Nusantara Journal of Multidisciplinary Informatics

Volume 1 Issue 1
https://doi.org/10.56705/njomi.v1i1.6

Research Article

Beyond Complete-Data Accuracy: Robustness and Calibration of Diabetes Status Classification Under Feature Unavailability

Rahma Puspitasari[1]
[1] Universitas Negeri Malang, Jl. Cakrawala No.5, Sumbersari, Kec. Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia, rahma.puspitasari.2505348@students.um.ac.id

Correspondence should be addressed to Rahma Puspitasari; rahma.puspitasari.2505348@students.um.ac.id

Received 08 February 2026 Accepted 29 March 2026; Published 30 April 2026

Abstract:
Introduction: Machine learning models developed under complete-data conditions may become unreliable when some predictors are unavailable during deployment. This study evaluates the robustness of survey-based prediabetes/diabetes status classification under random and structured feature unavailability. Method: The CDC Diabetes Health Indicators dataset containing 253,680 observations and 21 predictors was used. Logistic Regression and HistGradientBoosting were evaluated using standard complete-data training and missingness-aware training with simulated feature dropout and missingness indicators. Performance was assessed using AUROC, PR-AUC, Brier Score, Expected Calibration Error, sensitivity, specificity, and decision stability under complete-data, random feature-loss, and structured feature-loss scenarios. Results: Under complete data, standard and missingness-aware HistGradientBoosting achieved similar AUROC values of 0.8298 and 0.8297. At 40% random feature unavailability, the standard model declined to an AUROC of 0.7729 and ECE of 0.0561, while the missingness-aware model maintained an AUROC of 0.7938 and ECE of 0.0064. Sensitivity decreased from 0.8031 to 0.4689 for the standard model, whereas the missingness-aware model retained 0.7498. Cardiometabolic and functional-health feature loss produced greater degradation than lifestyle feature loss. Conclusion: Complete-data performance may overestimate model reliability under incomplete information. Missingness-aware training improved discrimination, calibration, and decision stability without materially reducing performance when all predictors were available.

Keywords: Diabetes Status Classification; Feature Unavailability; Missingness-Aware Learning; Model Robustness; Probability Calibration; BRFSS.

1. Introduction

Diabetes is a chronic disease influenced by various health, behavioral, and demographic factors. Identifying prediabetes and diabetes status using easily accessible indicators can support population-level risk screening. One widely used data source for this purpose is the Behavioral Risk Factor Surveillance System (BRFSS), managed by the Centers for Disease Control and Prevention [1]. Its derived dataset, the CDC Diabetes Health Indicators, contains 253,680 observations and 21 health, behavioral, and sociodemographic indicators [2]. This dataset has been widely used in machine learning studies for diabetes classification, risk factor identification, class imbalance handling, and model interpretability [3]–[6]. Related studies have further demonstrated that classifier performance may vary substantially with ensemble configuration, class-balancing strategy, dataset partitioning, and learning design across heterogeneous health and classification tasks [7]–[11].

Most previous studies evaluate models under the assumption that all predictor variables are available. However, this condition may not always represent real-world deployment, where some information can be unavailable because

of incomplete responses, differences in data collection procedures, or limitations in obtaining specific health indicators. The problem becomes more important when unavailable features follow a particular pattern, known as structured missingness. Previous studies have shown that structured missingness creates additional challenges for machine learning models and that increasing levels of missingness in test data can substantially reduce classifier performance [12], [13].

The impact of feature unavailability may also not be fully reflected by accuracy or AUROC alone. A model may retain acceptable discrimination while producing probabilities that no longer correspond well to the observed outcome distribution. In risk prediction, this problem is closely related to model calibration [14], [15]. Changes in predictor availability between model development and deployment can also affect classification performance and alter individual decisions [16]. Therefore, model robustness should be assessed not only through discrimination but also through calibration and decision stability when the availability of input information changes.

Based on the literature review, previous BRFSS-based diabetes studies have mainly focused on algorithm comparison, feature selection, class balancing, and model interpretability. However, studies that jointly evaluate random and structured feature unavailability, probability calibration, and decision stability using the CDC Diabetes Health Indicators dataset remain limited. This study compares conventional complete-data training with missingness-aware training using Logistic Regression and HistGradientBoosting. The models are evaluated under complete-data conditions and multiple feature-unavailability scenarios using AUROC, PR-AUC, Brier Score, Expected Calibration Error, sensitivity, specificity, and decision stability. Through this design, the study investigates whether models that perform well on complete data can maintain their reliability when some health indicators are unavailable.

