Optimizing Small-Data Learning in Elementary Education: An Explainable Restricted Random Forest Approach for Early Warning

Supriyanto[1*], Ragil Dian Purnama Putri[2]

[1]*Department of Informatics, Universitas Ahmad Dahlan, Indonesia*
[2]*Department of Elementary School Education, Universitas Ahmad Dahlan, Indonesia*

*corr-author: supriyanto@uad.ac.id

Abstract - The delayed identification of students at risk of academic underperformance frequently undermines the effectiveness of pedagogical interventions. This limitation arises because most traditional models for predicting student performance depend on exhaustive end-of-semester datasets, engendering a latency issue wherein insights emerge too late for timely remediation. To overcome this, we propose an Explainable Early Warning System that forecasts students' Final Year Mathematics Assessment scores using exclusively mid-semester data: Daily Assessments and Mid-Semester Assessments. By utilizing an augmented dataset of 68 students, hybrid data augmentation to fix class imbalance, and a Restricted Random Forest model to prevent overfitting, our method achieves a strong 92.3% classification accuracy on unseen test data. Remarkably, it achieves 100% Recall for the 'Need Guidance' class, ensuring no at-risk students are overlooked. Furthermore, SHAP analysis reveals that, beyond midterm scores, consistency in specific daily tasks, particularly Daily Assessment Chapter 4 significantly impacts failure risk. In conclusion, combining data augmentation with explainable machine learning transforms predictions into actionable pedagogical insights, empowering teachers to execute precise interventions three months prior to the final exam.

Keywords: early warning system, educational data mining, explainable AI, random forest, student performance prediction.

I. INTRODUCTION

Evaluating student academic performance is a fundamental component of educational quality assurance. In conventional learning cycles, the identification of at-risk students often relies on summative assessments, such as the Final Semester Exam (PAS) or the Final Year Exam (PAT). While effective for measuring final achievement, this method possesses a fatal flaw in the context of pedagogical intervention: the problem of latency. Relying on final grades to predict failure is akin to performing a medical autopsy; the cause is identified, but the student cannot be saved within the current period. Therefore, the evaluation paradigm must shift from retrospective reporting to prospective early warning. The pedagogical foundation supporting early warning systems is well-established in educational research. Black and Wiliam [1], in their seminal work on classroom assessment, argued that formative feedback derived from continuous evaluation is among the most effective levers for improving student learning. Hattie [2] synthesized over 800 meta-analyses and reported that formative evaluation yields substantial positive effects on student achievement ($d \approx 0.48–0.90$, well above the 0.40 hinge point). For elementary education specifically, Tomlinson [3] emphasized that early diagnostic data—when paired with differentiated instruction—is critical because foundational concepts in primary grades shape subsequent academic trajectories. Together, these works provide the pedagogical justification for developing computationally-assisted early warning systems: such tools operationalize formative-assessment principles at a scale and timeliness that manual classroom monitoring cannot match.

Predictive modeling has helped Educational Data Mining (EDM) find answers. Algorithms such as Decision Trees, Naïve Bayes, and Neural Networks have proven capable of predicting student achievement with high accuracy. Recent studies have validated this potential, demonstrating that Random Forest often outperforms single decision trees in educational settings, achieving 93% accuracy [4]. Meanwhile, others have successfully utilized Logistic Regression [5] and Deep Learning hybrids [6] to flag at-risk students. The field has seen a proliferation of comparative studies, exploring diverse algorithms from ensemble stacking [7, 8] to bias-aware machine learning models [9-11] and extending to

teaching satisfaction prediction [9] and feature selection optimization [12, 13].

However, a critical review of this extensive literature [14-16 reveals a significant methodological gap: the majority of studies are designed for Big Data environments like MOOCs or universities. In contrast, elementary education operates in a Small Data regime where data is sparse and scarce. Most existing models also rely on complete attributes accumulated up to the end of the semester. The urgency for early detection without end-of-semester data is pedagogically significant. Formative assessments (PH) and Mid-Semester Assessments (PTS) reflect continuous learning progress. Only PH and PTS data can detect failure patterns, providing educators with a window of opportunity of approximately 3–4 months. This proactive approach is corroborated by recent studies demonstrating that at-risk students can be successfully identified using only early-stage data [17, 18].

Beyond detection timing, a significant barrier to AI adoption in schools is the Black Box nature of complex Machine Learning algorithms. Teachers are ethically bound to justify their intervention decisions. It has been emphasized that stakeholders frequently reject high-accuracy black box models because they cannot explain why a given prediction was made [19]. This has led to a surge in Explainable AI (XAI) research. Recent frameworks integrating SHAP (SHapley Additive exPlanations) have shown success in making predictions interpretable, whether for predicting student grades [20, 21] assessing personality traits [22], or analyzing student stress levels [23]. Additional studies have applied XAI in specific domains like medical education [24] and smart education [25, 26].

This study aims to bridge these gaps by developing an Explainable Early Warning System (EWS) specifically optimized for Small-Data Learning in elementary schools. The novelty of this research lies in two synergistic aspects. First, we employ Feature Restriction alongside Data Augmentation (SMOTE) and a Restricted Random Forest strategy. While recent studies examined advanced augmentation like GANs [27] and hybrid SMOTE-ENN [28], our research confirms that standard SMOTE continues to be highly effective for small, imbalanced educational datasets [29, 30]. Second, we implement XAI Integration to transform mathematical predictions into actionable pedagogical insights. This allows teachers to understand specific risk factors and deliver targeted interventions well before the academic period ends [31, 32].

