Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2026, 2026:29
https://doi.org/10.28919/cmbn/9724
ISSN: 2052-2541

EVALUATION OF THE PERFORMANCE OF LASSO, GROUP LASSO, AND SPARSE GROUP LASSO IN IDENTIFYING FACTORS THAT HAVE A SIGNIFICANT INFLUENCE ON DENGUE HEMORRHAGIC FEVER IN INDONESIA

C. WIRDIASTUTI*, D. SULISTIOWATI, F.K. MUTYA, R.N. AMALINA, N. YAHPUTRI

Department of Statistics, University Negeri Padang, Padang, Indonesia

Abstract: Dengue Hemorrhagic Fever (DHF) remains a major public health challenge in Indonesia, driven by complex interactions among climatic, demographic, socio-economic, health capacity, and environmental factors. Identifying significant determinants in high-dimensional data with grouped explanatory variables requires appropriate regularization techniques. This study evaluates and compares the performance of LASSO, Group LASSO, and Sparse Group LASSO (SGL) in identifying factors influencing DHF incidence across Indonesian provinces. Three penalized regression models were applied to data comprising 29 predictors organized into five thematic groups. Model performance was assessed using the coefficient of determination (R^2) for data testing, with additional analysis of variable selection patterns to evaluate parsimony and epidemiological interpretability. LASSO was employed to identify dominant individual predictors, Group LASSO to assess the contribution of thematic variable groups, and SGL to simultaneously perform selection at both group and individual levels. The results indicate that SGL provides the best predictive performance, achieving an R^2 of 94.94%, followed by LASSO (88.91%) and Group LASSO (73.94%). LASSO produced the most parsimonious model, selecting only four socio-economic and demographic

*Corresponding author
E-mail address: chairinawirdi@unp.ac.id
Received December 01, 2025

variables, but overlooked meaningful group structures. Group LASSO retained three complete groups (Demographic, Socio-Economic, and Sanitation/Residential Environment) but exhibited reduced accuracy due to the inclusion of weakly contributing variables. In contrast, SGL selected 23 relevant variables while preserving group structures, revealing that DHF incidence is primarily associated with socio-economic factors (unemployment, labor force size, poverty), demographic pressure, sanitation and housing conditions, and rainfall. In conclusion, Sparse Group LASSO emerges as the most effective method for identifying significant DHF determinants, offering an optimal balance between predictive accuracy, model parsimony, and epidemiological interpretability. The findings advocate for the adoption of integrated, data-driven regularization approaches in developing targeted DHF prevention and control strategies in Indonesia.

Keywords: dengue hemorrhagic fever; regularization methods; LASSO; group LASSO; sparse group LASSO.

2020 AMS Subject Classification: 92B20.

1. INTRODUCTION

Dengue Hemorrhagic Fever (DHF) remains a major public health challenge in Indonesia. As a mosquito-borne disease transmitted by *Aedes aegypti*, DHF continues to impose a substantial burden of morbidity and mortality, with incidence rates fluctuating considerably from year to year [1]. Furthermore, data on the case fatality rate (CFR) of DHF (Figure 1) reveal an unstable pattern, highlighted by a sharp increase to 0.96% in 2021. This persistent fluctuation in mortality suggests that existing prevention, early detection, and clinical management strategies have not yet effectively addressed the factors contributing to disease severity [2].

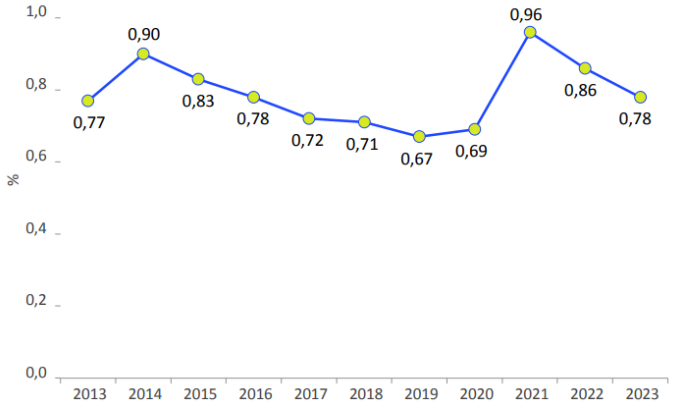

Figure 1. Case fatality rate of DHF

Spatially, the distribution of DHF in Indonesia shows a highly heterogeneous pattern. The distribution map of DHF cases in 2023 (Figure 2) reveals extreme disparities between regions, with case ranges from 48 to 13,844 per region. This variation reflects the complexity of disease determinants, which likely involve interactions between environmental, demographic, behavioral, and local health system capacity factors [3]. Without a comprehensive understanding of the dominant factors in each region, DHF control efforts risk being off-target and less effective.

