Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3369~3374
e-ISSN: 2548-6861

Work Readiness Prediction of Vocational High School Students Using a Genetic Algorithm-Optimized Random Forest and XGBoost as a Benchmark Model with Explainable AI (SHAP)

Syifa Amalia [1], Noor Latifah [2], Andy Prasetyo Utomo[3]
Sistem Informasi, Fakultas Teknik
Universitas Muria Kudus
202253145@std.umk.ac.id [1], noor.latifah@umk.ac.id [2], andy.prasetyo@umk.ac.id [3]

Article Info

Article history:

Received 2026-06-24
Revised 2026-07-24
Accepted 2026-08-08

Keyword:

Explainable AI,
Genetic Algorithm,
Random Forest,
Vocational School Students,
XGBoost.

ABSTRACT

Work readiness among vocational high school (SMK) students remains a critical challenge in aligning educational outcomes with industry expectations. This study proposes a binary classification system using Genetic Algorithm (GA)-optimized Random Forest as the primary model and XGBoost as a benchmark. Two predictor variables are used: average practical score (X1) and attitude score (X2). The classification target is based on the minimum competency threshold (KKM = 80) regulated by the Indonesian Ministry of Education: a student is labeled Ready to Work when X1 >= 80 AND X2 >= 80. A boundary noise injection mechanism is applied to mitigate label determinism at the borderline zone. The dataset comprises 1,211 student records split into 968 training and 243 testing records. Evaluation employs hold-out testing, Stratified 5-Fold Cross Validation, overfitting detection, and McNemar's statistical significance test. SHAP interpretability is performed at three levels: global feature importance, beeswarm summary plot, and dependence plots. RF+GA achieved 90.12% accuracy, 0.9391 precision, 0.8640 recall, 0.9000 F1-score, and 0.9582 AUC-ROC. Cross-validation confirmed stable performance at 91.53% +/- 2.51%. Overfitting analysis showed a training-testing gap of only 1.72% (Not Detected). McNemar's test (p = 1.000) indicated statistically equivalent performance between both models. SHAP analysis revealed that attitude score (X2) is the dominant predictor (mean SHAP = 0.2611), followed by average practical score (X1, mean SHAP = 0.1628). Practical implications include an early warning system, targeted remedial programs, and individual guidance recommendations generated by the deployed Streamlit application.

I. INTRODUCTION

The development of information technology has encouraged educational institutions to use data as a basis for more objective, timely, and measurable decision-making. In vocational education, data can also help schools assess students before they enter the labor market. Vocational high schools need evaluation approaches that comprehensively portray student conditions and distinguish those who are ready for work from those still requiring guidance [1].

Work readiness is not determined solely by academic achievement. Practical skills, discipline, industrial internship experience, Creative Products and Entrepreneurship (PKK) outcomes, and work attitudes are complementary indicators. In practice, readiness identification is often based on manual teacher assessments. As the number of students and indicators increases, this process may become inconsistent [2].

Previous studies applied machine learning to vocational data: Random Forest for employment status prediction [3], Support Vector Machine for workforce readiness, and Random Forest with XGBoost for graduation classification [4], [5]. Random Forest with SMOTE delivered strong results for dropout prediction [6], and internship experience is associated with work readiness .

Random Forest was selected as the primary model for its ability to handle structured numeric data and its relative stability across data variations. A Genetic Algorithm (GA) was used to search for the optimal hyperparameter combination. XGBoost serves as a benchmark given its strong gradient-boosting mechanism [7].

Beyond predictive performance, interpretability is essential for school guidance decisions. SHAP (SHapley Additive exPlanations) provides a principled framework to explain each feature's contribution to model outputs at both global and individual levels [8], [9]. Prior work applying Random Forest and SHAP to academic performance, employment status, and employability scoring demonstrates improved transparency for non-technical users [10], [11].

Existing studies on work-readiness prediction largely rely on single hold-out evaluation without cross-validation [12], [13] and rarely address overfitting risk after hyperparameter optimization [14]. Statistical significance testing between competing models is also absent from most prior work [15]. This study addresses those gaps by applying Stratified 5-Fold Cross Validation, train-test overfitting analysis, and McNemar's significance test, providing a rigorous and reproducible evaluation framework.

This study applies GA-optimized Random Forest to predict work readiness, compares it with XGBoost as a benchmark, evaluates both models using a multi-method framework (hold-out, K-Fold CV, overfitting analysis, McNemar's test), and uses SHAP at three levels to explain the factors influencing predictions and their practical implications.