2. Method

Research Design and Dataset

This study used an experimental design to evaluate the robustness of diabetes status classification models under partial feature unavailability. The *CDC Diabetes Health Indicators* dataset derived from the 2015 Behavioral Risk Factor Surveillance System was obtained from the UCI Machine Learning Repository. It contains 253,680 observations, 21 predictor variables, and one binary target, `Diabetes_binary`. A value of 0 represents no diabetes, while a value of 1 represents prediabetes or diabetes.

The original dataset contains no missing values. Therefore, missingness in this study was introduced as a controlled feature-unavailability stress test rather than treated as naturally occurring missing data. The dataset was stratified into 60% training, 20% validation, and 20% testing sets using a fixed random seed of 2026, resulting in 152,208 training, 50,736 validation, and 50,736 testing observations.

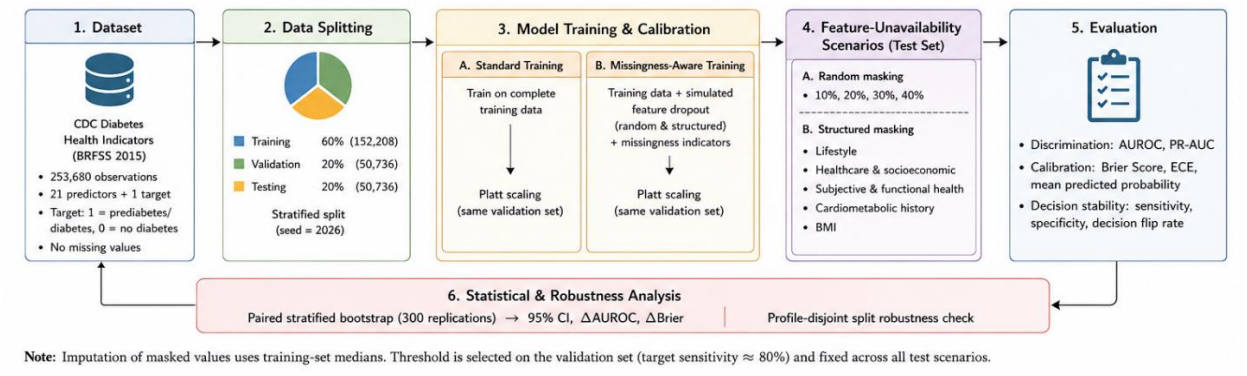

Figure 1. Research workflow for evaluating model robustness under random and structured feature unavailability.

Models and Training Strategies

Two classification algorithms were evaluated: Logistic Regression and HistGradientBoosting. Logistic Regression represented a linear baseline, while HistGradientBoosting was used to capture nonlinear relationships and interactions among health indicators [17].

Two training strategies were compared. The standard model was trained only on complete training data. When a feature became unavailable during testing, its value was replaced using the median calculated from the training set.

The missingness-aware model was trained using the original training data combined with an augmented copy containing simulated feature unavailability. Random masking was generated at 10%, 20%, or 30% feature-dropout levels, while structured masking removed predefined groups of related variables. Missing values were imputed using training-set medians, and 21 binary missingness indicators were added to identify unavailable features. The structured feature groups are presented in **Table 1**.

Table 1. Feature groups used for structured feature unavailability

Feature Group	Variables
Lifestyle	Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump
Healthcare and socioeconomic	AnyHealthcare, NoDocbcCost, Education, Income
Subjective and functional health	GenHlth, MentHlth, PhysHlth, DiffWalk
Cardiometabolic history	HighBP, HighChol, CholCheck, Stroke, HeartDiseaseorAttack
Anthropometric	BMI

Probability Calibration

Model probabilities were calibrated using Platt scaling [18]. To ensure a fair comparison, both standard and missingness-aware models were calibrated using the same complete validation set. Test data were not used during model training, probability calibration, or threshold selection.

Feature-Unavailability Scenarios

Robustness was evaluated under complete-data, random feature-unavailability, and structured feature-unavailability scenarios. Random feature masking was applied independently to 10%, 20%, 30%, and 40% of predictor values in the test set. Structured scenarios removed an entire feature group listed in **Table 1**.

Thus, the main test conditions consisted of complete data, four random masking levels, and five structured feature-loss scenarios. Feature masking was performed only after training and calibration were completed.