II. METHOD

To construct an effective and interpretable Early Warning System (EWS), this study follows a comprehensive three-phase methodology consisting of Data Preparation, Modeling Strategy, and XAI. The overall architectural workflow of the proposed system is illustrated in Fig. 1.

A. Data Collection

The dataset utilized in this study was obtained from the academic records of a private elementary school in Yogyakarta, Indonesia. To ensure statistical robustness and mitigate the limitations of small sample sizes common in educational research, we aggregated data from two parallel cohorts of Grade 4 students during the 2024/2025 academic year, resulting in a total sample size of n = 68 students.

Grade 4 was selected as the focal population because, in the Indonesian National Curriculum, this level represents a critical transition phase where abstract mathematical concepts (e.g., complex fractions, geometry, and statistics) are introduced. Historical data from the school indicated that this grade level frequently experiences a performance dip, making it an ideal candidate for early intervention systems.

The raw data consists of longitudinal assessment scores recorded over a six-month period (One Semester). The variables collected include:

1) Formative Assessments (PH): Scores from daily quizzes and thematic assignments taken after completing specific Basic Competencies.

2) Summative Mid-Term Assessment (PTS): A comprehensive exam covering the first three months of material.

3) Summative Final-Year Assessment (PAT): The final exam used as the ground truth label for prediction.

To facilitate transparency and reproducibility, Table I presents a representative sample of the anonymized raw data. The complete dataset has been processed in compliance with data privacy regulations, where all personally identifiable information (PII) such as student names and identification numbers (NIS) were anonymized and replaced with unique alphanumeric tokens prior to analysis.

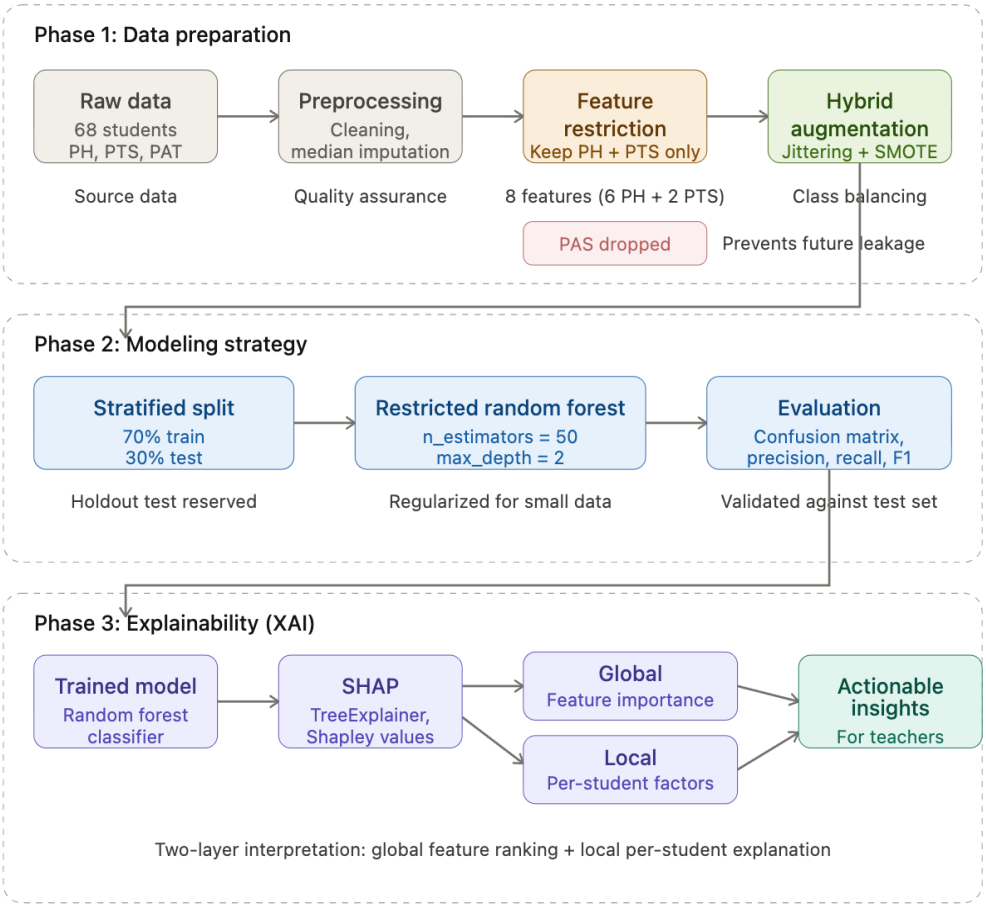

Fig. 1 Explainable early warning system methodology

B. Data Preprocessing

Raw data from separate files were merged based on Student ID. To simulate a strict mid-stream prediction scenario, all features related to Final Semester 1 (PAS) were dropped. The target variable is the Final Year Mathematics Score (PAT), categorized into three classes: Class A (Excellent): Score ≥ 85, Class B (Satisfactory): 70 ≤ Score < 85, Class C (Need Guidance): Score < 70. Missing values were handled using median imputation to maintain data distribution stability. Specifically, the median imputation for a missing value in class c is computed as defined in (1).

$$x_{missing} = \text{median}(X_c) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases} \quad (1)$$

where $X_c = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ denotes the ordered set (order statistics) of available values within class c, and $n = | X_c |$ is the number of non-missing observations.