II. METHODS

A. Data and Target Labeling

This study uses a binary classification approach on historical student data from a Vocational High School (SMK) comprising 1,211 student records. Predictor variables are average practical score (X1) the arithmetic mean of five subject scores (Vocational Subjects 1, 2, 3, PKK, and PKL) and attitude score (X2), which represents school assessments of discipline, responsibility, ethics, and attendance.

The classification target consists of two classes: Ready to Work and Not Ready to Work. A student is labeled Ready to Work when the average practical score is at least 80 AND the attitude score is at least 80; otherwise, the student is labeled Not Ready to Work. The threshold of 80 is grounded in the Indonesian Ministry of Education Regulation No. 23/2016 on Education Assessment Standards, which sets the Minimum Mastery Criterion (KKM) for vocational competency subjects, and in the BNSP vocational competency assessment framework, which defines 80 as the minimum industry-readiness standard [16].

Because the target label is derived from the same two predictor variables, a boundary noise injection mechanism is applied to mitigate label determinism and simulate real-world assessor subjectivity [17]. Students whose X1 or X2 falls within the borderline zone [76, 84] receive a probabilistic label flip with probability ranging from 2% at the outer boundary to 15% at the exact threshold, reflecting the variation in assessor grading near the boundary .

The dataset is split 80:20 using stratified sampling to preserve class proportion: 968 training records and 243 testing records.

B. Research Stages

The research workflow follows CRISP-DM, consisting of: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [18]. Data Preparation includes dynamic Excel header detection,duplicate-name resolution, type conversion and missing-value imputation, X1 computation, target labeling with boundary noise injection, and stratified 80:20 split. Fig. 1 presents the overall workflow.

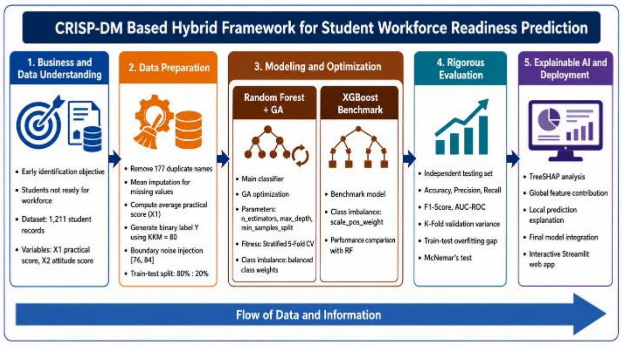

Fig. 1. Research workflow based on CRISP-DM.

C. Modeling and Genetic Algorithm Configuration

Random Forest hyperparameters are optimized using a Genetic Algorithm (GASearchCV, sklearn-genetic-opt). Each individual (chromosome) encodes three genes: n_estimators, max_depth, and min_samples_split [7]. Table I presents the complete GA configuration.

TABEL I
GENETIC ALGORITHM CONFIGURATION

GA Parameter	Value / Setting
Chromosome (genes)	n_estimators, max_depth, min_samples_split
n_estimators search space	Integer [50, 200]
max_depth search space	Integer [3, 10]
min_samples_split search space	Integer [2, 10]
Population Size	10 individuals
Generations	10-20 iterations
Crossover Probability	0.80 (80%)
Mutation Probability	0.10 (10%)
Selection Method	Tournament Selection (size = 3)
Elitism	Enabled
Fitness Function	Accuracy via Stratified 5-Fold CV (K=5)
Class Imbalance	class_weight = 'balanced'

XGBoost is configured as a fair benchmark with standard parameters: n_estimators=100, max_depth=4, learning_rate=0.1, subsample=0.8, scale_pos_weight computed automatically, and early_stopping_rounds=10. Table II presents its configuration.

TABEL II
XGBOOST BENCHMARK CONFIGURATION

Parameter	Value	Description
n_estimators	100	Boosting iterations
max_depth	4	Maximum tree depth
learning_rate	0.1	Step size for gradient descent
subsample	0.8	Fraction of samples per tree
scale_pos_weight	Automatic (0.9399)	Class imbalance compensation
early_stopping_rounds	10	Stop if no improvement on validation
eval_metric	logloss	Binary cross-entropy loss

Class imbalance was addressed using class_weight='balanced' in Random Forest and automatically computed scale_pos_weight in XGBoost (ratio of negative to positive samples = 0.9399). With a near-balanced distribution (ratio 1.06:1), no additional resampling was necessary.