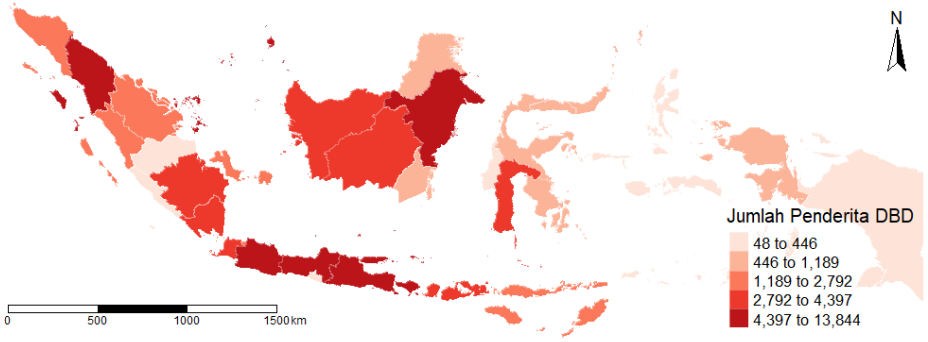

Figure 2. Distribution of DHF in Indonesia in 2023

Previous studies have attempted to identify factors influencing DHF incidence through various approaches. A study by [4] employed negative binomial regression to investigate the impact of climatic factors on DHF incidence in Yogyakarta. Meanwhile, research by [5] applied hierarchical models to analyze dengue risk factors using multiple levels of demographic and clinical determinants. However, most of these studies used conventional regression methods that have limitations in handling data with many correlated predictors.

In the context of risk factor analysis, researchers often encounter situations where the number of potential predictor variables is extremely large. Conventional regression methods, such as multiple linear regression, have limitations in handling high-dimensional data, as they are prone to overfitting and multicollinearity problems [6]. Additionally, these methods are not designed to perform automatic variable selection, so the resulting models may contain predictors that are actually insignificant.

Some recent studies have begun to adopt more advanced methods for DHF risk factor analysis. A study by Lim et al. [7] used LASSO regression within a Bayesian framework to identify climatic

factors influencing dengue transmission in Singapore, demonstrating the method's ability to handle high-dimensional data and produce parsimonious models. However, this study did not consider the group structure of predictor variables, which is often encountered in public health data. On the other hand, research by Ouhourane et al. [8] applied Group Penalized Quantile Regression, a variant of Group LASSO, to analyze risk factors, though not specifically for dengue, the methodological framework offers insights for grouped variable selection in epidemiological studies.

To overcome these limitations, regularization methods such as LASSO (Least Absolute Shrinkage and Selection Operator) have been developed [9]. LASSO is able to perform variable selection by penalizing regression coefficients, so that only the most informative variables are retained in the model. However, when predictor variables have a natural group structure, for example, climate variable groups, socio-economic variable groups, or environmental variable groups, the standard LASSO approach becomes less optimal.

Therefore, the Group LASSO method was developed to perform selection at the group level [10], and the Sparse group LASSO, which combines selection capabilities at both the group level and within-group individual level [11]. These three methods offer different approaches to handling high-dimensional data and group structures, but comparative evaluation of these three methods in the context of DHF risk factor modeling in Indonesia remains very limited.

Based on this background, this study aims to evaluate the performance of three regularization methods, LASSO, Group LASSO, and Sparse group LASSO, in identifying factors that have a significant influence on DHF incidence in Indonesia. Evaluation is conducted based on predictive ability, variable selection accuracy, and stability of the resulting models. This research will also compare findings with previous studies that used conventional methods.

The results of this study are expected to provide dual contributions. Methodologically, this research provides empirical evidence regarding the relative performance of the three regularization methods in the context of real public health data. Practically, the identification of dominant factors resulting from the best model can serve as a basis for formulating more targeted, effective, and

efficient policies and interventions in DHF control in Indonesia.

2. PRELIMINARIES

2.1 LASSO

Tibshirani introduced the least absolute shrinkage and selection operator (LASSO) method in 1996. The coefficient estimator in the LASSO method is obtained by minimizing the following Eq. (1) [9].

$$\hat{\beta} = \arg\ min_{\beta_0,\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

with penalty $\sum_{j=1}^{p}|\beta_j| \leq t$. The value of t is the quantity that controls the shrinkage of the coefficient estimator with $t \geq 0$, p is the number of explanatory variables, and N is the number of observations. The LASSO estimator can also be expressed in terms of a standardized parameter $s = t/t_0$ where $t_0 = \sum_{j=1}^{p}\left|\hat{\beta}_j^{\,0}\right|$ is sum of absolute values of the ordinary least squares (OLS) coefficients $\hat{\beta}_j^0$. If $t < t_0$, the LASSO coefficients are shrunk toward zero, with some becoming exactly zero. If $t \geq t_0$, then the LASSO estimator equals the OLS estimator $\left(\hat{\beta}_{LASSO} = \hat{\beta}_J^{\,0}\right)$. This causes LASSO to form an efficient model by maintaining the variables that affect the model. The LASSO method does not have a closed-form analytical solution. This limitation stems from the nature of its constraint function, which involves an absolute value term that is non-differentiable at zero [12]. Since a closed-form solution for the coefficient estimator does not exist, the estimation procedure relies on quadratic programming techniques [9].