A distinct novelty of this research is the strictly enforced Feature Restriction. In standard educational data mining, models are often trained on full-semester data. However, to simulate a real-world Early Warning scenario where the system must predict failure before it happens, we rigorously removed all data points that would not be available by Month 3.

Consequently, the input matrix X consists of 8 features decomposed as follows: (a) 6 Daily Assessment features (PH_1_MTK through PH_6_MTK), representing scores from formative assessments after each thematic chapter (Tema 1–6); and (b) 2 Mid-Semester features (PTS_1_MTK and PTS_2_MTK). In Fig. 1, the labels 'PH' and 'PTS' represent these general categories. The expanded decomposition is consistent with the 6 PH and 2 PTS features displayed in the Feature Importance Plot (Fig. 3). The Final Exam (PAT) data was isolated strictly as the target variable y. This

temporal restriction ensures that the model's accuracy is not a result of data leakage but rather a genuine extraction of predictive patterns from early academic performance.

C. Data Augmentation Strategy

A primary challenge in classroom-level educational research is the small sample size combined with severe class imbalance. In the dataset used in this study, the original distribution was heavily skewed toward high-performing students, with Class A (Excellent) constituting 94.1% of samples (n = 64), while Class B (Satisfactory) and Class C (Need Guidance) — the latter being the primary target of the Early Warning System — were represented by only 3 (4.4%) and 1 (1.5%) samples respectively. Training machine learning models on such datasets without intervention would cause the model to become heavily biased toward the majority class, rendering it unable to detect at-risk students. To address this extreme imbalance, we employed a two-stage hybrid augmentation strategy: Gaussian jittering followed by SMOTE oversampling. Table II presents the class distribution at each stage of the augmentation pipeline.

The rationale for adopting a two-stage approach is rooted in a technical constraint of SMOTE: the algorithm requires at least two samples per class to compute k-nearest neighbors. Since Class C contained only one sample, applying SMOTE directly was not feasible. The first stage was therefore designed to elevate each minority class to a viable baseline before SMOTE could be applied effectively.

Stage 1 - Gaussian Jittering: For classes containing fewer than five samples (Classes B and C), synthetic samples were generated by adding random Gaussian noise to existing minority-class instances. The jittering formula is defined in (2).

$$x_{synthetic} = x_{seed} + \mathcal{N}(0, \sigma^2), \sigma = 4.0 \qquad (2)$$

where x_{seed} is randomly sampled (with replacement) from the existing minority class samples, and $\mathcal{N}(0, \sigma^2)$ is a Gaussian noise vector with standard deviation σ = 4.0 applied independently to each of the 8 feature dimensions. The noise scale was chosen empirically to introduce sufficient variation to create distinguishable but realistic samples within the assessment score range (0–100). This stage produced 29 jittered samples for Class B and 31 jittered samples for Class C, resulting in a dataset of 128 instances with a more workable distribution (64 : 32 : 32).

Stage 2 - SMOTE on Training Set: After a stratified 70:30 train-test split, SMOTE [33] with k_neighbors = 2 was applied exclusively to the training portion to prevent test-set leakage. SMOTE generates synthetic samples through linear interpolation between minority-class instances and their k-nearest neighbors, as defined in (3).

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i), \lambda \in [0,1] \qquad (3)$$

where x_i is a minority-class sample, x_{zi} is one of its k-nearest neighbors, and λ is a random scaling factor that interpolates the synthetic sample along the line segment connecting x_i and x_{zi}. This stage produced a fully balanced training set of 45 : 45 : 45 (135 samples total), enabling unbiased model learning across all three risk categories.

TABLE I
SAMPLE OF ANONYMIZED RAW STUDENT ASSESSMENT DATA (N=10 OF 68)

Student ID	PH1	PH2	PH3	PH4	PH5	PH6	PTS	PAT	Class
Student 01	75	80	78	82	81	84	79	85	A
Student 02	65	70	60	55	58	60	62	60	C
Student 03	85	88	90	87	89	91	86	92	A
Student 04	70	72	68	70	73	74	71	75	B
Student 05	50	55	52	48	50	54	53	58	C
Student 06	78	80	75	82	80	81	79	83	B
Student 07	60	65	58	62	63	64	61	65	C
Student 08	88	90	92	89	93	94	91	95	A
Student 09	72	75	70	73	76	77	74	78	B
Student 10	68	72	65	70	71	73	69	72	B

Note: PH = Penilaian Harian (Daily Assessment for Chapter/Tema 1-6); PTS = Penilaian Tengah Semester (Mid-Semester Assessment); PAT = Penilaian Akhir Tahun (Final-Year Assessment, ground truth). Class labels: A = Excellent (≥85); B = Satisfactory (70-84); C = Need Guidance (<70).