Performance Evaluation

Model performance was evaluated using AUROC and Precision-Recall AUC to assess discrimination, particularly considering the class imbalance in the dataset [19], [20]. Probability quality was evaluated using the Brier Score [21]:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{1}$$

Calibration was further assessed using Expected Calibration Error:

$$ECE = \sum_{m=1}^{M}\frac{|B_m|}{N}|\,freq(B_m) - conf(B_m)\}\,| \tag{2}$$

The mean predicted probability was also compared with the actual positive-class prevalence to identify systematic shifts in predicted risk:

$$\tilde{p} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{3}$$

Decision Stability Analysis

A fixed classification threshold was selected from the complete validation set to achieve approximately 80% sensitivity. The threshold was then kept unchanged across all feature-unavailability scenarios. Sensitivity and specificity were calculated as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

Decision stability was quantified using the proportion of predictions that changed relative to the complete-data condition:

$$FlipRate = \frac{1}{N}\sum_{i=1}^{N} 1\left(\hat{y}_{i,s} \neq \hat{y}_{i,complete}\right) \tag{6}$$

A higher flip rate indicates greater instability of individual classifications when input information becomes unavailable.

Statistical and Robustness Analysis

Differences between the standard and missingness-aware strategies were evaluated using 300 paired stratified bootstrap replications. The 95% confidence intervals were obtained using the percentile bootstrap approach [22]. The difference in discrimination was expressed as:

$$\Delta AUROC = AUROC_{aware} - AUROC_{standard} \tag{7}$$

Brier Score improvement was calculated as:

$$\Delta Brier = Brier_{standard} - Brier_{aware} \tag{8}$$

Positive values indicate better robustness of the missingness-aware strategy.

As an additional robustness check, the experiment was repeated using a profile-disjoint split, ensuring that observations with identical combinations of the 21 predictors did not appear across different data subsets. This produced 152,285 training, 50,724 validation, and 50,671 testing observations. The main analysis nevertheless used the stratified 60:20:20 split because it maintained the target distribution more consistently.

Overall, the experimental workflow consisted of dataset preparation, stratified splitting, standard and missingness-aware training, probability calibration, feature-unavailability stress testing, performance and decision-stability evaluation, and statistical robustness analysis.

3. Result and Discussion

Performance Under Random Feature Unavailability

Both training strategies performed almost identically when all predictors were available. Standard HistGradientBoosting achieved an AUROC of 0.8298, while the missingness-aware model achieved 0.8297. Their Brier Scores were also identical at 0.0967, indicating that missingness-aware training did not introduce a meaningful performance penalty under complete-data conditions.

The difference became clearer as feature availability decreased. At 40% random feature unavailability, the standard model declined to an AUROC of 0.7729 and an ECE of 0.0561, whereas the missingness-aware model maintained an AUROC of 0.7938 and an ECE of 0.0064. Similar improvements were observed for PR-AUC and Brier Score.

Table 2. HistGradientBoosting performance under random feature unavailability

Scenario	Model	AUROC	PR-AUC	Brier	ECE
Complete	Standard	**0.8298**	**0.4312**	0.0967	0.0037
Complete	Missingness-aware	0.8297	0.4311	0.0967	**0.003**
Random 20%	Standard	0.804	0.3881	0.1023	0.0295
Random 20%	Missingness-aware	**0.8141**	**0.405**	**0.0993**	**0.0027**
Random 40%	Standard	0.7729	0.3452	0.1099	0.0561
Random 40%	Missingness-aware	**0.7938**	**0.3756**	**0.1024**	**0.0064**

A notable effect was also observed in the average predicted probability. The positive-class prevalence in the test set was approximately 13.93%. Under complete data, the standard model produced a mean probability of 13.92%, but this decreased to only 8.32% at 40% feature unavailability. In contrast, the missingness-aware model maintained a mean probability of 13.44%. This indicates that feature loss caused the standard model to systematically underestimate positive-class probability. Previous studies have similarly shown that missing information may affect both classification performance and calibration [12]–[15].

Structured Feature Unavailability

The impact of missing information varied across feature groups. Removing lifestyle or socioeconomic indicators caused relatively small changes, while the absence of subjective and functional health, cardiometabolic history, or BMI produced greater degradation.