After its first publication in 1996, the LASSO paper did not receive much attention until 2004, when Efron, Hastie, Johnstone, and Tibshirani introduced the Least Angle Regression (LAR) algorithm [13]. A modification of LAR for LASSO yields an efficient algorithm for estimating the LASSO coefficient solution with faster computation than quadratic programming. The original LAR algorithm is as follows [14]. The original LAR algorithm is as follows [15] :

1. Standardize each independent variable to have a median of zero and a variance of one. Initialize the residual as $r = y - \bar{y}$, and set all coefficients $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.
2. Find the independent variable x_j that is most correlated with the current residual r.
3. Increase the value of β_j from zero toward its least squares coefficient $\langle x_j, r \rangle$, until another variable x_k has as much correlation with the current residual as x_j.
4. Update β_j and β_k simultaneously in the direction defined by the joint least squares fit of the current residual on (x_j, x_k), until another variable x_l attains the same correlation with the current residual.
5. Continue this procedure until all $p$independent variables have entered the active set. After the min (*N-1*,*p*) steps, the OLS solution is obtained.

To obtain the LASSO solution, the LAR algorithm is modified at Step 4 as follows:

4a. If any non-zero coefficient reaches zero, remove that variable from the active set and recompute the joint least squares direction using only the remaining active variables.

2.2 Group LASSO

The Group LASSO method extends the standard LASSO framework by applying the variable selection mechanism to predefined groups of explanatory variables. This grouping structure enables the simultaneous identification and retention of variables that share common characteristics. One of the key advantages of Group LASSO is its ability to accommodate groups of varying sizes [16]. The coefficient estimator in Group LASSO is obtained by minimizing a penalized residual sum of squares, as formulated in Equation (2) [10]:

$$\hat{\beta} = \arg min_{\beta} \left\{ \frac{1}{2} \left\| \boldsymbol{Y} - \sum_{j=1}^{k} \boldsymbol{X}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{k} \sqrt{p_j} \left\| \boldsymbol{\beta}_j \right\|_2 \right\}$$

In this formulation, *k* denotes the total number of groups, X_j represents the matrix of explanatory variables for the j^{th} group, β_j is the corresponding vector of regression coefficients, and p_j denotes the number of variables within the j^{th} group. The tuning parameter $\lambda \geq 0$ controls the strength of the penalty. When $\lambda = 0$, the model reduces to the standard OLS formulation. As

λ increases, the estimated coefficients are progressively shrunk toward zero, with all coefficients converging to zero as $\lambda \to \infty$.

2.3 Sparse group LASSO

Sparse group LASSO (SGL) is a further extension of the Group LASSO method, designed to produce sparsity at two simultaneous levels: the group level and the within-group level of individual variables [11]. This method integrates the standard LASSO penalty and the Group LASSO penalty into a unified objective function, allowing both group-wise and individual variable selection.

Similar to Group LASSO, SGL is particularly suitable for modeling data in which predictor variables naturally form groups, such as single-nucleotide polymorphisms grouped by genes or dummy variables representing categorical factors [17]. A key limitation of the Group LASSO lies in its inability to perform variable selection within an active group; once a group is selected, all variables within that group are retained, regardless of their individual relevance. SGL addresses this limitation by introducing an additional sparsity-inducing penalty at the within-group level [11].

The coefficient estimator in SGL is obtained by minimizing the residual sum of squares subject to two combined penalties, as expressed in Equation (3) [11].

$$\hat{\beta} = \arg min_{\beta} \left\{ \frac{1}{2} \left\| \boldsymbol{Y} - \sum_{j=1}^{k} \boldsymbol{X}_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda \left((1-\alpha) \sum_{j=1}^{k} \left\| \boldsymbol{\beta}_j \right\|_1 + \alpha \sum_{j=1}^{k} \sqrt{p_j} \left\| \boldsymbol{\beta}_j \right\|_2 \right) \right\}$$

In this formulation, k denotes the total number of groups, $\boldsymbol{X}_j$ represents the matrix of explanatory variables for the j^{th} group, $\boldsymbol{\beta}_j$ denotes the corresponding vector of regression coefficients, and p_j indicates the number of variables within the j^{th} group. The tuning parameter $\lambda \geq 0$ controls the overall strength of regularization, while $\alpha \in [0,1]$ governs the balance between the LASSO penalty $\left\| \boldsymbol{\beta}_j \right\|_1$ and the Group LASSO penalty $\sqrt{p_j} \left\| \boldsymbol{\beta}_j \right\|_2$.

Through this dual-penalty structure, SGL enables simultaneous group-level sparsity and within-group sparsity. Specifically, the ℓ_2-norm penalty promotes the exclusion of entire groups, whereas the ℓ_1-norm penalty facilitates the elimination of individual variables within selected

groups. Extreme values of α yield familiar special cases: when $\alpha = 1$, SGL reduces to the standard Group LASSO, and when $\alpha = 0$, it becomes equivalent to the conventional LASSO. By appropriately tuning α, typically via cross-validation, SGL offers a flexible framework for modeling high-dimensional data with inherent group structures.