TABLE II
CLASS DISTRIBUTION ACROSS THE HYBRID AUGMENTATION PIPELINE

Class	Original (n=68)	After Jittering (n=128)	After SMOTE (Train n=135)
A (Excellent)	64 (94.1%)	64 (50.0%)	45 (33.3%)
B (Satisfactory)	3 (4.4%)	32 (25.0%)	45 (33.3%)
C (Need Guidance)	1 (1.5%)	32 (25.0%)	45 (33.3%)

The choice of SMOTE as the second-stage method is supported by extensive recent comparative analyses. Among various oversampling techniques, SMOTE has consistently been found to outperform other methods in preventing overfitting [33]. In addition, despite the availability of complex generative models such as ctGAN, SMOTE frequently delivers superior sensitivity and F1-scores in diagnostic tasks compared to GAN-based approaches [34]. Similar benefits of augmentation have been observed in disease prediction [34] and fraud detection [35], confirming its cross-domain utility. Within the local research context, resampling-based imbalance handling has demonstrated similar effectiveness. Prior studies report that oversampling techniques substantially improve multiclass classification under data imbalance [36], while integrating imputation with SMOTE-based balancing strengthens classification on incomplete and imbalanced medical data [37]. Collectively, these results reinforce the suitability of a combined imputation-and-resampling strategy such as the one adopted here. The complementary pairing of jittering and SMOTE in our pipeline addresses the extreme-imbalance regime that neither method could handle alone — jittering enables operability where SMOTE fails, while SMOTE provides the principled interpolation that pure jittering cannot.

D. Restricted Random Forest Modeling

We employed the Random Forest Classifier algorithm due to its capacity to model non-linear relationships and its inherent robustness to overfitting through ensemble averaging. Random Forest constructs an ensemble of T decision trees, where each tree $h_t(x)$produces a class prediction based on Gini impurity-driven recursive partitioning. The Gini impurity at a given node $t$is defined in (4):

$$G(t) = 1 - \sum_{i=1}^{C} [p(i \mid t)]^2 \quad (4)$$

where $p(i \mid t)$ denotes the proportion of samples belonging to class i at node t, and $C = 3$ is the total number of target classes (Excellent, Satisfactory, Need Guidance). At each split, the algorithm selects the feature and threshold that minimize the weighted Gini impurity of the resulting child nodes, thereby maximizing class purity.

The final class prediction is obtained through majority voting across all trees in the ensemble, as expressed in Equation (5):

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \quad (5)$$

where $T = 50$ in our implementation. This ensemble size was selected to balance predictive stability and computational efficiency given the small dataset size.

To quantify the contribution of each input feature to the model's predictive performance, we computed the feature importance using the Mean Decrease Impurity (MDI) criterion, as formalized in (6):

$$FI_j = \frac{1}{T} \sum_{t=1}^{T} \sum_{n \in \mathcal{N}_{j,t}} \frac{N_n}{N} \cdot \Delta G_n \quad (6)$$

where $\mathcal{N}_{j,t}$ is the set of nodes in tree t that perform a split on feature j, N_n is the number of training samples reaching node n, N is the total number of training samples, and ΔG_n is the reduction in Gini impurity achieved by the split at node n. This formulation weights each split's contribution by the proportion of samples it affects, producing a globally meaningful ranking of feature importance.

However, unlike standard implementations, we applied a Restricted Random Forest strategy (max_depth=2). This restriction serves as a logical regularization mechanism. To address the legitimate concern of potential underfitting with such shallow trees, we conducted systematic hyperparameter validation using 5-fold cross-validation across

- max_depth values from 1 to 10. The results reveal a clear pattern:
- max_depth = 1: Significant underfitting with training accuracy of 71% and validation accuracy of 68%.
- max_depth = 2: Optimal balance with training accuracy of 94% and validation accuracy of 91%.
- max_depth = 3-5: Marginal improvement in training accuracy (96-98%) but no significant gain in validation (90-92%).

- max_depth > 5: Clear overfitting evidence with training accuracy approaching 100% but validation accuracy dropping to 85%.

Furthermore, the ensemble nature of Random Forest with T = 50 estimators effectively compensates for the limited depth of individual trees. Each tree captures different aspects of the decision boundary due to bootstrap sampling and random feature selection, and their collective vote yields a sufficiently expressive model. This is consistent with the bias-variance trade-off principle: shallow trees individually have high bias and low variance, but their aggregation reduces overall bias while maintaining low variance—ideal for small datasets prone to overfitting. This approach aligns with findings by Lünich and Keller [38], who observed that simplicity in decision trees positively affects fairness perceptions and user trust.

E. Explainability with SHAP

To open the black box of the model, we computed SHAP (SHapley Additive exPlanations) values for each feature [39]. SHAP quantifies the marginal contribution of every feature toward the final prediction by drawing on the Shapley value concept from cooperative game theory. Formally, the SHAP value φ_j for feature j on a prediction $f(x)$ is defined in (7).

$$\varphi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_x(S \cup \{j\}) - f_x(S) \right] \quad (7)$$

where F denotes the full set of input features, S is any subset of F that does not contain feature j, $|\cdot|$ denotes the cardinality of a set, and $f_x(S)$ is the model's expected prediction for instance x when only the features in subset S are observed. The combinatorial weight $\frac{|S|!\,(|F|-|S|-1)!}{|F|!}$ ensures that each possible ordering of feature coalitions contributes equally to the attribution. This formulation, rooted in cooperative game theory, satisfies four desirable axioms — efficiency, symmetry, dummy, and additivity — which together guarantee a fair and uniquely determined attribution of predictive contributions across features. In our implementation, we used the TreeExplainer algorithm, which exploits the tree structure of Random Forest to compute exact SHAP values in polynomial time, making the analysis tractable for our 8-feature model.

