Received: May 23, 2025. Revised: July 13, 2025. 352

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Clustering-Based and Multicollinearity Feature Selection in Macroeconomic Indicators for Predicting Loan Losses

Evelyn Sierra[1] Erick Delenia[1] Agus Tri Haryono[1] Riyanarto Sarno[1]
Hilmil Pradana[1] Diana Purwitasari[1]*

Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Indonesia
* Corresponding author's Email: diana@its.ac.id

Abstract: Macroeconomic indicators and their transformations are critical in credit risk modeling but often suffer from high multicollinearity, distorting model estimates and reducing predictive accuracy. Since these indicators are collected at varying frequencies, such as 3-month, 6-month intervals, transformations with lags and differences are required to align them, resulting in multiple derived features from the same indicator. Traditional methods like Principal Component Regression (PCR) reduce multicollinearity but compromise interpretability by converting features into latent components. To address this, we propose a hybrid feature selection framework that combines clustering algorithms (K-means, Agglomerative, DBSCAN) with Variance Inflation Factor (VIF) filtering. This approach preserves the original economic meaning of features—such as GDP growth, exchange rate changes, and inflation—while reducing dimensionality and multicollinearity. Experimental results show that the K-means + VIF method achieves best performance out of our experiments, with a test R-squared of 0.95653, MSE of 0.00395, RMSE of 0.06288, and a maximum VIF of 4.08053. These metrics demonstrate both high predictive accuracy and low multicollinearity. By retaining interpretable features and validating across pre- and post-pandemic periods, our framework offers a transparent solution for macroeconomic-based loan losses prediction.

Keywords: Clustering macroeconomic indicator, Default rates, Financial risk, Variance inflation factor.

1. Introduction

Financial risk—the potential for monetary loss due to market shifts, economic downturns, or borrower defaults—poses significant challenges for banks and financial institutions [1]. Among its forms, credit risk (the likelihood of loan losses) is particularly sensitive to macroeconomic conditions, shown as macroeconomic indicators, such as Gross Domestic Product (GDP), Inflation, and Exchange Rate [2], as economic health directly influences borrowers' repayment capacity [3, 4]. While regulatory standards like International Financial Reporting Standards (IFRS) require banks to proactively estimate loan losses using macroeconomic indicators, existing predictive models face critical limitations [5, 6]. Many studies rely on machine learning techniques, such as Decision Trees with SMOTE [3], but these struggle with large datasets, multicollinearity—a statistical issue that distorts risk models, among predictors, and a trade-off between accuracy and interpretability. This gap highlights the need for a more robust yet transparent approach to macroeconomic-based credit risk modeling—one that simplifies feature selection without compromising predictive power. To address these challenges, researchers have increasingly turned to machine learning models that incorporate macroeconomic indicators to improve prediction accuracy and risk assessment.

Several studies have explored the prediction of loan losses by incorporating macroeconomic indicators, employing machine learning techniques. For instance, one study used a Decision Tree model alongside SMOTE (Synthetic Minority Over-Sampling Technique) to address data imbalance in mortgage loan default prediction [3]. The research examined the impact of pandemic-related and

macroeconomic factors, but there are limitations in handling large datasets efficiently with Decision Tree. The author suggested that a more robust or deep learning method could improve performance. In contrast, our proposed method uses a straightforward regression model with selected features rather than all variables simultaneously, reducing complexity while maintaining predictive accuracy. Another example from Hariharan [4] utilized deep learning to enhance predictive accuracy but encountered computational and interpretability challenges, reinforcing the value of simpler regression-based approaches like ours. While some studies have focused on improving predictive accuracy through complex models or machine learning techniques, others have explored preprocessing strategies to handle challenges like multicollinearity and time-series structure in macroeconomic data.

The researchers applied Principal Component Regression (PCR) for feature selection, leveraging factor loadings to reduce multicollinearity [5]. However, a key limitation was the absence of stationary testing for time-series data and reliance solely on PCR to reduce multicollinearity. Our method addresses multicollinearity differently by using a similar method with PCR and applying the Variance Inflation Factor (VIF) calculations after feature selection, ensuring a more reliable variable combination in a regression model. Additionally, a different approach was introduced: Preptimize, an automated framework for time-series data preprocessing and forecasting. This method combines statistical and machine learning techniques to recommend optimal prediction models [6]. However, it demands substantial computational resources and assumes normally distributed data, limiting its applicability. Despite these advances, many existing methods still overlook the practical challenges of handling macroeconomic indicators—particularly the need for consistent feature transformation across varying data frequencies, which can introduce new issues like multicollinearity.

In loan losses modeling using macroeconomic indicators, feature transformations are often necessary due to the heterogeneity in data collection frequency—some indicators are recorded quarterly, others biannually—leading to the need for alignment through techniques such as lags, differences, or rolling statistics. While these transformations are essential for temporal consistency, they result in multiple derived features from the same underlying indicator. This redundancy can introduce severe multicollinearity, where features exhibit high intercorrelations that distort model coefficients and impair prediction stability. Although traditional multicollinearity filtering methods like the Variance Inflation Factor (VIF) are commonly employed, they focus solely on pairwise linear correlations and neglect more complex interdependencies that can exist in high-dimensional datasets. Moreover, relying exclusively on VIF may lead to the removal of informative variables or the retention of redundant ones without considering their broader relational structure. To address these limitations, we propose clustering-based feature selection that groups feature with similar behavior while explicitly avoiding the merging of features originating from the same macroeconomic source. Unlike PCR, which projects features into abstract components and sacrifices interpretability—a critical aspect in financial domains—our method maintains the original economic meaning of each selected feature. This ensures a balance between dimensionality reduction, multicollinearity mitigation, and interpretability, enhancing both model transparency and predictive performance.

This research starts by Section 2 reviews related works and existing methodologies in loan losses modeling, focusing on macroeconomic integration and feature selection techniques. Section 3 will explain about employing two feature selection methodologies that integrate clustering techniques, and combination features to perform regression analysis. The process begins with correlation analysis to eliminate highly redundant variables, followed by clustering to group similar macroeconomic indicators based on shared patterns, thereby reducing dimensionality and mitigating multicollinearity. Section 4 will explain the resulting clusters, and correlation-filtered features are then evaluated in an ensemble framework to determine the optimal combination for linear regression, and Section 5 concludes the paper with a summary of contributions and potential future research directions.

2. Related works

The International Financial Reporting Standards (IFRS) have influenced how financial institutions estimate credit losses through the Expected Credit Loss (ECL) model. Prior research has examined IFRS 9's shift from an "incurred loss", actual default happened, and recalculate the risk, to a "forward-looking", predicting future default rate, approach, emphasizing the need to incorporate macroeconomic factors into credit risk assessments [7].