3. RESEARCH METHODOLOGY

3.1 Data Sources

This study employs secondary data from the Badan Pusat Statistik Indonesia and official dengue surveillance reports for the year 2023. The response variable (Y) is the annual number of DHF cases recorded in each Indonesian province. A comprehensive set of 29 predictor variables was compiled from the BPS dataset and theoretically organized into five distinct groups based on their relevance to DHF transmission ecology: Climate, Demographic, Socio-economic, Healthcare Capacity, Sanitation, and Residential Environment.

3.2 Procedure of Analysis

This study begins with a data exploration stage to understand the dataset's characteristics. First, a thematic map is constructed to visualize the spatial distribution of DHF cases across Indonesian provinces, providing an overview of geographic patterns and case intensity. Subsequently, multicollinearity among the explanatory variables is assessed using a Pearson correlation matrix visualized through a heatmap. The heatmap illustrates the magnitude and direction of pairwise linear relationships, with coefficients near ±1 indicating strong multicollinearity. This observed correlation structure motivates the application of regularization methods, LASSO, Group LASSO, and SGL, which are inherently robust to multicollinearity through coefficient shrinkage and structured variable selection.

Following exploration, data preprocessing is performed. Predictor variables are standardized to have a mean of zero and a standard deviation of one. This step ensures that the regularization penalties are applied uniformly across variables, preventing those with larger scales from dominating the model. The dataset is then randomly split into a training set (75% of observations) for model building and a testing set (the remaining 25%) for final, unbiased evaluation.

The core analytical stage involves the implementation and comparison of three regularization techniques. First, the Least Absolute Shrinkage and Selection Operator (LASSO) is applied using the glmnet package in R. The optimal regularization parameter (λ) is determined via 5-fold cross-validation to minimize the prediction error. Variables with non-zero coefficients are retained as significant predictors.

Second, the Group LASSO method is implemented using the gglasso package. This approach extends LASSO by performing selection at the level of pre-defined groups of variables (e.g., all climate variables as one group). The optimal λ is similarly selected through 5-fold cross-validation.

Third, the Sparse group LASSO method is employed, utilizing packages such as sgl or msgl. Sparse group LASSO integrates the penalties of LASSO and Group LASSO, enabling simultaneous sparsity at both the group level (selecting relevant groups) and the individual variable level within selected groups. This allows for a more flexible and parsimonious model structure.

Finally, the performance of the three models is rigorously evaluated and compared. The comparison is based on two primary criteria: (1) predictive performance measured by the coefficient of determination (R^2) calculated on the held-out testing set, which quantifies the proportion of variance in DHF cases explained by each model; and (2) model parsimony, assessed by the number of selected variables and groups. A higher R^2 value indicates better explanatory power, while greater parsimony (fewer selected variables/groups) reflects a more interpretable and generalizable model. The model that achieves the optimal balance between high explanatory power (R^2) and parsimony will be selected as the best-performing model. This model is then interpreted to identify the most influential factors associated with DHF incidence in Indonesia, providing an empirical foundation for targeted, evidence-based policy recommendations.

4. RESULT

4.1 Multicollinearity

As an initial step in the analysis, prior to the implementation of LASSO, Group LASSO, and Sparse Group LASSO, the pairwise Pearson correlation coefficients were computed among all explanatory variables to examine the linear dependence structure of the design matrix and to assess

the presence of multicollinearity. This step is particularly relevant given that the predictors are organized into predefined thematic groups and may exhibit strong correlations both within and across groups. Figure 3 displays the resulting correlation matrix visualized as a heatmap, where color intensity represents the magnitude and sign of the Pearson correlation coefficients.

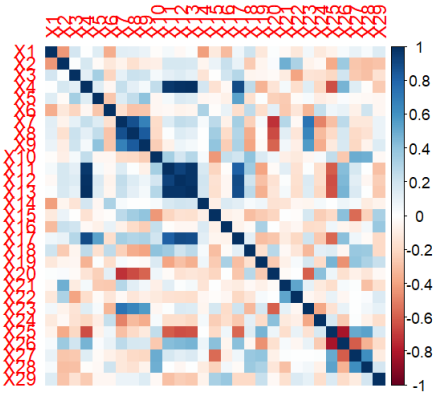

Figure 3. Correlation between explanatory variables

The predominance of high-magnitude correlations in Figure 3 indicates that the design matrix is potentially ill-conditioned, thereby violating the non-multicollinearity assumption of ordinary least squares regression. This empirical evidence provides strong methodological justification for adopting penalization-based regularization methods, such as LASSO, Group LASSO, and Sparse Group LASSO, which address multicollinearity and high dimensionality by imposing sparsity through coefficient shrinkage and structured variable selection.

4.2 LASSO

The LASSO coefficient estimates are computed using an efficient iterative algorithm derived from the Least Angle Regression (LAR) framework, commonly referred to as the LARS algorithm. This approach is computationally more efficient than traditional quadratic programming methods. The estimation procedure begins by initializing all regression coefficients to zero. At each iteration, the predictor exhibiting the highest absolute correlation with the current model residual is identified and its coefficient is updated, while previously selected predictors are adjusted accordingly. The evolution of coefficient estimates across iterations can be visualized through a solution path plot, as illustrated in Figure 4.