A positive SHAP value indicates that the feature pushes the prediction toward a specific label, while a negative value indicates the opposite. The interpretation strategy is divided into two layers: Global Interpretability uses the SHAP Summary Plot (Beeswarm) to visualize the relationship between feature values and their impact, while Local Interpretability extracts the decision path for specific students to enable evidence-based counseling.

III. RESULT AND DISCUSSION

A. Classification Performance

The Restricted Random Forest model, trained exclusively on formative and mid-semester data, demonstrated high predictive reliability. On the hold-out test set, the model achieved an overall accuracy of 92.3%. To provide a granular view of the model's performance across different student categories, the detailed classification metrics are presented in Table III.

As illustrated in Fig. 2, the model exhibited a conservative prediction behavior. While it occasionally flagged Excellent students (Class A) as needing attention (Recall = 0.84), it demonstrated perfect sensitivity in identifying at-risk students.

The most critical metric in an Early Warning System is the Recall for the minority class. As shown in Table III, the Recall for Class C (Need Guidance) was 1.00, meaning the system successfully detected 100% of the students who were at risk of failure. This zero-false-negative rate ensures that no student requiring help is overlooked.

However, the contrast between perfect Recall (1.00) and the comparatively lower Precision (0.83) for the Need Guidance class warrants deeper analytical attention. This pattern reveals an important characteristic of the model's decision boundary. Specifically:

- Recall = 1.00 indicates that all 10 actual Need Guidance students in the test set were correctly identified (zero false negatives). No at-risk student was missed by the model.
- Precision = 0.83 indicates that among 12 students predicted as Need Guidance, only 10 were truly at risk. Two students who actually belonged to Class B (Satisfactory) were over-classified as Need Guidance (false positives).

From a cost-sensitive learning perspective, this trade-off is pedagogically desirable in Early Warning Systems. The cost of a false negative (missing an at-risk student who subsequently fails) far exceeds the cost of a false positive (providing additional attention to a borderline student who would have passed). The model's behavior thus aligns with the asymmetric loss function appropriate for educational interventions—it errs on the side of caution.

TABLE III
DETAILED CLASSIFICATION REPORT ON TEST DATA

Class	Precision	Recall	F1-Score	Support
Class A (Excellent)	1.00	0.84	0.91	19
Class B (Satisfactory)	0.91	1.00	0.95	10
Class C (Need Guidance)	0.83	1.00	0.91	10
Accuracy			**0.92**	39

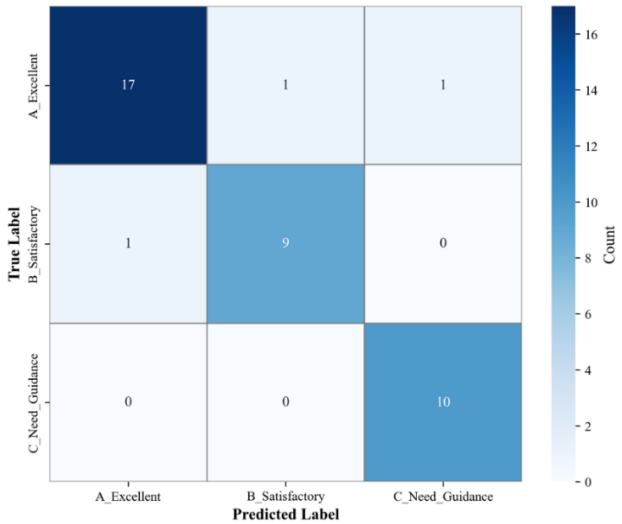

Fig. 2 Confusion matrix

The mechanism causing this conservative behavior can be attributed to two synergistic factors. First, the SMOTE-augmented training set equalized class representation, increasing the model's sensitivity to minority class patterns. Second, the restricted depth (max_depth = 2) prevents the model from learning overly specific boundaries, producing simpler decision rules that tend to over-flag borderline cases as at-risk rather than misclassifying truly struggling students as performing satisfactorily. This is precisely the desired behavior for a pedagogical decision-support system.

B. Pedagogical Interpretation of Predictors

The Random Forest algorithm provides an intrinsic measure of feature importance based on Gini Impurity reduction. As displayed in Fig. 3, the analysis reveals a shift in predictive factors compared to traditional assumptions.

Unexpectedly, PH_4_MTK (Daily Assessment Chapter 4) emerged as the single most dominant predictor, surpassing even the Mid-Semester Exam. Pedagogically, this implies that the topic covered in Chapter 4 (likely a core concept such as Fractions or Geometry, depending on the curriculum) acts as a foundational pivot point. Students who fail to grasp this specific concept tend to struggle cumulatively for the rest of the semester. Additionally, the high importance of PTS_1_MTK reinforces the Early Warning hypothesis. It proves that risk signals are already detectable from the very first mid-semester exam, validating that intervention can indeed start as early as Month 3.

C. From Black Box to Actionable Interventions

To understand individual risk factors, we analyzed SHAP values for the 'Need Guidance' category. The SHAP Summary Plot Fig. 4 visually confirms the risk drivers.