2.1 Macroeconomic indicators

The selection and preparation of macroeconomic indicators data are critical in developing machine

learning models for credit risk assessment. Key variables such as Gross Domestic Product (GDP) Growth, Unemployment rate in a country, Import/Export Rate, and Exchange Rates exhibit a relationship with default probabilities that will be captured in the modelling process [8]. [9] prepared macroeconomic data from 11 Southeast Asian countries before applying machine learning models. First, the researchers collected historical data for key indicators like GDP Growth, Inflation Rate, Unemployment Rate, and Exchange Rates. Since these indicators come from different sources and may have missing values or inconsistencies, they use interpolation to fill the gaps. They may have performed normalization, scaling numbers to a similar range [10] or standardization, adjusting data to have a mean of zero and a standard deviation of one [11]. For a given macroeconomic indicator column X, the standardized value Z. Given t $\in\{1,2,3,\ldots,n\}$, i is the index of the data and n is the total sample data, this formula is calculated as shown in Eq. (1):

$$Z_t = \frac{X_t - \mu}{\sigma}, for\ t = 1,2,3,\ldots,n \quad (1)$$

Where, Xi is the original value of the macroeconomic indicator, μ is the mean of the macroeconomic indicator column, and σ as the standard deviation of the macroeconomic indicator column. Previous studies by [12] have demonstrated that incorporating macroeconomic indicators enhances the accuracy of the loan losses model by accounting for external economic shocks. However, macroeconomic indicators often exhibit heterogeneous frequencies; some are recorded monthly, like the Inflation Rate, others are quarterly, like GDP, while some are annually. This inconsistency complicates direct integration into predictive models, necessitating data transformation to ensure temporal alignment and stationarity [13].

Macroeconomic indicators require transformation for three key reasons:

- **Frequency Alignment**: Converting all indicators to a uniform frequency in monthly to match the loan losses data.
- **Stationarity**: Many macroeconomic indicators exhibit trends or seasonality, which can distort predictive models if not adjusted [14].
- **Interpretability**: Transforming raw macroeconomic indicator values into relative transformation improves model comparability and robustness [15].

This study applies the following transformation to extract economic signals while mitigating multicollinearity and non-stationarity. Given Δr_t as the difference at time *t* and *p* is the lag period number.

1. Periodic difference

Calculates the change in a macroeconomic indicator over a fixed period [16]. In which, r_t is the current value of the macroeconomic indicator and r_{t-p} is the value of *p* periods ago. We use this when we want to remove trends, and the macroeconomic indicators' raw value isn't as meaningful as how much it grew, as shown in Eq. (2).

$$\Delta r_t = r_t - r_{t-p} \quad (2)$$

For example, a sudden drop in GDP (Δr_t) result is negative. signals economic trouble, which may lead to higher loan losses.

2. Periodic Percentage Change

The periodic percentage change is used to measure relative growth over a period. Sometimes, a percentage change can make comparisons fair [17]. For example, banks care more about growth rates than the actual value (e.g., 4% GDP Growth is healthy, whereas 0% signals stagnation). Periodic change uses the formula in Eq. (3).

$$\Delta r_t = \left(\frac{r_t}{r_{t-p}} - 1\right) \times 100\% \quad (3)$$

3. Lagged Change

Economic policies take time to impact loans. High employment today might cause losses 6 months later. Using the current data can falsely imply instant effects [18]. Hence why we need lagged change for more stable modeling than raw data.

4. Logarithmic Transformation

Macroeconomic indicators can be in large-scale data, such as billions. By changing the transformation using logarithmic, it makes the model less sensitive to extreme values [19]. For example, if GDP grows 10% annually, logarithmic GDP adds a fixed amount each year with a linear trend. The formula is shown in Eq. (4).

$$r_t = log(r_t) \quad (4)$$

2.2 Preprocessing for time series data

Time series preprocessing is crucial for accurate modeling, especially for macroeconomic indicators like GDP or exchange rates, which often have trends, seasonality, and noise. Since non-stationary data can lead to unreliable results, methods like stationarity

adjustments are commonly used. Research by Usmani, Mehak, et. al. introduced Preptimize [6]. Preptimize is an automated time series preprocessing framework that addresses common data challenges through an integrated pipeline of statistical and machine learning techniques.

Preptimize starts with a data quality assessment, handling missing values using context-aware strategies like linear interpolation for trends or ML-based imputation for volatile data. It detects outliers via robust methods, IQR for normal data, and MAD for non-Gaussian [20], and checks stationarity using the Augmented Dickey-Fuller (ADF) [21] and the KPSS tests [22]. Based on results, it automatically applies differencing or log transformations to remove trends, ensuring optimal preprocessing without manual input. Another approach by Bülte, Christopher, et al. improves missing data imputation by first learning the underlying distribution, capturing temporal trends, and cross-variable relationships [23]. While the method improves imputation accuracy, it has some limitations. First, learning the data distribution before imputation can be computationally expensive, especially for large-scale datasets. Second, the model's performance depends on the quality of the initial distribution learning—if this step fails, imputation may be less accurate.

The Augmented Dickey-Fuller (ADF) test has been widely adopted in time series analysis to assess stationarity, a fundamental requirement for many forecasting models [24]. Stationary processes, characterized by constant mean, variance, and autocorrelation over time, enable reliable parameter estimation in autoregressive models. The ADF test specifically checks for a unit root—a cause of non-stationarity—where trends or random walks distort time series analysis. The test's null hypothesis states that the time series has a unit root (i.e., it is non-stationary), while the alternative hypothesis suggests stationarity. To reject the null hypothesis (and confirm stationarity), the test statistic must be more negative than the critical value (e.g., −2.868 at 5% significance), and the p-value must be below the significance level (e.g., 0.05). The experiment by Sreehari et. al. demonstrated that the ADF Test outcome successfully converted a non-stationary series into a stationary one. However, their demonstration does not clarify whether higher-order differencing or seasonal adjustments were needed for more complex patterns. Additionally, while differencing resolves trends, it may amplify noise or remove meaningful long-term patterns if overapplied. Future steps might involve testing alternative transformations (e.g., logarithms or decomposition) if differencing alone proves insufficient.

This brings to our attention that the preprocessing process will be gone through with transformed variables and observe their ADF Test. Our preprocessing pipeline will apply first-order differencing to non-stationary variables identified by the ADF test, followed by retesting to confirm stationarity of the variables.

2.3 Feature clustering and multicollinearity in regression model

In regression modeling, especially within macroeconomic contexts, multicollinearity among predictor variables poses significant challenges. Multicollinearity occurs when independent variables are highly correlated, leading to unreliable coefficient estimates and inflated standard errors. This issue complicates the interpretation of regression results and can diminish the predictive power of the model. Addressing multicollinearity is crucial for developing robust financial risk prediction models. A commonly used diagnostic tool for detecting multicollinearity is the VIF represented in τj, which is calculated as shown in Eq. (5).

$$\tau_j = \frac{1}{1- R^2{}_j} \tag{5}$$

Where R2j is the coefficient of determination obtained by regressing the j-th independent variable against all other independent variables. A τj value above 10 is often considered a sign of severe multicollinearity [25].