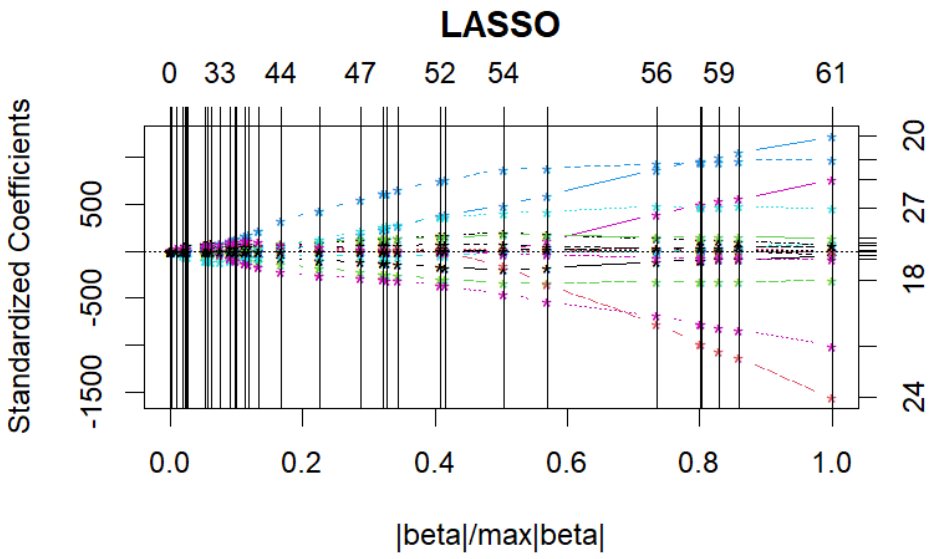

Figure 4. Plot the estimated LASSO regression coefficient with the LARS algorithm at each iteration

The plot of the estimated LASSO regression coefficients with the LARS algorithm represents the LASSO regression coefficients at each step of the LARS algorithm. After all the independent variables are selected, the next step is to choose the best model. In Figure 4, the variable X13 is included in the model in the first iteration, meaning that the variable X13 has the highest correlation with the remainder compared to the other variables. Furthermore, in the second iteration, X12 is entered into the model. In the second iteration, the model has two variables, namely X13 and X12. And so on until the variable X29 enters the model at stage 61.

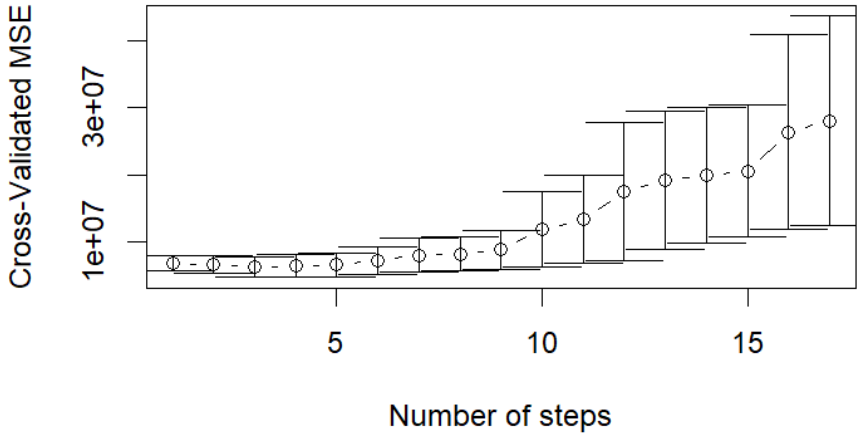

Figure 5. Cross-validation value using step mode

Figure 5 shows Cross-Validates-MSE minimun in the 4th iteration. Therefore, the best

LASSO model selected in this data is the model in the 4th iteration. At this iteration, four variables enter the model, namely X6, X12, X13, and X14. These variables are area, number of unemployed, number of workforce, and average monthly per capita expenditure of the population for health.

4.2 GROUP LASSO

Within the Group LASSO framework, a group of variables is considered influential and retained in the model only when its estimated group coefficient is non-zero. A non-zero group coefficient indicates that the variables within the group jointly contribute to explaining the variation in dengue fever case counts across Indonesia. Accordingly, groups with non-zero coefficients are interpreted as key thematic factors that are significantly associated with incidence of DHF. The selected groups and their relative influences are summarized visually in Figure 6.