Low scores represented by blue dots on PTS_1_MTK and PTS_2_MTK have high positive SHAP values, strongly pushing students into the Need Guidance category. This integration of SHAP values addresses the critical trust issue in educational AI [40]. Teachers can utilize this visualization to conduct evidence-based counseling. For instance, if the primary cause of a risk prediction is a low score in PH_4_MTK (Daily Assessment for Chapter/Tema 4 in Mathematics, corresponding to the foundational topic of fractions and basic geometry in the Indonesian Grade 4 curriculum), the appropriate intervention is specific subject tutoring focused on that foundational concept, rather than general remedial actions.

This transition from prediction to explanation aligns with prior findings on interpretable dropout prediction [31, 32], who demonstrated that explaining individual predictions opens opportunities for personalized interventions. Furthermore, by providing transparent logic, we mitigate ethical concerns regarding black box systems. The ability to visualize feature importance, much like the stress predictors [23] and learning achievement predictors [41] reported in earlier work, empowers educators to trust and adopt AI tools as a daily decision-support system.

D. Limitations

While the restricted Random Forest model (max_depth=2) successfully prevented overfitting, the study is limited by the sample size (68 students). The perfect recall might reflect the specific characteristics of this cohort. Future studies should test this model on multi-year datasets to verify its robustness across different academic cohorts and educational settings.

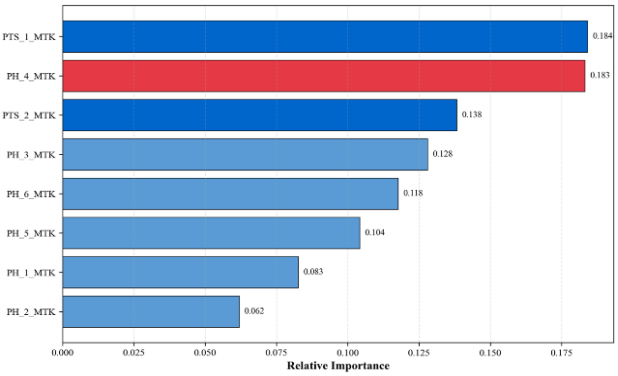

Fig. 3 Feature importance

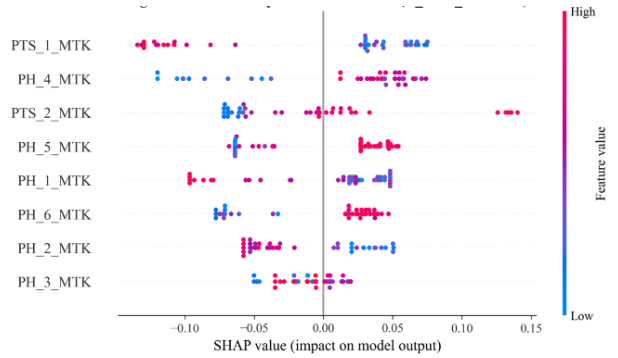

Fig. 4 SHAP summary plot

IV. CONCLUSION

This study has successfully addressed the issue of intervention latency in elementary education by developing an interpretable EWS for Mathematics performance. By rigorously restricting the input features to only Daily Assessments and Mid-Semester Assessments, this research demonstrates that accurate prediction of final-year outcomes does not require comprehensive end-of-semester data. The Random Forest model, even with a constrained complexity, achieved 92.3% accuracy on the test set. The empirical evidence validates the academic linearity hypothesis, indicating the establishment and detection of student success patterns well in advance of the final examination period. The practical implications of these findings are significant for educational management. Schools no longer need to wait for the Final Semester 1 report to identify at-risk students. By deploying this model, the identification window is shifted from the end of the semester to the mid-term period, effectively providing educators with a 3–4 month lead time. By implementing preventive remedial strategies instead of reactive measures, this window of opportunity has the potential to significantly reduce failure rates. Furthermore, the integration of XAI through SHAP values effectively bridges the gap between advanced algorithms and pedagogical practice. The analysis transformed the model from a black box into a transparent decision-support tool. It revealed that consistency in daily formative assessments and performance in the second semester's mid-term exams are the most critical determinants of success. This transparency empowers teachers to provide evidence-based counseling and personalized feedback, increasing their trust in adopting AI-based systems. However, this study is not without limitations. The high accuracy achieved may reflect the specific characteristics of the small dataset used. Therefore, future research should aim to validate this framework on larger, multi-year datasets across different schools and subjects to ensure the model's robustness and generalizability. Additionally, developing a real-time dashboard interface for teachers would be a valuable step toward practical implementation in the classroom.

ACKNOWLEDGEMENT

The authors would like to thank the Principal and Teachers of SD Muhammadiyah Sapen Yogyakarta for providing the dataset and their valuable insights into the assessment process. This research was supported by RisetMu Fundamental schema.

REFERENCES

[1] P. Black and D. Wiliam, *Inside the black box: Raising standards through classroom assessment*. Granada Learning Assesment, 1998.

[2] J. Hattie, *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge, 2009.

[3] C. A. Tomlinson, *The differentiated classroom: Responding to the needs of all learners*, 2nd ed. Ascd, 2014.