Table 3. Performance under structured feature unavailability

Unavailable Features	Standard AUROC	Aware AUROC	Standard ECE	Aware ECE
Lifestyle	0.8287	0.8284	0.0044	**0.0038**
Healthcare and SES	0.8278	0.8277	0.0068	**0.0029**
Subjective/function	0.8026	**0.8072**	0.0488	**0.0025**
Cardiometabolic history	0.8089	**0.8112**	0.0614	**0.0056**
BMI	0.8118	**0.8124**	0.0215	**0.0017**

Cardiometabolic feature loss produced the largest calibration error in the standard model. Although its AUROC remained relatively high at 0.8089, ECE increased to 0.0614. This result shows that discrimination alone may provide an overly optimistic view of model reliability. Structured missingness should therefore be evaluated separately from random feature loss because the importance of unavailable information differs across feature groups [12].

Recent studies similarly emphasize that reliable model evaluation should consider validation design, data transformation, and predictor structure. Clinical recommendation models evaluated through cross-validation, medical image classifiers subjected to different augmentation strategies, and multivariate models with strongly interdependent predictors have all demonstrated that apparent model performance can depend substantially on how the available information is represented and evaluated [23]–[25].

Decision Stability and Statistical Robustness

Feature unavailability also changed individual classification decisions. Using thresholds fixed from the clean validation set, the standard model achieved a sensitivity of 0.8031 under complete data. At 40% random feature unavailability, sensitivity fell to 0.4689 and the decision flip rate reached 20.12%. The missingness-aware model retained a sensitivity of 0.7498 with a lower flip rate of 13.38%.

Table 4. Decision stability under selected feature-unavailability scenarios

Scenario	Model	Sensitivity	Decision Flip Rate
Complete	Standard	0.8031	–
Complete	Missingness-aware	0.7984	–
Random 40%	Standard	0.4689	20.12%
Random 40%	Missingness-aware	**0.7498**	**13.38%**
Subjective/function unavailable	Standard	0.6124	13.74%
Subjective/function unavailable	Missingness-aware	**0.7734**	**11.01%**
Cardiometabolic unavailable	Standard	0.5411	17.03%
Cardiometabolic unavailable	Missingness-aware	**0.7779**	**10.26%**

Paired stratified bootstrap analysis confirmed that the observed differences were stable. Under complete data, the AUROC difference between the two strategies was negligible at −0.00016 with a 95% CI of −0.00071 to 0.00041. However, at 40% random feature unavailability, the missingness-aware model improved AUROC by 0.02089 with a 95% CI of 0.01833–0.02328. The Brier Score improvement was 0.00751 with a 95% CI of 0.00709–0.00797.

The profile-disjoint robustness check produced a similar pattern. At 40% random feature unavailability, standard HistGradientBoosting achieved an AUROC of 0.7695 and ECE of 0.0582, while the missingness-aware model achieved an AUROC of 0.7905 and ECE of 0.0089. Logistic Regression also showed the same overall trend, indicating that the findings were not specific to a single classifier.

Discussion

The results demonstrate that strong complete-data performance does not necessarily guarantee reliable behavior when some predictors become unavailable. The largest degradation was observed in calibration and sensitivity rather than AUROC alone. This is important because a model may still rank individuals reasonably well while systematically shifting predicted probabilities downward and consequently changing classification decisions.

Missingness-aware training reduced these effects without materially lowering complete-data performance. The benefit was particularly clear under severe random feature loss and when clinically relevant feature groups were unavailable. Therefore, robustness evaluation for survey-based risk models should consider discrimination, calibration, and decision stability together rather than relying solely on conventional classification metrics.

This observation is also relevant to recent medical AI studies. Gender-aware liver disease prediction has demonstrated the importance of evaluating model behavior across population subgroups [26], while health-informatics clustering experiments have shown that incomplete feature information can directly influence the structure and interpretation of machine-learning outputs [27]. These findings further support evaluating data completeness as part of model reliability rather than treating it only as a preprocessing issue.

These findings should not be interpreted as evidence of clinical diagnostic performance. The target represents survey-based prediabetes/diabetes status, and feature unavailability was generated through controlled simulation. Nevertheless, the results provide evidence that incorporating feature-unavailability scenarios during model development can improve reliability when deployment data are incomplete.

4. Conclusion

This study evaluated the robustness of survey-based prediabetes/diabetes classification under random and structured feature unavailability. Under complete-data conditions, standard and missingness-aware HistGradientBoosting models produced nearly identical performance, with AUROC values of 0.8298 and 0.8297, respectively. However, when 40% of features were unavailable, the standard model declined to an AUROC of 0.7729 and an ECE of 0.0561, while the missingness-aware model maintained an AUROC of 0.7938 and an ECE of 0.0064.