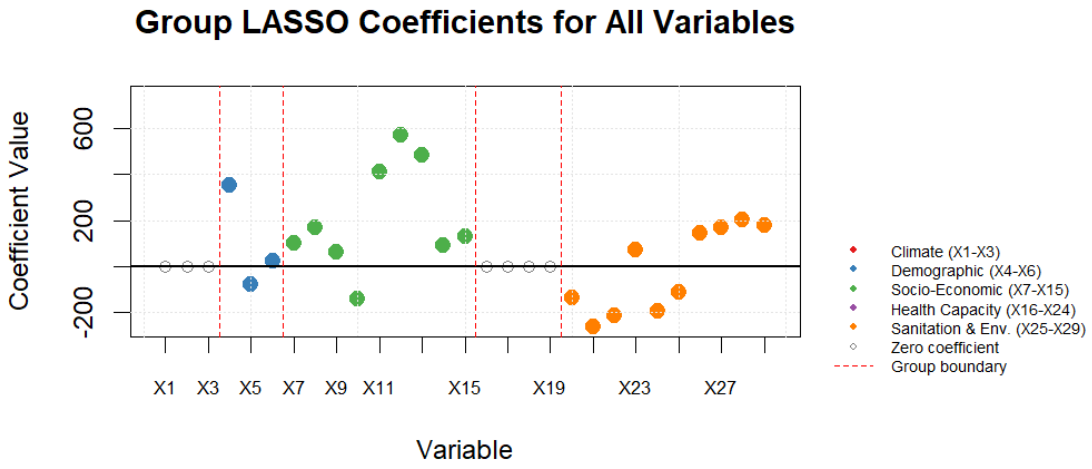

Figure 6. Group LASSO coefficient plots

Figure 6 shows that the variable groups retained by the Group LASSO model are Groups 2, 3, and 5, corresponding to the Demographic, Socio-Economic, and Sanitation and Residential Environment groups, respectively. In the Group LASSO framework, a group is retained when at least one of its constituent variables has a non-zero coefficient, indicating a collective contribution to explaining dengue fever incidence. The Demographic group (X4–X6) includes variables related to population size, population density, and land area, reflecting the influence of human settlement patterns on dengue transmission dynamics. The Socio-Economic group (X7–X14) comprises indicators of educational attainment, poverty, employment, and health expenditure, highlighting

the role of community socio-economic conditions in shaping vulnerability and preventive capacity. The Sanitation and Residential Environment group (X20–X29) consists of variables representing healthcare resources, funding, and accessibility, emphasizing the importance of health system readiness in responding to dengue outbreaks.

In contrast, Group 1 (Climate: X1–X3) and Group 4 (Health Capacity: X15–X19) were not retained in the model, suggesting that, when considered jointly with other grouped factors, their collective contribution was relatively weaker and insufficient for inclusion in the most parsimonious Group LASSO model.

4.3 Sparse Group LASSO

Within the SGL framework, a group is regarded as relevant when at least one variable within the group has a non-zero coefficient, indicating that the group contributes to explaining variations in dengue fever incidence. At the same time, the method enables the identification of specific influential variables within each group, resulting in a more selective and parsimonious model. The groups and variables retained by SGL are presented in Figure 7.

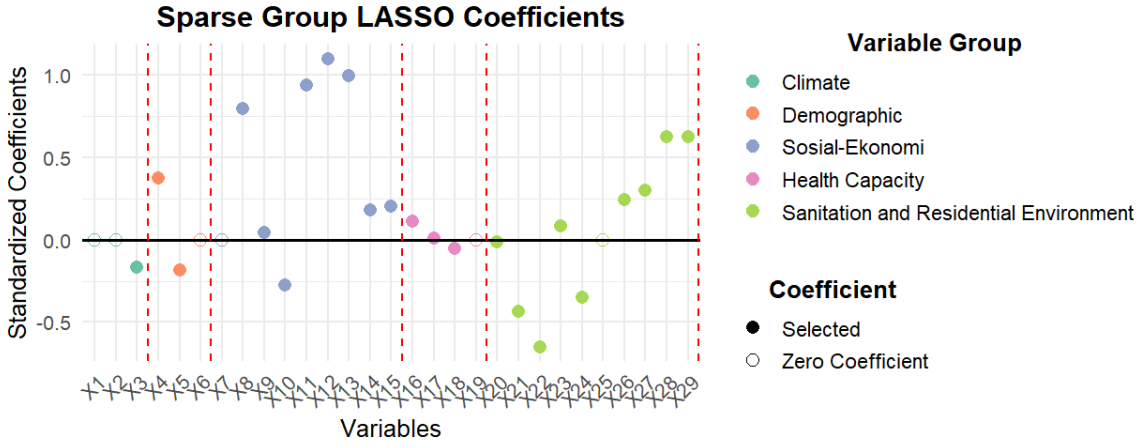

Figure 7. SGL coefficient plots

Figure 7 displays the standardized coefficient estimates from the SGL model used to identify factors influencing the incidence of DHF in Indonesia. Each point represents an explanatory variable, with its position along the horizontal axis indicating the variable index and its vertical position reflecting the magnitude and direction of its effect on the response variable. The horizontal

reference line at zero denotes no effect; variables with coefficients exactly on this line are shrunk to zero and excluded from the model. Point colors correspond to the predefined variable groups, thereby illustrating the contribution of each thematic factor. By simultaneously performing variable selection at both the group and individual levels, SGL retains only relevant groups and key variables within those groups, as evidenced by the presence of non-zero coefficients for selected variables alongside eliminated variables within the same group due to the regularization process.