Another method that is commonly used but still crucially implemented is the Principal Component Analysis (PCA) technique. By transforming the original variables into a new set of uncorrelated components, PCA captures the maximum variance in the data with fewer variables [26]. This transformation aids in mitigating multicollinearity and simplifies the modeling process. PCA has been effectively applied in various domains, including macroeconomic analysis, to enhance model performance. PCA computes the covariance matrix of C of the standardized data Z, which we explained in Section 2.1, within the n sample data, the formula as shown in Eq. (6):

$$\boldsymbol{C_t} = \frac{1}{n} Z^t Z \tag{6}$$

The covariance matrix quantifies pairwise relationships between indicators. For example, a positive off-diagonal entry in C indicates that two

variables (e.g., GDP Growth) tend to move together, signaling potential multicollinearity. The core of PCA lies in the eigendecomposition, where C is factorized into eigenvalues (λ) and eigenvectors (v), as shown in Eq. (7):

$$\boldsymbol{C}v = \lambda v \tag{7}$$

Eigenvalues represent the variance explained by each principal component (P), while eigenvectors (v) define their directions (loadings). For instance, the first eigenvector (v1) might assign high weights to GDP and inflation.

Finally, the original data is projected onto the top-k eigenvectors (vk) to derive the principal components, as shown in Eq. (8):

$$P_k = Z \cdot v_k \tag{8}$$

Though PCA is commonly used in classification problems, it is also applicable to multivariate time series data. In this context, PCA helps extract dominant patterns or temporal features from high-dimensional time series. One drawback of PCA is that the resulting components are linear combinations of all original features, which may lack interpretability [27]. To illustrate the interpretability limitation of PCA, Table 1 presents the PCA loadings derived from our macroeconomic dataset. Each principal component is a linear combination of multiple transformed indicators, such as EXPORT, GOVERNMENT BOND, and various lagged values of USDIDR. For instance, PC2 is dominated by GOVERNMENT BOND (loading = 0.855), but also includes contributions from USDIDR and its lagged changes. This blending of features makes it difficult to trace predictive influence back to specific economic indicators, which is problematic in financial modeling where interpretability is essential.

Additionally, the presence of similar loadings across multiple lagged USDIDR features suggests redundancy and potential multicollinearity. These findings support our argument that PCA sacrifices interpretability and fails to preserve the economic meaning of features—especially when indicators are transformed and lagged. This motivates our use of clustering-based feature selection, which retains original feature identities while reducing dimensionality and multicollinearity.

Feature clustering groups correlated features into homogeneous clusters, offering an alternative to PCA for dimensionality reduction. One popular technique is hierarchical clustering, which computes pairwise distances between features using correlation or Euclidean distance. The linkage criterion determines the distance between clusters, such as in Eq. (9). Given the two macroeconomic indicators represented as vectors x and y; xi and yi is the individual observation. the Euclidean distance d is shown in Eq. (9):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}\left(x_i - y_j\right)^2} \tag{9}$$

This method produces a dendrogram that helps identify representative features from each cluster, reducing redundancy and improving model robustness [28]. Combining clustering with multicollinearity diagnostics enhances feature selection. For instance, after clustering, one can apply VIF filtering to each representative variable to ensure low multicollinearity. The final feature subset is shown in Eq. (10).

$$\tau^* = \arg max\left\{x_i \in x_{selected} \middle| \tau_j < \theta\right\} \tag{10}$$

Table 1. Principal Component Analysis' Loadings

	PC1	PC2	PC3	PC4	PC5	PC6
EXPORT	-0.1219	-0.1261	0.0632	-0.0458	0.2224	-0.0047
GOVERNMENT BOND	0.0950	0.0855	-0.1843	0.2604	0.1405	-0.0057
USDIDR	-0.1326	0.1098	-0.0768	0.1947	-0.0632	-0.0714
USDIDR_chng_lag5	0.0713	0.1429	-0.1131	-0.1923	0.1107	-0.1589
…	…	…	…	…	…	…
GOVERNMENT BOND_log_lag9	0.1585	0.0535	0.1457	-0.0436	-0.02261	0.1330
GOVERNMENT BOND_log_lag10	0.1531	0.0269	0.1765	0.0449	-0.0214	0.0322
GOVERNMENT BOND_log_lag11	0.1431	-0.0028	0.1952	0.1377	-0.0443	-0.0540

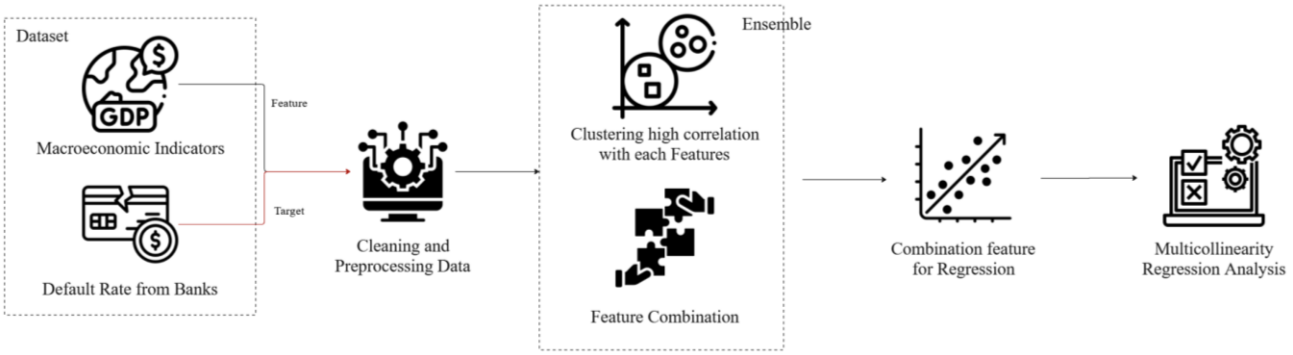

Figure. 1 Clustering-based and Multicollinearity Analysis for Financial Risk Framework

Where θ is a predefined VIF threshold (e.g., 5 or 10). This dual-stage approach enhances model stability [29].

In conclusion, PCA and feature clustering are effective tools for addressing multicollinearity and reducing feature dimensionality in macroeconomic regression models. However, few studies have integrated these techniques into a unified framework specifically designed for financial risk prediction. This research aims to bridge that gap by proposing a clustering-based and multicollinearity-aware feature selection strategy tailored to macroeconomic indicators.

3. Methodology

Based on the discussion in Section 2, it can be determined that the ensemble feature selection with clustering and multicollinearity. This section has an explanation of the proposed method. This research employs an ensemble feature selection methodology that integrates clustering, a combination of the features, and checks their multicollinearity.

The methodology is structured into four main phases: (1) data cleaning and preprocessing, (2) ensemble-based feature selection through clustering and statistical testing, (3) feature combination for regression modeling, and (4) multicollinearity and regression analysis, as shown in Fig. 1.

3.1 Cleaning and preprocessing data

The first phase focuses on data cleaning and preprocessing. This step ensures that the time-series data are suitable for modeling. Missing values are handled through interpolation techniques or forward filling to preserve temporal continuity. All datasets are aligned to a common monthly frequency. For variables such as GDP that are originally reported quarterly, linear interpolation is used to approximate monthly values, as shown in Table 1. Furthermore, to prepare the data for regression analysis, the stationarity of each time series is evaluated using the Augmented Dickey-Fuller (ADF) test. Stationarity is essential in time-series modeling to avoid spurious correlations. Any non-stationary variables are differenced or log-transformed accordingly.