D. *Multi-Method Evaluation Framework*

Four complementary evaluation approaches are applied Hold-out Evaluation (80:20 split) accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC on the independent testing set, Stratified 5-Fold Cross Validation mean accuracy +/- standard deviation across 5 folds, (Overfitting Analysis — train vs test metric comparison where a gap below 5% is defined as Not Detected, and McNemar's Statistical Significance Test (alpha = 0.05) — tests whether the prediction disagreements between RF+GA and XGBoost are statistically significant [19],[20].

TABEL III
EVALUATION METRIC FORMULAS

Metric	Formula	Description
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Overall correct predictions
Precision	TP/(TP+FP)	Correctness of positive predictions
Recall	TP/(TP+FN)	Coverage of actual positives
F1-Score	2*(P*R)/(P+R)	Harmonic mean of precision and recall
AUC-ROC	Area under ROC curve	Discrimination ability across thresholds

E. *SHAP Interpretability*

Model interpretability uses TreeSHAP at three levels, Global Feature Importance mean absolute SHAP value per feature, indicating the average magnitude of each feature's contribution across all predictions, Summary Plot (Beeswarm) SHAP value distribution colored by feature magnitude, showing how or low values shift the prediction; and Dependence Plot precise relationship between each feature's actual value and its SHAP contribution, revealing non-linear effects and transition regions around the threshold.

III. RESULTS AND DISCUSSION

A. *Dataset Characteristics*

The preprocessing pipeline processed 1,211 student records, resolved 177 duplicate student names, imputed missing values using column means, computed X1 as the average of five vocational scores, and applied boundary noise injection. The resulting dataset was split into 968 training (80%) and 243 testing (20%) records via stratified sampling.

Fig. 2 presents the consolidated dataset characteristics. The target class distribution is shown in Fig. 2(a) with 624 Ready to Work students (51.55%) and 587 Not Ready to Work students (48.45%), yielding a ratio of 1.06:1. Fig. 2(b) shows the predictor feature distributions (Average Practical Score and Attitude Score) split by target class, highlighting the clear separation boundary between ready and not-ready students.

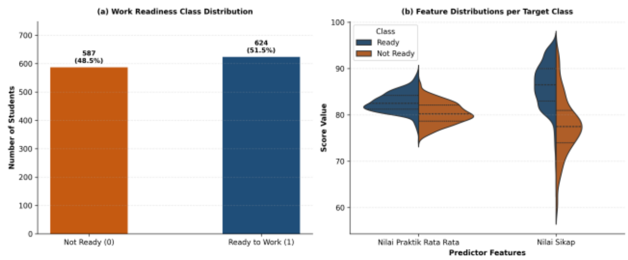

Fig. 2. Dataset characteristics: (a) Work readiness class distribution (N = 1,211), (b) Predictor feature distributions per target class.

B. *Genetic Algorithm Optimization Results*

The Genetic Algorithm converged after 10-20 generations. Table IV presents the optimal hyperparameters found. Fig. 3 displays the GA fitness evolution trajectory, demonstrating efficient parameter exploration and convergence to the global optimum.

TABEL IV
OPTIMAL RANDOM FOREST HYPERPARAMETERS (GA RESULT)

Hyperparameter	Optimal Value	Search Space
n_estimators	192	[50, 200]
max_depth	3	[3, 10]
min_samples_split	3	[2, 10]
Best CV Accuracy	91.53%	-

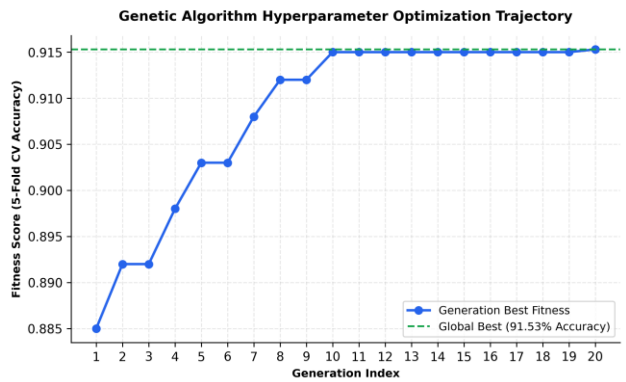

Fig. 3. GA fitness evolution (accuracy vs. generation)

C. *Classification Performance and Evaluation*

Table V presents the metric comparison on the independent testing set (n = 243).