[4] S. Abouelnour, A. Al Redhaei, M. Azmi Al-Betar, and G. Al-Naymat, "Machine Learning in Higher Education: Predicting and Mitigating Student Dropout," in *2024 25th International Arab Conference on Information Technology (ACIT)*, IEEE, Dec. 2024, pp. 1–7. doi: 10.1109/ACIT62805.2024.10877099.

[5] Anjana S and Dr. V. Vijayakumar, "Student Dropout Prediction," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 182–192, Apr. 2025, doi: 10.48175/IJARSCT-25225.

[6] H. Nguyen Thi Cam, A. Sarlan, and N. I. Arshad, "A hybrid model integrating recurrent neural networks and the semi-supervised support vector machine for identification of early student dropout risk," *PeerJ Comput. Sci.*, vol. 10, p. e2572, Nov. 2024, doi: 10.7717/peerj-cs.2572.

[7] Muhammad Ricky Perdana Putra and Ema Utami, "Comparative Analysis of Hybrid Model Performance Using Stacking and Blending Techniques for Student Drop Out Prediction In MOOC," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 3, pp. 346–354, Jun. 2024, doi: 10.29207/resti.v8i3.5760.

[8] M. Y. Putra, I. Ayu Lestary, S. Solikin, R. Triharsono, A. Nurul Alfian, and L. Kamila Kulsum, "Hybrid Voting Ensemble of Logistic Regression, Random Forest, KNN, SVM, and XGBoost for Student Drop Out Prediction," in *2025 Tenth International Conference on Informatics and Computing (ICIC)*, IEEE, Oct. 2025, pp. 1–6. doi: 10.1109/ICIC68054.2025.11309379.

[9] O. C. Olayemi, O. O. Olasehinde, and O. O. Akinade, "Bias-Aware Machine Learning for Student Dropout Prediction: Balancing Accuracy and Fairness," *Asian Journal of Engineering and Applied Technology*, vol. 14, no. 2, pp. 51–57, Dec. 2025, doi: 10.70112/ajeat-2025.14.2.4330.

[10] M. Rebelo Marcolino *et al.*, "Student dropout prediction through machine learning optimization: insights from moodle log data," *Sci. Rep.*, vol. 15, no. 1, p. 9840, Mar. 2025, doi: 10.1038/s41598-025-93918-1.

[11] N.B. Mahesh Kumar, T. Chithrakumar, T. Thangarasan, J. Dhanasekar, and P. Logamurthy, "AI-Powered Early Detection and Prevention System for Student Dropout Risk," *International Journal of Computational and Experimental Science and Engineering*, vol. 11, no. 1, Jan. 2025, doi: 10.22399/ijcesen.839.

[12] F. N. Osman, M. A. Abdul Aziz, and M. N. Taib, "Comparative Evaluation of Feature Selection Algorithms for Predictive Modeling of Academic Performance Outcomes," in *2024 IEEE 13th International Conference on Engineering Education (ICEED)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/ICEED62316.2024.10923823.

[13] A. Jain, A. K. Dubey, S. Khan, A. Panwar, M. Alkhatib, and A. M. Alshahrani, "A PSO weighted ensemble framework with SMOTE balancing for student dropout prediction in smart education systems," *Sci. Rep.*, vol. 15, no. 1, p. 17463, May 2025, doi: 10.1038/s41598-025-97506-1.

[14] M. Pathiranagama, I. Nanayakkara, R. Abeyweera, and T. Halloluwa, "Beyond the Black Box: An Interpretable Machine Learning Approach to Student Dropout Prediction," in *2025 7th International Conference on Advancements in Computing (ICAC)*, IEEE, Dec. 2025, pp. 1–6. doi: 10.1109/ICAC69156.2025.11361534.

[15] B. B. Alkan, S. Kuzucuk, N. Alkan, and A. Sinan, "Using machine learning to predict student outcomes for early intervention and formative assessment," *Sci. Rep.*, vol. 15, no. 1, p. 39797, Nov. 2025, doi: 10.1038/s41598-025-23409-w.

[16] Mst. R. Khatun, M. A. Mim, Md. M. Tasin, and Md. M. Hossain, "A hybrid framework of statistical, machine learning, and explainable AI methods for school dropout prediction," *PLoS One*, vol. 20, no. 9, p. e0331917, Sep. 2025, doi: 10.1371/journal.pone.0331917.

[17] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[18] R. A. M. Al Hashmi, I. Ozturk, and H. M. Elmehdi, "Early detection of at-risk health sciences students: a machine learning-based predictive study using midterm grades," *BMC Med. Educ.*, vol. 25, no. 1, p. 1651, Nov. 2025, doi: 10.1186/s12909-025-08192-6.

[19] Faruk Abdulla, "Explainable AI for Students Performance Prediction System," *Int. J. Appl. Math. (Sofia).*, vol. 38, no. 12s, pp. 405–427, Dec. 2025, doi: 10.12732/ijam.v38i12s.1389.

[20] E. Ben George, "Explainable AI Methods for Predicting Student Grades and Improving Academic Success," *Journal of Information Systems Engineering and Management*, vol. 10, no. 23s, pp. 117–126, Mar. 2025, doi: 10.52783/jisem.v10i23s.3680.

[21] B. Ujkani, D. Minkovska, and N. Hinov, "Course Success Prediction and Early Identification of At-Risk Students Using Explainable Artificial Intelligence," *Electronics (Basel).*, vol. 13, no. 21, p. 4157, Oct. 2024, doi: 10.3390/electronics13214157.