In the second phase, an ensemble feature selection strategy is implemented. This method integrates multiple statistical and machine learning-based techniques to reduce dimensionality and identify the most informative macroeconomic features. The first step is to perform a Pearson correlation analysis among all features. Features exhibiting strong correlation (typically above 0.85) are considered redundant. As part of the ensemble feature selection strategy, a Pearson correlation analysis was conducted to measure the linear relationships between all pairs of macroeconomic indicators. The goal of this step is to identify variables that are highly correlated, which may contribute redundant information to the regression model.

Fig. 2 shows the distribution of the Pearson correlation coefficients across all feature pairs.

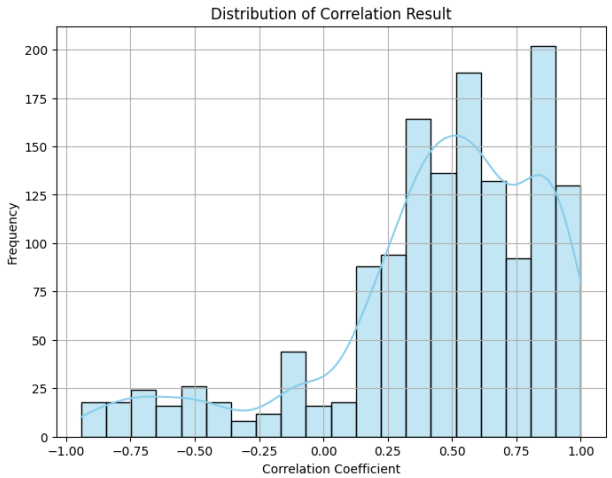

Figure. 2 Distribution of the Correlated Macroeconomic Indicators to Default Rate

Table 2. Preview of Macroeconomic Indicators with Transformation

DATE	EXPORT	GOVERNMENT BOND	IMPORT	USDIDR	USDIDR_chng	USDIDR_chng_lag1
31/10/2013	0.0244	0.0747	-0.0891	11234	0.1683	0.2112
30/11/2013	-0.0232	0.0866	-0.1054	11977	0.2469	0.1683
31/12/2013	0.1022	0.0845	-0.0081	12189	0.2604	0.2469
31/01/2014	-0.0587	0.0903	-0.0346	12226	0.2606	0.2604
31/07/2022	0.3151	0.0712	0.3985	14958	0.0322	0.0242
31/08/2022	0.3025	0.0712	0.3281	14875	0.0348	0.0322
30/09/2022	0.2011	0.0737	0.2202	15247	0.0657	0.0348
31/10/2022	0.1193	0.0753	0.1744	15542	0.0945	0.0657

The histogram indicates that most correlations fall in the range of 0.4 to 1.0, with a notable concentration between 0.6 and 0.9. This suggests the presence of substantial multicollinearity among the indicators, which could potentially distort regression estimates. To address this, pairs of variables exhibiting a strong positive correlation (greater than 0.85) were grouped together, and only one representative feature from each group was retained for further analysis. This process helps in simplifying the model while preserving the essential information content from the dataset. These highly correlated variables are then grouped using a clustering algorithm to identify clusters of similar indicators. In parallel, the ADF test results are used to eliminate non-stationary features that may undermine model reliability. The ensemble approach allows for a more balanced feature selection process by combining both statistical validity (stationarity and correlation) and structural similarity (clustering).

3.2 Feature clustering and combination

Following feature selection, the third phase involves combining selected features for regression modeling. From each cluster of correlated indicators, a representative feature is selected—often the most statistically robust or the most interpretable in economic terms. These representative features are then combined to form a feature set that captures the core information from the original high-dimensional dataset. we begin by generating a comprehensive set of transformed features from the original macroeconomic indicators. Let the original feature set be denoted by $F= \{F_1,F_2,\ldots,F_i\}$, where i is the index of raw variables, each $F_i$represents an index in macroeconomic indicators such as GDP, USDIDR, Inflation, and others. Each feature undergoes a predefined set of transformations, denoted by $T= \{T_1,T_2,\ldots,T_i\}$, which includes logarithmic transformation $T_1(F_i) = log(F_i)$ to normalize skewed distributions; Differencing $T_2(F_i) = \Delta (F_i)$ to remove trends; Lagged values $3_2(F_i) = F_{i-p}$ to capture delayed effects. The transformed feature $F_{i,j}$ is defined as the result of apply transformation T_j to the original indicator F_i.

This dataset typically has high dimensionality and multicollinearity, as correlated transformations (e.g., GDP and its lagged values) inflate feature space redundantly. We apply a clustering algorithm (e.g., K-means or hierarchical clustering) to group features that share similar eigenvector loadings. This results in a set of clusters $\boldsymbol{C}= \{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_p\}$, where $\boldsymbol{C}_p$ contains features with similar temporal patterns. For instance:

- Cluster $\boldsymbol{C}_1$ might group GDP, Inflation, and their lagged values.
- Cluster $\boldsymbol{C}_2$ could contain Unemployment Rate and its differenced transformations.

The Euclidean Distance $d(F_a,F_b)$ between two transformed features is computed as shown in Eq. (12):

$$d(F_a, F_b; n) = \sqrt{\sum_{t=1}^{n}(F_a - F_b)^2} \quad (12)$$

We use Silhouette Score to determine the best cluster for each of our clustering methods. For K-Means and Agglomerative clustering, we used the Silhouette Score to determine the ideal number of clusters. This score measures how well data points are grouped within their clusters and separated from other clusters. The highest average Silhouette Score indicated that K=9 was optimal for K-Means, and K=6 was optimal for Agglomerative clustering. These findings are illustrated in Figs. 3 and 4, respectively. For DBSCAN, we used epsilon by analyzing nearest density. This process led to an optimal value of 2.2, as shown in Fig. 5.

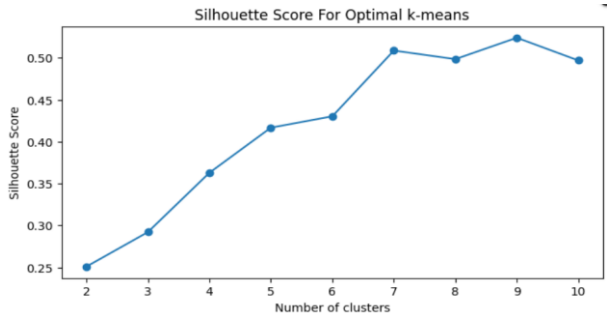

Figure. 3 Silhouette Score for K-Means

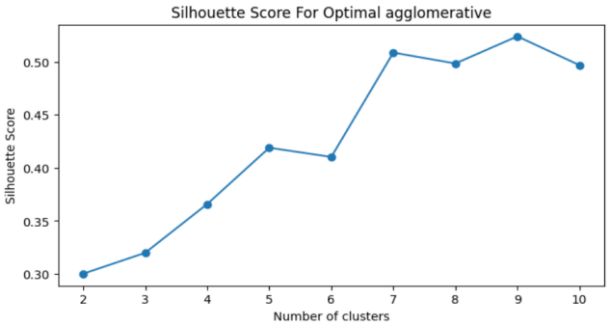

Figure. 4 Silhouette Score for Agglomerative Cluster

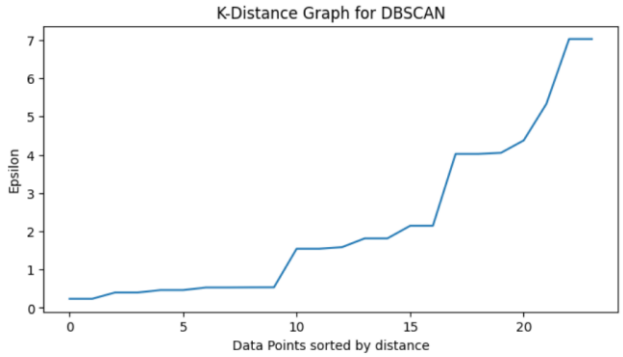

Figure. 5 Epsilon for DBScan Cluster

3.3 Multicollinearity analysis

The final phase of the methodology addresses the issue of multicollinearity within the regression model. From each cluster $\boldsymbol{C}_p$,we select the most representative feature while ensure a multicollinearity constraint. Additionally, each valid P must exhibit low multicollinearity. The VIF is computed for all features in P, and the combination is rejected if any feature exceeds a threshold θ (e.g., $\theta = 10$), as shown in Eq. (13).