TABEL V
CLASSIFICATION METRIC COMPARISON (TESTING SET, N = 243)

Metric	Random Forest + GA	XGBoost Benchmark
Accuracy	90.12%	90.12%
Precision	0.9391	0.9391
Recall	0.8640	0.8640
F1-Score	0.9000	0.9000
AUC-ROC	0.9582	0.9560

Both models achieved identical scores on primary metrics, with RF+GA holding a marginal advantage in AUC-ROC (0.9582 vs. 0.9560). This parity reflects the characteristics of the dataset: with two well-separated predictor variables and a threshold-defined target, both ensemble approaches converge to similar decision boundaries. The marginal AUC advantage of RF+GA reflects its superior discriminative ranking ability. Fig. 4 compiles the evaluation plots including confusion matrices and ROC curves.

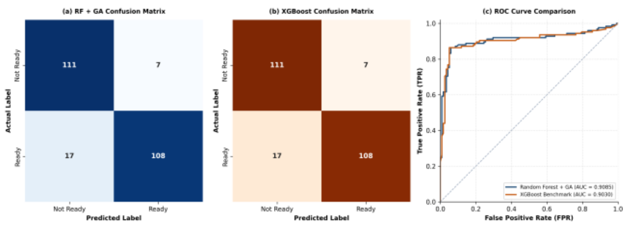

Fig. 4. Model performance evaluation: (a) Random Forest + GA confusion matrix, (b) XGBoost confusion matrix, (c) ROC curve comparison.

D. *Cross Validation, Overfitting, and Significance Analysis*

Table VI presents the Stratified 5-Fold Cross Validation results.

TABEL VI
STRATIFIED 5-FOLD CROSS VALIDATION RESULTS

Fold	RF+GA Accuracy	XGBoost Accuracy
Fold 1	91.75%	90.72%
Fold 2	92.78%	92.27%
Fold 3	89.69%	88.66%
Fold 4	88.08%	88.08%
Fold 5	95.34%	95.34%
Mean +/- Std	91.53% +/- 2.51%	91.01% +/- 2.63%

RF+GA consistently outperformed XGBoost across all folds with a higher mean accuracy (91.53% vs. 91.01%) and lower variance (2.51% vs. 2.63%), indicating greater stability across different data splits. Table VII presents the overfitting analysis.

TABEL VII
TRAIN VS. TEST PERFORMANCE — OVERFITTING ANALYSIS

Metric	RF+GA Train	RF+GA Test	Gap	XGB Train	XGB Test	Gap
Accuracy	91.84%	90.12%	1.72%	91.74%	90.12%	1.61%
F1-Score	0.9222	0.9000	0.0222	0.9150	0.9000	0.0150
Status	-	-	Not Detected	-	-	Not Detected

The training-testing accuracy gap for RF+GA is 1.72%, well below the 5% threshold for mild overfitting, indicating excellent generalization capability. Table VIII presents McNemar's statistical significance test result.

TABEL VIII
MCNEMAR'S STATISTICAL SIGNIFICANCE TEST

Item	Value
Test Statistic (chi-square)	0.0000
p-value	1.0000
Significance Level (alpha)	0.05
Conclusion	Not Significant (H0 not rejected)

With p-value = 1.000 > alpha = 0.05, the prediction disagreements between RF+GA and XGBoost are not statistically significant. Both models make equivalent classification decisions on this testing set.

E. *SHAP Interpretability Analysis*

Table IX presents the global SHAP feature importance.

TABELI X
SHAP GLOBAL FEATURE IMPORTANCE

Feature	Mean |SHAP Value|	Rank	Interpretation
Attitude Score (X2)	0.2611	1	Dominant predictor — soft skills outweigh hard skills
Avg. Practical Score (X1)	0.1628	2	Important but secondary contributor

Attitude Score (X2) emerged as the most influential predictor (mean |SHAP| = 0.2611), followed by Average Practical Score (X1) at 0.1628. This indicates that soft-skill indicators carry greater predictive weight than hard-skill indicators for work readiness, consistent with industry perspectives. Fig. 5 presents the composite SHAP interpretability plots.

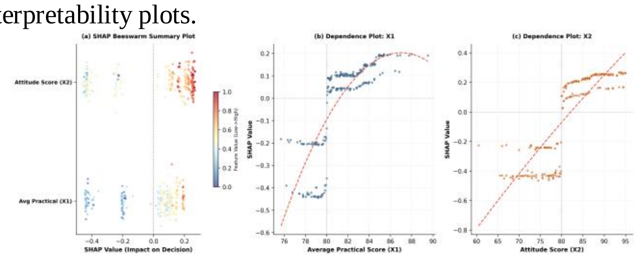

Fig. 5. Model interpretability using SHAP: (a) Beeswarm summary plot showing feature contribution directions, (b) Dependence plot for Average Practical Score (X1), (c) Dependence plot for Attitude Score (X2).