Based on Figure 7, the SGL estimation indicates that only a subset of variables contributes to dengue hemorrhagic fever (DHF) incidence in Indonesia. Within the climate group, rainfall (X3) is the only variable retained in the model, while average temperature (X1) and average humidity (X2) are eliminated. In the demographic group, total population (X4) and population density (X5) have non-zero coefficients. In the socio-economic group, all variables are retained except primary school completion (X7), with the number of unemployed individuals (X12), labor force size (X13), and number of poor populations (X11) exhibiting the largest coefficient magnitudes. Furthermore, in the health capacity group, three out of four variables, number of hospitals (X16), number of medical personnel (X17), and the percentage of the population covered by health insurance (X18), are retained in the model. Finally, in the sanitation and residential environment group, nine out of ten variables have non-zero coefficients, including indicators related to open defecation practices, protected and unprotected drinking water sources, handwashing and public sanitation facilities, slum housing conditions, access to improved sanitation, access to improved drinking water, and flood occurrence (X20–X25 and X26-X29).

4.4 Performance Evaluation of LASSO, Group LASSO and Sparse group LASSO

The performance of LASSO, Group LASSO, and SGL in this study is evaluated using the coefficient of determination (R^2). The R^2 metric quantifies the proportion of variance in DHF incidence explained by the explanatory variables included in the model. It is employed to compare the predictive performance of the three methods, particularly in the context of high-dimensional data with grouped explanatory variables. A higher R^2 value indicates superior model performance.

The comparative R^2 results for LASSO, Group LASSO, and SGL are presented in Table 1.

Table 1. Measures of the goodness

Analysis Used	R^2	Model Complexity
LASSO	88.91%	4 variables
Group LASSO	73.96%	22 variables
Sparse group LASSO	94.94%	23 variables

Based on Tabel 1, The results indicate that SGL achieves the best performance, with an R^2 value of 94.94%, followed by LASSO (88.91%) and Group LASSO (73.96%). These findings suggest that an approach combining group-level and individual-level variable selection is more effective in capturing the complexity of factors influencing DHF incidence.

From the perspective of variable selection, LASSO produces the most parsimonious model by selecting only four variables, making it effective in identifying dominant factors. However, this approach may overlook relevant group structures and correlations among variables that are epidemiologically meaningful. Group LASSO retains three variable groups comprising 22 variables, but exhibits lower predictive performance, indicating that group-wise selection may include variables with weak contributions and reduce overall model accuracy.

In contrast, SGL selects 23 relevant variables while preserving group structures, thereby providing a more comprehensive representation of DHF risk factors. These results highlight the trade-off between model simplicity and predictive accuracy, with SGL emerging as the most optimal method in this study, as it achieves the highest accuracy without fully compromising epidemiological interpretability. Consequently, this method shows strong potential for supporting data-driven policy formulation and strategic planning in DHF prevention and control.

5. CONCLUSION

This study evaluates the performance of LASSO, Group LASSO, and Sparse Group LASSO (SGL) in identifying factors that significantly influence the incidence of Dengue Hemorrhagic Fever (DHF), based on predictive ability and variable selection characteristics. The results indicate that SGL is the most optimal approach for identifying influential factors of DHF incidence in

Indonesia, achieving the highest predictive performance with an R^2 value of 94.94%, outperforming LASSO (88.91%) and Group LASSO (73.96%). The superiority of SGL lies in its ability to perform variable selection simultaneously at two levels: at the group level, preserving epidemiologically meaningful thematic structures, and at the individual level within groups, ensuring that only truly informative variables are retained. This dual mechanism yields a model that is accurate, parsimonious, and interpretable, consistent with the theoretical advantages of SGL in high-dimensional and highly correlated data settings [11].

From a substantive perspective, the SGL results reveal the multifactorial nature of DHF determinants. Socio-economic factors—particularly unemployment (X12), labor force size (X13), and poverty level (X11)—emerge as the dominant contributors, followed by demographic pressure reflected by population density (X5) and environmental housing conditions captured through sanitation, access to clean water, and housing quality indicators (X20–X29). These findings are consistent with previous epidemiological studies highlighting that socio-economic vulnerability and inadequate living environments substantially increase dengue risk by constraining preventive capacity and access to health services [18, 19, 20]. Climatic factors show a more selective influence, with only rainfall (X3) retained in the final model, supporting prior evidence that climate effects on dengue transmission are often indirect and mediated through socio-environmental pathways [21].

Methodologically, this study highlights a clear trade-off between parsimony and predictive accuracy. LASSO produces the most parsimonious model by selecting only four variables but tends to overlook important group structures and correlated predictors, while Group LASSO retains entire variable groups (22 variables), potentially reducing predictive performance due to the inclusion of weakly contributing variables. SGL effectively bridges these two extremes by selecting 23 key variables within relevant groups, achieving an optimal balance between model simplicity, predictive power, and scientific interpretability, in line with previous methodological findings [11]. From a policy perspective, these results emphasize the need for integrated and evidence-based DHF control strategies that address not only vector and climatic factors but also

underlying socio-economic and environmental determinants. The SGL framework thus provides a robust decision-support tool for prioritizing interventions and can be further extended into a spatio-temporal, prediction-based early warning system to support timely and targeted DHF mitigation efforts.