$$\tau^* = \arg max \{F_a \in \boldsymbol{C}_p | \tau_j < \theta\} \quad (13)$$

Features with a VIF value exceeding the threshold of 10 are considered problematic, as they can inflate the standard errors of regression coefficients and lead to unreliable interpretations. Such features are reviewed and removed, if necessary, followed by re-estimation of the regression model. The final model is evaluated using performance metrics such as R-squared (R^2) and Root Mean Squared Error (RMSE) to assess goodness-of-fit and predictive accuracy. Through this structured methodology, the study ensures that the selected features are not only statistically valid and non-redundant but also effective in explaining and predicting financial risk. In this study, we adopted a VIF threshold of 10, which is widely accepted in regression modeling, and alternative threshold of 5 as comparison. This threshold was applied consistently across all feature selection strategies to ensure comparability. Furthermore, we employed the Shapiro-Wilk test to assess whether the residuals follow a normal distribution [30]. This test proves particularly effective for small to moderate sample sizes and significantly supports the validity of our statistical inferences when its p-value exceeds 0.05. Additionally, we applied the Durbin-Watson test to detect autocorrelation in the residuals. A statistic close to 2 confirms no significant autocorrelation, which is crucial for ensuring unbiased coefficient estimates [31].

Our ensemble feature selection framework implicitly performs a form of sensitivity analysis. Each macroeconomic indicator undergoes multiple transformations—such as differencing, percentage change, logarithmic scaling, and lagging—resulting in a wide range of derived features. By clustering these transformed features and applying VIF-based filtering, the model evaluates various transformation combinations and selects those that optimize predictive performance while minimizing multicollinearity. This process ensures that the final feature set reflects the most effective transformation strategy for each indicator.

4. Experiments

4.1 Data collection

In this study, we gathered economic data from Indonesia to help predict loan losses, which shows how much money banks might lose if borrowers can't repay their loans. We focused on monthly data from the past 10 years (2013–2022) to make sure the analysis reflects current economic conditions. The data includes important indicators like inflation, interest rates, unemployment, GDP growth, car sales, and property prices.

Table 3. Macroeconomic Indicators

Code	Indicator	Purpose
BI_RATE	Central Bank Interest Rate	Measures monetary policy stance; influences borrowing costs and inflation.
CPI	Consumer Price Index	Tracks inflation by measuring changes in the price level of consumer goods.
UNEMPLOYMENT_RATE	Unemployment Rate	Indicates labor market health and economic stability.
GDP_GROWTH	GDP Growth Rate	Reflects overall economic expansion or contraction.
CAR_SALES_MTH	Monthly Car Sales	Gauges consumer demand and manufacturing sector performance.
RESIDENTIAL_INDEX	Residential Property Index	Tracks real estate market trends and housing demand.
COAL_PRICE	Coal Price	Influences energy costs and industrial production.
OIL_PRICE	Oil Price	Affects transportation costs, inflation, and trade balances.
GOVERNMENT BOND	Government Bond Yield	The yield on government bonds reflects the cost at which the government can borrow money.
INTL_RESERVE	International Reserves	Measures a country's foreign currency liquidity and financial stability.
EXPORT	Export Value	Indicates trade performance and external demand for domestic goods.
IMPORT	Import Value	Indicates total value of goods and services imported by a country
CCI	Consumer Confidence Index	Assesses household spending sentiment and future economic activity.
USDIDR	Currency Exchange Rate	Indicates the value of international transactions
RETAIL_SALES	Retail Sales Volume	Reflects consumer spending trends and domestic demand.

These indicators give us a picture of how the economy is doing and how it might affect people's ability to repay loans. Details of the macroeconomic purpose are shown in Table 3 and can be obtained from the Central Bureau of Statistics in Indonesia (Badan Pusat Statistik Indonesia). To obtain the relevant datasets, researchers may use the search functionality provided on the website by entering keywords such as "macroeconomic indicators monthly". This query yields a list of available datasets categorized by indicator type and reporting frequency, along with the corresponding years of data availability. The selected datasets can then be downloaded in spreadsheet format for further preprocessing and analysis. Alongside this, we used loan default data. This data shows the percentage of loans that were not paid back. It depicts as a key signal of financial stress in the banking system. While detailed default data is usually kept private by banks, the Financial Services Authority of Indonesia shares summary reports that researchers can use. The dataset used in this study includes a comprehensive set of macroeconomic indicators, such as Gross Domestic Product (GDP), inflation rates, and other key economic variables. Another challenge that we faced is not all indicators are reported monthly. For example, GDP is usually reported every three months. To match everything to a monthly timeline, we used a filling method called forward-fill. This means we assume the same value for each month in a quarter, which avoids creating fake or misleading data. By aligning all the data to a monthly format, we can build a model that looks at how changes in the economy affect loan repayments over time as such, helping banks and regulators make better decisions.

4.2 Clustering results

Clustering plays an important role in grouping features that share similar characteristics. In this study, clustering is used to help reduce multicollinearity by identifying groups of features that are highly correlated with each other. After identifying these groups, the next step is to select a representative feature from each cluster to be used in the predictive model.

To do this, two different strategies are applied for feature selection based on clustering results. The first strategy is based on centroid-based clustering, where the aim is to divide features into a predefined number of clusters. In this approach, the centroid (or center point) of each cluster is used to represent the group. Features that are closest to the centroid are selected as representative features. The clustering is performed using a distance metric—commonly

Euclidean distance—to assign each feature to the nearest centroid. This assignment and centroid recalculation process continues iteratively until the positions of the centroids become stable, indicating that optimal cluster formation has been achieved.

The second strategy is a random feature combination method. In this approach, after clustering, several random combinations of features are selected from each cluster. These combinations are then evaluated using the VIF to check for multicollinearity. A VIF threshold of 10 is used to filter out highly collinear features. The selected combinations that pass this multicollinearity check are then used in regression modeling, and the model performance is compared to find the best feature set. This study uses three clustering algorithms:

- K-Means Clustering, with the optimal number of clusters determined as K = 9, shown in Fig. 6.
- Agglomerative Clustering, with the optimal number of clusters as K = 6, as shown in Fig. 7, and
- DBSCAN Clustering, with an optimal ε value of 2.2, as shown in Fig. 8.