The effect of feature loss was particularly evident in calibration and decision stability. Standard-model sensitivity decreased from 0.8031 under complete data to 0.4689 at 40% random feature unavailability, whereas the missingness-aware model retained a sensitivity of 0.7498. Structured experiments also showed that unavailable cardiometabolic and functional-health indicators had a greater impact than missing lifestyle information.

Overall, the findings demonstrate that complete-data performance alone may overestimate model reliability. Missingness-aware training improved robustness to incomplete information without materially reducing performance when all predictors were available. However, the simulated feature-unavailability scenarios do not represent naturally occurring missingness, and the target reflects survey-based diabetes status rather than clinical diagnosis. Future studies should validate this approach using external datasets with real-world missingness patterns.

References

[1] C. M. Pickens, C. Pierannunzi, W. Garvin, and M. Town, "Surveillance for certain health behaviors and conditions among states and selected local areas—Behavioral Risk Factor Surveillance System, United States, 2015," MMWR Surveillance Summaries, vol. 67, no. 9, pp. 1–90, 2018, doi: https://doi.org/10.15585/mmwr.ss6709a1.

[2] Centers for Disease Control and Prevention, "CDC Diabetes Health Indicators," UCI Machine Learning Repository, 2017, doi: https://doi.org/10.24432/C53919.

[3] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building risk prediction models for type 2 diabetes using machine learning techniques," Preventing Chronic Disease, vol. 16, Art. no. E130, 2019, doi: https://doi.org/10.5888/pcd16.190109.

[4] Z. Ullah, F. Saleem, M. Jamjoom, B. Fakieh, F. Kateb, A. M. Ali, and B. Shah, "Detecting high-risk factors and early diagnosis of diabetes using machine learning methods," Computational Intelligence and Neuroscience, vol. 2022, Art. no. 2557795, 2022, doi: https://doi.org/10.1155/2022/2557795.

[5] M. M. Chowdhury, R. S. Ayon, and M. S. Hossain, "An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset," Healthcare Analytics, vol. 5, Art. no. 100297, 2024, doi: https://doi.org/10.1016/j.health.2023.100297.

[6] V. P. Nugraha and A. Febrian, "Comprehensive diabetes risk prediction using BRFSS data: Performance, explainability, fairness, and calibration," Journal of Applied Informatics and Computing, vol. 10, no. 3, pp. 2357–2368, 2026, doi: https://doi.org/10.30871/jaic.v10i3.12740.

[7] H. Azis, M. Abdullah, and S. Ismail, "Performance comparison of various weight and combination function on heterogeneous parallel ensemble machine learning model," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 56, no. 4, pp. 342–353, 2025, doi: https://doi.org/10.37934/araset.56.4.342353.

[8] A. Latip, H. Azis, H. Himawan, and D. A. Kurnia, "SMOTE technique utilization in cirrhosis classification: A comparison of Gradient Boosting and XGBoost," *International Journal of Informatics and Computing*, vol. 1, no. 1, pp. 34–40, 2025.

[9] Purnawansyah, A. P. Wibawa, T. Widiyaningtyas, Haviluddin, H. Darwis, and H. Azis, "An in-depth exploration of supervised and semi-supervised learning on face recognition," *Open Computer Science*, vol. 15, no. 1, pp. 817–840, 2025, doi: https://doi.org/10.1515/comp-2025-0029.

[10] H. Azis, A. Alisma, Purnawansyah, and N. Nirmala, "Analisis kinerja algoritma pembelajaran mesin ensemble pada dataset multi kelas citra JAFFE," *NERO (Networking Engineering Research Operation)*, vol. 9, no. 2, pp. 107–118, 2024, doi: https://doi.org/10.21107/nero.v9i2.27872.

[11] H. Darwis, Z. Ali, Purnawansyah, H. Lahuddin, and H. Azis, "Deep dive into PubMed RCT: Leveraging tribrid embedding recurrent neural network model," *ICIC Express Letters*, vol. 19, no. 1, pp. 111–118, 2025, doi: https://doi.org/10.24507/icicel.19.01.111.

[12] R. Mitra et al., "Learning from data with structured missingness," Nature Machine Intelligence, vol. 5, no. 1, pp. 13–23, 2023, doi: https://doi.org/10.1038/s42256-022-00596-z.