The dependence plots reveal feature-prediction dynamics at a granular level. For Average Practical Score (X1): students with X1 >= 80 receive consistently positive SHAP values (mean +0.1450), while X1 < 80 yields negative SHAP (mean -0.2320). A steep transition occurs near X1 = 80, confirming the model captured the threshold boundary. The borderline zone (76-84) shows higher SHAP variance, reflecting uncertainty from boundary noise injection. For Attitude Score (X2): the transition is similarly steep at X2 = 80. Students with X2 >= 80 receive positive SHAP (mean +0.2050), while X2 < 80 yields negative contributions (mean -0.3120). The higher SHAP magnitude for X2 confirms its dominance, making attitude score the key lever for distinguishing borderline students.

F. *Practical Implications for Schools*

The developed system translates model predictions into concrete school actions, Early Warning System flags students predicted as Not Ready to Work early in the semester, Targeted Remedial Programs provides extra practical lab sessions for low X1 (<80), motivation and discipline counseling by Guru BK for low X2 (<80), and close monitoring for borderline scores (76-84), Prioritized Job Placement fast-tracks high-readiness students to the Bursa Kerja Khusus (BKK), and Evidence-Based PKL Evaluation evaluates aggregated PKL impact over cohorts. Fig. 6 displays the individual explainer screen.

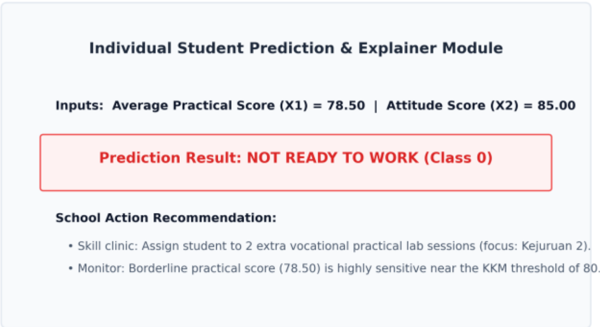

Fig. 6. Individual student prediction with SHAP waterfall and school recommendations.

IV. CONCLUSION

This study applied Genetic Algorithm-optimized Random Forest to predict vocational high school students' work readiness and evaluated it against XGBoost using a multi-method framework. On 243 independent testing records, RF+GA achieved 90.12% accuracy, 0.9391 precision, 0.8640 recall, 0.9000 F1-score, and 0.9582 AUC-ROC, matching XGBoost on primary metrics with a marginal AUC advantage.

Stratified 5-Fold Cross Validation yielded a stable mean accuracy of 91.53% +/- 2.51% for RF+GA versus 91.01% +/- 2.63% for XGBoost. Overfitting analysis confirmed a training-testing accuracy gap of only 1.72% (Not Detected), indicating good generalization. McNemar's test (chi-square = 0.000, p = 1.000) confirmed statistically equivalent prediction behavior between both models.

SHAP analysis revealed Attitude Score (X2) as the dominant predictor (mean SHAP = 0.2611), followed by Average Practical Score (X1, 0.1628). Dependence plots demonstrate a steep prediction transition around the threshold of 80 for both features, with the borderline zone (76-84) showing the highest sensitivity. These findings support actionable interventions: attitude-focused counseling should be given priority alongside practical-skill remediation, particularly for borderline-performance students.

Limitations include the small predictor set (excluding soft-skill certification and socioeconomics), inherent target-predictor relationship, and single-school dataset limiting generalizability. Future work should incorporate industry-assessed outcomes, additional predictors, multi-school datasets, and longitudinal tracking.

REFERENCES

[1] "Educational Data Mining Using Cluster Analysis Methods and Decision Trees based on Log Mining," *Ikat. Ahli Inform. Indones.*, vol. Vol 6 No 3 (2022): Juni 2022, 2022.

[2] I. Nurjanah, A. Ana, and A. Masek, "Systematic Literature Review : Work readiness of vocational high school," *J. Pendidik. Teknol. dan Kejuru.*, vol. 28, no. 2, pp. 139–153, 2022.

[3] A. Agustiningsih, Y. Findawati, and I. Alnarus Kautsar, "Classification Of Vocational High School Graduates' Ability In Industry Using Extreme Gradient Boosting (Xgboost), Random Forest, And Logistic Regression," *J. Tek. Inform.*, vol. 4, no. 4 SE-Articles, pp. 977–985, Sep. 2023, doi: 10.52436/1.jutif.2023.4.4.945.