These clustering results are visualized in Figs. 6-8. The silhouette score is used to evaluate and compare the quality of clustering in K-Means and Agglomerative methods, ensuring that the chosen number of clusters leads to well-defined and distinct groups.

Based on the analysis of three clustering methods in Table 3, K-Means with 9 clusters provides a granular and distinct partitioning of features, indicating high similarity within compact, spherical groups. Agglomerative Clustering, using 6 clusters, forms broader, more generalized groupings, which can be beneficial for dimensionality reduction and interpretability, though it may sacrifice some feature specificity. In contrast, DBSCAN, with an epsilon of 2.2, identified 5 density-based clusters and a significant number of outliers ("noise" points), highlighting its effectiveness with irregularly shaped clusters and noisy data, but suggesting that many features may not align with dominant groups.

Table 4. Modeling Results of the Feature Selection Scenarios

Feature Selection	R-squared Train.	R-squared Test.	MSE	RMSE	Highest VIF	Shapiro Wilk p-value	Durbin Watson stats
VIF 5	0.9173	0.8433	0.0142	0.1193	4.6827	0.1221	2.0980
VIF 10	0.9174	0.8436	0.0142	0.1193	9.8432	0.1273	2.0792
K-means centroid	0.8862	0.9418	0.0052	0.0727	24.0830	0.3965	2.3366
Agglomerative_Centroid	0.8862	0.9418	0.0052	0.0727	24.0830	0.3965	2.3366
DBSCAN Centroid	0.8951	0.9144	0.0077	0.0882	6.3249	0.8219	2.0184
Agglomerative + VIF 10	0.8889	0.9565	0.0039	0.0628	4.0805	0.3485	2.1866
DBSCAN + VIF 10	0.8801	0.9496	0.0045	0.0676	6.4573	0.6138	2.2018
Principal Component Regression [5]	0.6464	0.8979	0.1005	0.1002	1.2300	0.6483	1.5543
Preptimize [6]	0.6501	0.8757	0.0110	0.1053	1.2300	0.6881	1.5926
K-means + VIF 10 (ours)	**0.8889**	**0.9565**	**0.0039**	**0.0628**	**4.0805**	**0.3485**	**2.1866**

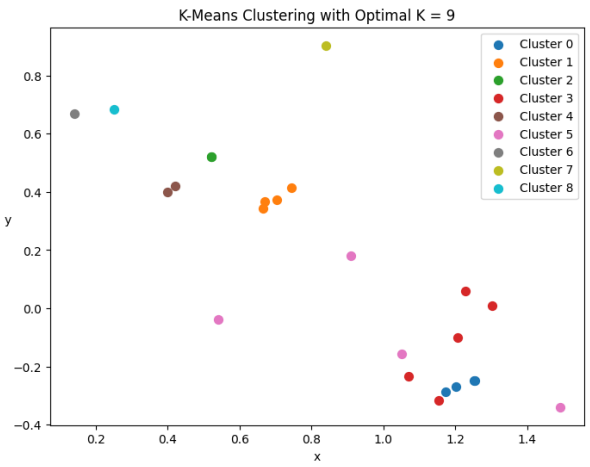

Figure. 6 Distribution of Features based on KMeans

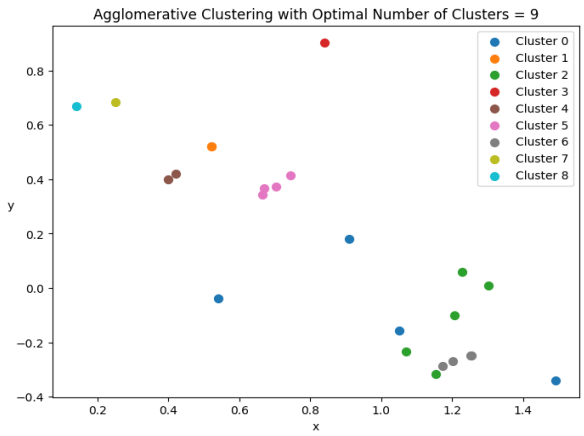

Figure. 7 Distribution of Features based on Agglomerative

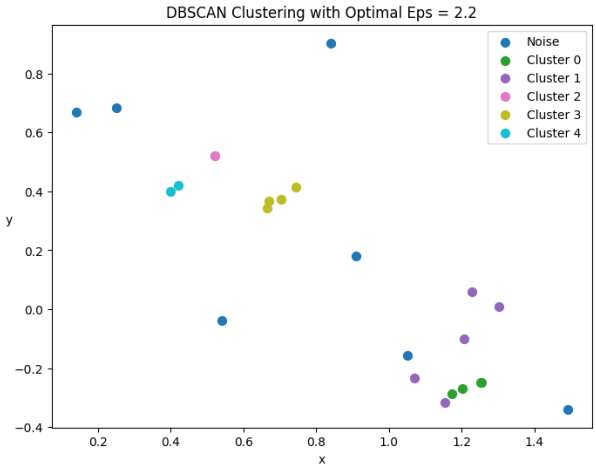

Figure 8. Distribution of Features based on DBScan

4.3 Discussion

This study tested several feature selection strategies to improve prediction accuracy and reduce multicollinearity in loan loss modeling. Table 4 shows the performance of different methods, including traditional VIF filtering, clustering-based selection, and hybrid approaches. Among all methods, the combination of K-Means clustering with VIF filtering (threshold ≤ 10) gave the best results. It achieved the highest test R-squared value of 0.95653, the lowest Mean Squared Error (MSE) of 0.00395, and the lowest RMSE of 0.06288. It also maintained a low maximum VIF of 4.08053, which means the selected features were not highly correlated. This shows that the hybrid method is effective in selecting meaningful features while keeping the model stable and accurate. Other clustering methods like Agglomerative and DBSCAN also performed well, especially when combined with VIF filtering. However, they did not outperform the K-Means hybrid. For example, DBSCAN + VIF 10 had a test R-squared of 0.94968 and RMSE of 0.06766, which is slightly lower than the K-Means hybrid. While alternative threshold like 5 may offer different trade-off, our results show that the selected features consistently exhibit low multicollinearity, with maximum VIF values well below the cutoff. In our best-performing model (K-Means + VIF 10), the Shapiro-Wilk test returned a p-value of 0.3485, and the Durbin-Watson statistic was 2.1866. These results confirm that we adequately met both assumptions.

We also compared our method with Principal Component Regression (PCR) and Preptimize, which are commonly used in time-series modeling. While PCR had a low RMSE, it uses abstract components that are hard to interpret in economic terms. Preptimize also showed good performance but requires more computational resources and assumes normal data distribution. In contrast, our method keeps the original economic meaning of each feature, making it more practical for financial analysis and regulatory reporting. To test the model's robustness, we used a time-based holdout evaluation. The training data covered the pre-pandemic period (2013–2019), and the testing data included the post-pandemic recovery (2020–2022). This setup allowed us to see how well the model performs during economic shocks. The results confirmed that our method generalizes well and avoids overfitting to historical patterns.

In summary, the proposed clustering-based feature selection combined with VIF filtering improves both prediction accuracy and interpretability. It performs well across different economic conditions and offers a reliable approach for credit risk modeling.