[13] T. Shadbahr et al., "The impact of imputation quality on machine learning classifiers for datasets with missing values," Communications Medicine, vol. 3, Art. no. 139, 2023, doi: https://doi.org/10.1038/s43856-023-00356-z.

[14] Y. E. Shin, M. H. Gail, and R. M. Pfeiffer, "Assessing risk model calibration with missing covariates," Biostatistics, vol. 23, no. 3, pp. 875–890, 2022, doi: https://doi.org/10.1093/biostatistics/kxaa060.

[15] K. Luijken, R. H. H. Groenwold, B. Van Calster, E. W. Steyerberg, and M. van Smeden, "Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective," Statistics in Medicine, vol. 38, no. 18, pp. 3444–3459, 2019, doi: https://doi.org/10.1002/sim.8183.

[16] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," BMC Medicine, vol. 17, Art. no. 230, 2019, doi: https://doi.org/10.1186/s12916-019-1466-7.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: https://doi.org/10.1214/aos/1013203451.

[18] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 2000, pp. 61–74.

[19] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006, doi: https://doi.org/10.1016/j.patrec.2005.10.010.

[20] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: https://doi.org/10.1371/journal.pone.0118432.

[21] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950, doi: https://doi.org/10.1175/1520-0493(1950)078<0001>2.0.CO;2.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proceedings of the 34th International Conference on Machine Learning, vol. 70, 2017, pp. 1321–1330.

[23] M. R. F. Rayhan, Harlinda, H. Darwis, and R. R. Raja, "Comparing ECLAT and Decision Tree for drug therapy recommendation rules on multi label clinical data," *Indonesian Journal of Data and Science*, vol. 7, no. 2, pp. 168–182, 2026, doi: https://doi.org/10.56705/ijodas.v7i2.410.

[24] A. A. P. Agustavada, A. P. Wibawa, D. F. Hilmi, A. Sholum, and F. A. Dwiyanto, "Effect of spatial, intensity, and hybrid augmentation on kidney CT image classification," *Indonesian Journal of Data and Science*, vol. 7, no. 2, pp. 300–317, 2026, doi: https://doi.org/10.56705/ijodas.v7i2.443.

[25] N. F. Azevedo Neto, O. P. Gomes, L. P. Piedade, F. S. Miranda, and R. S. Pessoa, "Interpreting reactive species interdependencies in plasma-activated saline: An exploratory multivariate workflow," *Indonesian Journal of Data and Science*, vol. 7, no. 2, pp. 376–387, 2026, doi: https://doi.org/10.56705/ijodas.v7i2.417.

[26] U. Zaky, M. Habibi, A. Priadana, and T. E. Tarigan, "Gender-aware prediction of liver disease using machine learning and clinical laboratory data," *International Journal of Artificial Intelligence in Medical Issues*, vol. 4, no. 1, pp. 110–130, 2026, doi: https://doi.org/10.56705/wtsdw234.

[27] A. P. Wibowo, M. L. Radhitya, E. Faizal, and I. Arfiani, "Machine learning-based clustering of viruses using taxonomic and genomic features for health informatics applications," *International Journal of Artificial Intelligence in Medical Issues*, vol. 4, no. 1, pp. 72–92, 2026, doi: https://doi.org/10.56705/qstvhw47.

[28] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.

[29] U. Held, A. Kessels, J. Garcia Aymerich, X. Basagaña, G. ter Riet, K. G. M. Moons, and M. A. Puhan, "Methods for handling missing variables in risk prediction models," American Journal of Epidemiology, vol. 184, no. 7, pp. 545–551, 2016, doi: https://doi.org/10.1093/aje/kwv346.

[30] M. I. Gabr, Y. M. Helmy, and D. S. Elzanfaly, "Effect of missing data types and imputation methods on supervised classifiers: An evaluation study," Big Data and Cognitive Computing, vol. 7, no. 1, Art. no. 55, 2023, doi: https://doi.org/10.3390/bdcc7010055.

[31] A. Aich et al., "A copula based supervised filter for feature selection in machine learning driven diabetes risk prediction," Scientific Reports, vol. 16, Art. no. 12132, 2026, doi: https://doi.org/10.1038/s41598-026-41874-9.

[32] M. Talebi Moghaddam et al., "Predicting diabetes in adults: Identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm," BMC Medical Research Methodology, vol. 24, Art. no. 220, 2024, doi: https://doi.org/10.1186/s12874-024-02341-z.