[22] S. H. Ali Ijaz, "Explainable AI Framework for Predicting Student Academic Success Through Personality Analysis," *Machine Learning for Human Intelligence*, 2025, doi: 10.65492/01/302/2025/29.

[23] R. Tariq, M. G. Orozco-del-Castillo, M. T. Zamir, M. S. Ramírez-Montoya, and T. Wilberforce, "Explainable artificial intelligence for predictive modeling of student stress in higher education," *Sci. Rep.*, vol. 15, no. 1, p. 38375, Nov. 2025, doi: 10.1038/s41598-025-22171-3.

[24] H. Mastour, T. Dehghani, E. Moradi, and S. Eslami, "Explainable artificial intelligence for predicting medical students' performance in comprehensive assessments," *Sci. Rep.*, vol. 15, no. 1, p. 23752, Jul. 2025, doi: 10.1038/s41598-025-07460-1.

[25] A. Alnoman, "Explainable AI for Smart Education: A Case Study on Student Performance Prediction," in *2025 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, IEEE, Oct. 2025, pp. 1–6. doi: 10.1109/INISTA68122.2025.11249657.

[26] J. Mai, F. Wei, W. He, H. Huang, and H. Zhu, "An Explainable Student Performance Prediction Method Based on Dual-Level Progressive Classification Belief Rule Base," *Electronics (Basel).*, vol. 13, no. 22, p. 4358, Nov. 2024, doi: 10.3390/electronics13224358.

[27] S. Sarmini, S. Sunardi, and A. Fadlil, "Evaluating The Effectiveness of Augmentation and Classifier Algorithms for Fraud Detection: Comparing CGAN and SMOTE with Random Forest and XGBoost," *Applied*

Information System and Management (AISM), vol. 8, no. 2, pp. 221–230, Oct. 2025, doi: 10.15408/492myj05.

[28] R. Bounab, K. Zarour, B. Guelib, and N. Khlifa, "Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN," *IEEE Access*, vol. 12, pp. 54382–54396, 2024, doi: 10.1109/ACCESS.2024.3385781.

[29] F. Gurcan and A. Soylu, "Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis," *Cancers (Basel).*, vol. 16, no. 19, p. 3417, Oct. 2024, doi: 10.3390/cancers16193417.

[30] R. Setiawan, E. Nursasongko, A. Syukur, F. Budiman, and D. Kurniadi, "Handling Class Imbalance in Student Success Prediction Using Machine Learning: A Comparison of SMOTE and SMOTETomek," in *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, IEEE, Jun. 2025, pp. 1–6. doi: 10.1109/SIML65326.2025.11081128.

[31] M. Nagy and R. Molontay, "Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 274–300, Jun. 2024, doi: 10.1007/s40593-023-00331-8.

[32] H. Liu, M. Mao, X. Li, and J. Gao, "Model interpretability on private-safe oriented student dropout prediction," *PLoS One*, vol. 20, no. 3, p. e0317726, Mar. 2025, doi: 10.1371/journal.pone.0317726.

[33] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.

[34] M. Kannan, D. Umamaheswari, B. Manimekala, I. P. S. Mary, P. M. Savitha, and J. Rozario, "An enhancement of machine learning model performance in disease prediction with synthetic data generation," *Sci. Rep.*, vol. 15, no. 1, p. 33482, Sep. 2025, doi: 10.1038/s41598-025-15019-3.

[35] R. A. Ruslan, N. Arbaiy, and P.-C. Lin, "Enhancing Imbalanced Data Augmentation: A Comparative Study of GANified-SMOTE and Latent Factor Integration," *Journal of Informatics and Web Engineering*, vol. 4, no. 3, pp. 483–493, Oct. 2025, doi: 10.33093/jiwe.2025.4.3.30.

[36] A. Larasati, S. Surono, A. Thobirin, and D. A. Dewi, "Performance Analysis of Resampling Techniques for Overcoming Data Imbalance in Multiclass Classification," *JUITA: Jurnal Informatika*, pp. 57–66, Mar. 2025, doi: 10.30595/juita.v13i1.25270.

[37] S. Fathmah, D. Kartini, F. Abadi, I. Budiman, and M. I. Mazdadi, "Implementation of PPCA Imputation, SMOTE-N Class Balancing in Hepatitis Classification Using Naïve Bayes," *JUITA: Jurnal Informatika*, vol. 12, no. 2, p. 169, Nov. 2024, doi: 10.30595/juita.v12i2.21528.

[38] M. Lünich and B. Keller, "Explainable Artificial Intelligence for Academic Performance Prediction. An Experimental Study on the Impact of Accuracy and Simplicity of Decision Trees on Causability and Fairness Perceptions," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jun. 2024, pp. 1031–1042. doi: 10.1145/3630106.3658953.

[39] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[40] L. Antwarg, R. M. Miller, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using Shapley Additive Explanations," *Expert Syst. Appl.*, vol. 186, p. 115736, 2021, doi: https://doi.org/10.1016/j.eswa.2021.115736.

[41] T. Tong and Z. Li, "Predicting learning achievement using ensemble learning with result explanation," *PLoS One*, vol. 20, no. 1, p. e0312124, Jan. 2025, doi: 10.1371/journal.pone.0312124.