5. Conclusion

The results of the experiments demonstrate that clustering-based feature selection, particularly when combined with multicollinearity filtering using the VIF, effectively uncovers underlying structures in macroeconomic data and improves loan loss prediction models. For example, the grouping of indicators through K-means clustering highlights how certain variables, such as exchange rate movements and their lags, tend to behave similarly, suggesting the presence of common economic forces influencing multiple indicators. This provides valuable insights into how economic shocks or policy changes ripple through the system.

Reducing multicollinearity through the combination of clustering and VIF filtering enhances the interpretability of the model. By selecting features with lower VIF values, each variable contributes more distinct information to the prediction of default risk. This clarity is crucial in economic decision-making, allowing analysts and policymakers to identify which specific indicators are most influential in driving credit risk and to avoid redundancy in economic narratives or risk models. Moreover, the improved prediction performance—evidenced by higher R-squared values and lower MSE and RMSE in the K-means + VIF 10 configuration—suggests strong generalization capabilities. This is particularly important during periods of economic uncertainty or financial instability, where reliable credit risk forecasts enable timely interventions.

When compared to Principal Component Regression (PCR), the proposed method not only offers comparable or better predictive performance but also retains the interpretability of the original macroeconomic indicators. While PCR compresses data into latent components that lack direct economic meaning, our approach preserves real-world variables, making it more practical for regulatory reporting, credit assessments, and stress testing. This transparency is critical for justifying decisions to stakeholders.

Overall, this study contributes a novel hybrid framework that integrates clustering algorithms (K-Means, Agglomerative, DBSCAN) with multicollinearity diagnostics (VIF filtering) to enhance feature selection in macroeconomic time-series modeling. The proposed method outperforms traditional techniques such as Principal Component Regression and Preptimize in both predictive accuracy and interpretability. By validating the model across structurally different economic periods and confirming key regression assumptions, this research provides a statistically robust and practically interpretable approach for credit risk forecasting under IFRS 9.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

This study can run well and successfully because of the following research contributions: Conceptualization by Prof. RS and ATH; Methodology by ATH, ES, and ED; Data Collecting and Preprocessing by ES; Experiment by ED; validation and analysis by Prof. RS, ATH, HP, ES and ED; visualization, and editing by ES and ED; preparation of the original draft by ES and ED; supervision by Prof. RS and HP.

Acknowledgments

This research was funded by Institut Teknologi Sepuluh Nopember (ITS) under Penelitian Dana Departemen Program, Penelitian Keilmuan, Penelitian Flagship, Penelitian Kolaborasi Pusat and under the Project Scheme of the Publication Writing and Intellectual Property Rights (IPR) Incentive (Penulisan Publikasi dan Hak Kekayaan Intelektual/PPHKI) Program 2024.

References

[1] A. Colasante and L. Riccetti, “Financial and non-financial risk attitudes: What does it matter?”, *Journal of Behavioral and Experimental Finance*, p. 100494, 2021.

[2] M. M. Khyareh and N. Rostami, “Macroeconomic Conditions, Innovation and Competitiveness”, *Journal of the Knowledge Economy*, Vol. 13, pp. 1321-1340, 2021.

[3] B. Fazlija, “Credit Risk Assessment via Machine Learning: Impact of Pandemic and Macroeconomic Variables on Mortgage Loan Default Prediction”, *Zürcher Hochschule für Angewandte Wissenschaften*, 2022.

[4] H. P. Kothandapani, “Application of Machine Learning for Predicting U.S. Bank Deposit Growth: A Univariate and Multivariate Analysis of Temporal Dependencies and Macroeconomic Interrelationships”, *Journal of Empirical Social Science Studies*, Vol. 4, No. 1, pp. 1-20, 2020.

[5] D. G. Breed, J. Hurter, M. Marimo, M. Raletjene, H. Raubenheimer, V. Tomar, and T. Verster, “A Forward-Looking IFRS 9 Methodology, Focusing on the Incorporation of Macroeconomic and Macroprudential Information into Expected Credit Loss Calculation”, *Risks*, Vol. 11, No. 3, p. 59, 2023.

[6] M. Usmani, Z. A. M. A. Zulfiqar, and R. Qureshi, “Preptimize: Automation of Time Series Data Preprocessing and Forecasting”, *Algorithms*, Vol. 17, No. 8, p. 332, 2024.

[7] Q. Jin, and S. Wu, “Shifting from the incurred to the expected credit loss model and stock price crash risk”, *Journal of Accounting and Public Policy*, Vol. 42, No. 2, p. 107014, 2023.

[8] E. Yeboah, H. C. Chibalamula, and F. Atiso, “The effect of international trade on economic growth: Evidence from Ghana”, *Global Journal of Business, Economics and Management: Current Issues*, Vol. 13, No. 1, pp. 91-105, 2023.

[9] L. D. T. Ngoc, Ngo-Thi-Thu-Trang, K.-M. Kim, V. Pham, and H.-N. Nguyen, “Studying of Machine Learning Models for Forecasting Macroeconomic Indicators”, In: *Proc. of 2025 27th International Conference on Advanced Communications Technology (ICACT), Pyeong Chang*, 2025.

[10] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, “Normalization Techniques in Training DNNs: Methodology, Analysis and Application”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 8, pp. 10173-10196, 2023.

[11] A. Elen, and E. Avuçlu, “Standardized Variable Distances: A distance-based machine learning method”, *Applied Soft Computing*, Vol. 98, p. 106855, 2021.

[12] G. Jiménez, S. Ongena, J.-L. Peydró, and J. Saurina, "Macroprudential Policy, Countercyclical Bank Capital Buffers, and Credit Supply: Evidence from the Spanish Dynamic Provisioning Experiments", *Journal of Political Economy*, Vol. 125, No. 6, pp. 2126-2177, 2017.

[13] A. M. Elberry, R. Garaffa, A. Faaij, and B. V. D. Zwaan, "A review of macroeconomic modelling tools for analysing industrial transformation", *Renewable and Sustainable Energy Reviews*, Vol. 199, p. 114462, 2024.

[14] J. D. Hamilton, "Why You Should Never Use the Hodrick-Prescott Filter", *The Review of Economic and Statistics*, Vol. 100, No. 5, pp. 831-843, 2018.

[15] A. Kanas, and P. Molyneux, "Macro stress testing the U.S. banking system", *Journal of International Financial Markets, Institutions and Money*, Vol. 54, pp. 204-227, 2018.

[16] F. Yurdakul, "Macroeconomic Modelling of Credit Risk for Banks", In: *Proc. of Procedia - Social and Behavioral Sciences*, Vol. 109, pp. 748-793, 2014.

[17] I. E. Ceylan, "The Impact of Firm-Specific and Macroeconomic Factors on Financial Distress Risk: A Case Study from Turkey", *Journal of Accounting and Finance*, Vol. 9, No. 3, pp. 506-517, 2021.

[18] Z. Liu, Y. Luo, and M. Duan, "Macroeconomic factors, industrial enterprises, and debt default prediction: Based on the VAR-GRU model", *Finance Research Letters*, Vol. 78, p. 107122, 2025.