ACKNOWLEDGEMENTS

The authors would like to thank Lembaga Penelitian dan Pengabdian Masyarakat Universitas Negeri Padang for funding this work with a contract number: 1951/UN35.15/LT/2025.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] Ministry of Health of the Republic of Indonesia, Indonesia Health Profile 2022, Ministry of Health of the Republic of Indonesia, Jakarta, (2023). https://kemkes.go.id/eng/profil-kesehatan-indonesia-2022.

[2] World Health Organization, Dengue and Severe Dengue, (2024). https://www.who.int/health-topics/dengue-and-severe-dengue.

[3] M.G. Guzman, E. Harris, Dengue, Lancet 385 (2015), 453-465. https://doi.org/10.1016/s0140-6736(14)60572-9.

[4] A.L. Ramadona, L. Lazuardi, Y.L. Hii, Å. Holmner, H. Kusnanto, et al., Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data, PLOS ONE 11 (2016), e0152688. https://doi.org/10.1371/journal.pone.0152688.

[5] N.S. da Silva, E.A. Undurraga, E.R. da Silva Ferreira, C.F. Estofolete, M.L. Nogueira, Clinical, Laboratory, and Demographic Determinants of Hospitalization Due to Dengue in 7613 Patients: A Retrospective Study Based on Hierarchical Models, Acta Trop. 177 (2018), 25-31. https://doi.org/10.1016/j.actatropica.2017.09.025.

[6] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer, New York, 2021. https://doi.org/10.1007/978-1-0716-1418-1.

[7] J.T. Lim, B.S. Dickens, S. Haoyang, N.L. Ching, A.R. Cook, Inference on Dengue Epidemics with Bayesian Regime Switching Models, PLOS Comput. Biol. 16 (2020), e1007839. https://doi.org/10.1371/journal.pcbi.1007839.

[8] M. Ouhourane, Y. Yang, A.L. Benedet, K. Oualkacha, Group Penalized Quantile Regression, Stat. Methods Appl. 31 (2021), 495-529. https://doi.org/10.1007/s10260-021-00580-8.

[9] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, J. R. Stat. Soc. Ser. B: Stat. Methodol. 58 (1996), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[10] M. Yuan, Y. Lin, Model Selection and Estimation in Regression with Grouped Variables, J. R. Stat. Soc. Ser. B: Stat. Methodol. 68 (2005), 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x.

[11] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A Sparse-Group Lasso, J. Comput. Graph. Stat. 22 (2013), 231-245. https://doi.org/10.1080/10618600.2012.681250.

[12] M. Robbani, F. Agustiani, N. Herrhyanto, Regresi Least Absolute Shrinkage and Selection Operator (Lasso) Pada Kasus Inflasi di Indonesia Tahun 2014-2017, J. EurekaMatika 7 (2019), 1-16.

[13] A.M. Soleh, Aunuddin, LASSO: Solusi Alternatif Seleksi Peubah Dan Penyusutan Koefisien Model Regresi Linier, Forum Stat. Komput. 18 (2013), 21-27.

[14] C. Wirdiastuti, U.D. Syafitri, I M. Sumertajaya, E. Rohaeti, M. Rafi, Application of Lasso For Identification of Functional Groups with Significant Contributions to Antioxidant Activities of Centella Asiatica, Commun. Math. Biol. Neurosci. 2023 (2023), 15. https://doi.org/10.28919/cmbn/7843.

[15] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer New York, 2008. https://doi.org/10.1007/978-0-387-84858-7.

[16] J. Huang, T. Zhang, The Benefit of Group Sparsity, Ann. Stat. 38 (2010), 1978-2004. https://doi.org/10.1214/09-aos778.

[17] J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, J. Stat. Softw. 33 (2010), 1-22. https://doi.org/10.18637/jss.v033.i01.

[18] G.E. Mendoza, C.J. Moreira, B.R. Sornoza, L. Merchán, M.K. Andrade, Determinantes Sociales de la Salud Asociados Al Dengue en América Del sur: Revisión Sistemática, Rev. Gregor. Cienc. Salud 2 (2025), 138-149. https://doi.org/10.36097/rgcs.v2i1.3146.

[19] P.J. Hotez, M.E. Bottazzi, C. Franco-Paredes, S.K. Ault, M.R. Periago, The Neglected Tropical Diseases of Latin America and the Caribbean: A Review of Disease Burden and Distribution and a Roadmap for Control and Elimination, PLoS Neglected Trop. Dis. 2 (2008), e300. https://doi.org/10.1371/journal.pntd.0000300.

[20] World Health Organization, Global vector control response 2017–2030, (2017). https://www.who.int/publications/i/item/9789241512978.

[21] M.A. Johansson, F. Dominici, G.E. Glass, Local and Global Effects of Climate on Dengue Transmission in Puerto Rico, PLoS Neglected Trop. Dis. 3 (2009), e382. https://doi.org/10.1371/journal.pntd.0000382.