[4] S. Jayachandran and B. Joshi, "Customized support vector machine for predicting the employability of students pursuing engineering," *Int. J. Inf. Technol.*, vol. 16, no. 5, pp. 3193–3204, 2024, doi: 10.1007/s41870-024-01818-w.

[5] R. Haque, A. Quek, C. Y. Ting, H. N. Goh, and M. R. Hasan, "Classification Techniques Using Machine Learning for Graduate Student Employability Predictions," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 1, pp. 45–56, 2024, doi: 10.18517/ijaseit.14.1.19549.

[6] L. G. R. Putra, D. D. Prasetya, and M. Mayadi, "Student Dropout Prediction Using Random Forest and XGBoost Method," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 9, no. 1, pp. 147–157, 2025, doi: 10.29407/intensif.v9i1.21191.

[7] P. R. Togatorop, M. Sianturi, D. Simamora, and D. Silaen, "Optimizing Random Forest using Genetic Algorithm for Heart Disease Classification," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 13, no. 1, p. 60, 2022, doi: 10.24843/lkjiti.2022.v13.i01.p06.

[8] Y. Guan, F. Wang, and S. Song, "Interpretable machine learning for academic performance prediction: A SHAP-based analysis of key influencing factors," *Innov. Educ. Teach. Int.*, pp. 1–20, Jul. 2025, doi: 10.1080/14703297.2025.2532050.

[9] M. M. Islam, F. H. Sojib, M. F. H. Mihad, M. Hasan, and M. Rahman, "The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system," *Telemat. Informatics Reports*, vol. 18, no. February, p. 100203, 2025, doi: 10.1016/j.teler.2025.100203.

[10] G. Caroline, P. Ningsih, F. Liantoni, and Y. Sujana, "Telematika A Systematic Analysis of the Impact of Non-Academic Factors on Student Academic Performance Prediction using Data Mining," vol. 19, no. 1, pp. 116–125, 2026.

[11] Swono Sibagariang, "Interpretable Machine Learning for Job Placement Prediction: A SHAP-Based Feature Analysis," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 14, no. 3, pp. 190–198, 2025,

doi: 10.22146/jnteti.v14i3.20516.

[12] C. Surya, E. Yubarda, K. Ameliza, and W. Simatupang, “Analyzing the Influence of Academic Competence and Soft Skills on Vocational Students ’ Work Readiness Using Regression and Machine Learning Approaches,” vol. 13, no. 1, pp. 32–43, 2026.

[13] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[14] M. F. Al Hakim, S. Wahyuni, K. Budiman, A. Marianti, and B. E. Susilo, “Optimization of Machine Learning Model using Grid and Random Search Algorithms for Predicting Student Dropout,” *J. Tek. Inform.*, vol. 7, no. 3, pp. 3012–3024, 2026.

[15] M. Martinović, K. Dokic, and D. Pudić, “Comparative Analysis of Machine Learning Models for Predicting Innovation Outcomes: An Applied AI Approach,” *Appl. Sci.*, vol. 15, no. 7, pp. 1–44, 2025, doi: 10.3390/app15073636.

[16] G. R. Saputra and M. Mujiyono, “Evaluating the implementation of the mechanical engineering student skills competency,” *J. Pendidik. Vokasi*, vol. 14, no. 2, pp. 194–207, 2024, doi: 10.21831/jpv.v14i2.71359.

[17] S. E. Davis, M. E. Matheny, S. Balu, and M. P. Sendak, “A framework for understanding label leakage in machine learning for health care,” *J. Am. Med. Informatics Assoc.*, vol. 31, no. 1, pp. 274–280, 2024, doi: 10.1093/jamia/ocad178.

[18] S. Studer *et al.*, “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology,” *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 392–413, 2021, doi: 10.3390/make3020020.

[19] I. V. Tetko, R. van Deursen, and G. Godin, “Be aware of overfitting by hyperparameter optimization!,” *J. Cheminform.*, vol. 16, no. 1, 2024, doi: 10.1186/s13321-024-00934-w.

[20] S. Widodo, H. Brawijaya, and S. Samudi, “Stratified K-fold cross validation optimization on machine learning for prediction,” *Sinkron*, vol. 7, no. 4, pp. 2407–2414, 2022, doi: 10.33395/sinkron.v7i4.11792.