[19] Y. Luo, H. Cui, H. Zhong, and C. Wei, "Business environment and enterprise digital transformation", *Finance Research Letters*, Vol. 57, p. 104250, 2023.

[20] J. Wang, "An Intuitive Tutorial to Gaussian Process Regression", *Computing in Science & Engineering*, Vol. 25, No. 4, pp. 4-11, 2023.

[21] M. Ahmed, M. Irfan, A. Meero, M. Tariq, U. Comite, A. A. A. Rahman, M. S. Sial, and S. B. Gunnlaugsson, "Bubble Identification in the Emerging Economy Fuel Price Series: Evidence from Generalized Sup Augmented Dickey-Fuller Test", *Processes*, Vol. 10, No. 1, p. 65, 2022.

[22] M. K. Hassan, H. Kazak, U. Adıgüzel, M. A. Gunduz, and A. T. Akcan, "Convergence in Islamic financial development: Evidence from Islamic countries using the Fourier panel KPSS stationarity test", *Borsa Istanbul Review*, Vol. 23, No. 6, pp. 1289-1302, 2023.

[23] C. Bülte, M. Kleinebrahm, H. Ü. Yilmaz, and J. Gómez-Romero, "Multivariate time series imputation for energy data using neural networks", *Energy and AI*, Vol. 13, p. 100239, 2023.

[24] D. A. Dickey, and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series With a Unit Root", *Journal of the American Statistical Association*, Vol. 74, No. 366, pp. 427-431, 1979.

[25] A. Yanke, N. E. Zendrato, and A. M. Soleh., "Handling Multicollinearity Problems in Indonesia's Economic Growth Regression Modeling Based on Endogenous Economic Growth Theory", *Indonesian Journal of Statistics and Its Applications*, Vol. 6, No. 2, p. 214-230, 2022.

[26] S. Crépey, N. Lehdili, N. Madhar, and M. Thomas, "Anomaly Detection in Financial Time Series by Principal Component Analysis and Neural Networks", *Algorithms*, Vol. 15, No. 10, p. 385, 2022.

[27] X. Wan, H. Li, L. Zhang, and Y. J. Wu, "Dimensionality reduction for multivariate time-series data mining", *The Journal of Supercomputing*, Vol. 78, No. 4, pp. 9862-9878, 2022.

[28] H. Li, and T. Du, "Multivariate time-series clustering based on component relationship networks", *Expert Systems with Applications*, Vol. 173, p. 114649, 2021.

[29] M. R. Machado, and S. Karray, "Assessing credit risk of commercial customers using hybrid machine learning algorithms", *Expert Systems with Applications*, Vol. 200, p. 116889, 2022.

[30] E. González-Estrada, J. A. Villaseñor, and R. Acosta-Pech, "Shapiro-Wilk test for multivariate skew-normality", *Computational Statistics*, Vol. 37, pp. 1985-2001, 2022.

[31] W. Krämer, "Durbin-Watson Test", *International Encyclopedia of Statistical Science*, p. 762-764, 2025.

[32] R. S. Koijen and M. Yogo, "The Fragility of Market Risk Insurance", *The Journal of Finance*, Vol. 77, No. 2, pp. 815-862, 2022.

[33] H. Zhao, Z. Jiao, J. Wang, and A. Kamar, "Corporate Social Responsibility and Firm Liquidity Risk: U.S. Evidence", *Sustainability*, Vol. 13, No. 22, p. 12894, 2021.

[34] W. Lin, A. Panaretou, G. Pawlina, and C. Shakespeare, "What can we learn about credit risk from debt valuation adjustments?", *Review of Accounting Studies*, Vol. 28, pp. 2556-2588, 2023.

[35] O. Raiter, "Macro-Economic and Bank-Specific Determinants of Credit Risk in Commercial Banks", *Empirical Quests for Management Essences*, Vol. 1, No. 1, pp. 36-50, 2021.

[36] Y. Zhukova, and O. Sobolieva-Tereshchenko, "Modeling macroeconomic indicators in unstable economies", *Journal of International Studies*, Vol. 14, No. 2, pp. 128-148, 2021.

[37] A. Maulana, M. Dwita, M. Fitriyani, D. Sunaryo and Y. Adiyanto, "Risk management as a determinant of indonesian banking financial performance: A systematic literature approach", *Indo-Fintech Intellectuals: Journal of Economics and Business*, Vol. 4, No. 5, pp. 2523-2537, 2024.

[38] J. U. Akpan, I. S. Akinadewo, and Y. A. Satuyi, "International Financial Reporting Standards (IFRS) and Small and Medium-sized Enterprises (SMEs): Assessing the impact of IFRS adoption on SMEs", *World Journal of Finance and Investment Research*, Vol. 7, No. 4, pp. 34-56, 2023.

[39] Y. Salazar, P. Merello, and A. Zorio-Grima, "IFRS 9, banking risk and COVID-19: Evidence from Europe", *Finance Research Letters*, Vol. 56, p. 104130, 2023.

[40] O. Ozgur, E. T. Karagol, and F. C. Ozbugday, "Machine learning approach to drivers of bank lending: evidence from an emerging economy", *Financial Innovation*, Vol. 7, pp. 1-29, 2021.

[41] J. Quast, and M. H. Wolters, "Reliable Real-Time Output Gap Estimates Based on a Modified Hamilton Filter", *Journal of Business & Economic Statistics*, Vol. 40, pp. 152-168, 2022.

[42] Y. Tu, and Y. Wang, "Spurious functional-coefficient regression models and robust inference with marginal integration", *Journal of Econometrics*, Vol. 229, No. 2, pp. 396-421, 2022.

[43] M. I. Ahmed, and S. P. Cassou, "Asymmetries in the effects of unemployment expectation shocks as monetary policy shifts with economic conditions", *Economic Modelling*, Vol. 100, p. 105502, 2021.

[44] T. C. Wilson, "Portfolio Credit Risk I", *Economic Policy Review*, Vol. 4, No. 3, pp. 71-81, 1997.

[45] J. Fan, K. Zhang, Y. Huang, Y. Zhu, and B. Chen, "Parallel spatio-temporal attention-based TCN for multivariate time series prediction", *Neural Computing and Applications*, Vol. 35, No. 18, pp. 13109-13118, 2023.

[46] S. Grieder, and M. D. Steiner, "Algorithmic jingle jungle: A comparison of implementations of principal axis factoring and promax rotation in R and SPSS", *Behavior Research Methods*, Vol. 54, pp. 54-74, 2022.

[47] G. P. Yee, M. S. Rusiman, M. A. Shafi, "K-Means Clustering Analysis and Multiple Linear Regression Model on Household Income in Malaysia", *IAES International Journal of Artificial Intelligence*, Vol. 12, No. 2, pp. 731-738, 2023.

[48] D. G. Breed, J. Hurter, M. Marimo, M. Raletjene, H. Raubenheimer, V. Tomar, and T. Verster, "A Forward-Looking IFRS 9 Methodology, Focussing on the Incorporation of Macroeconomic and Macroprudential Information into Expected Credit Loss Calculation", *Risks*, Vol. 11, No. 3, p. 59, 2023.

[49] W. Li, S. Ding, Y. Chen, and S. Yang, "Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China", *IEEE Access*, Vol. 6, pp. 54396-54406, 2018.