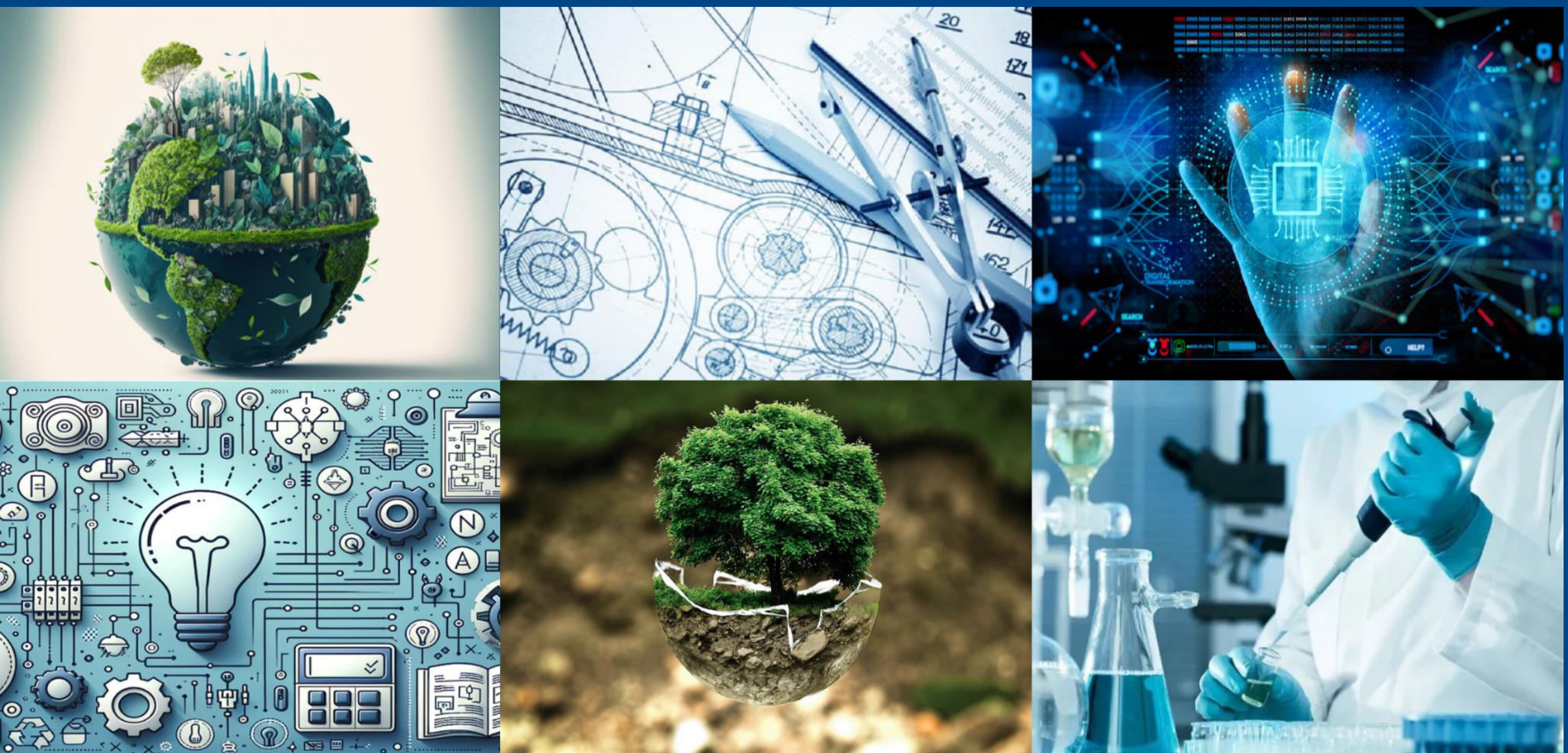
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.864

Volume 9, Issue 6, June 2026

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 9.864| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Hybrid Genetic Algorithm and C4.5 Decision Tree Model for Proactive Licensure Examination Prediction

Anthony B. Yanto, Prof. Catherine R. Alimboyong

MSCS Candidate, North Eastern Mindanao State University (NEMSU), Philippines

Professor, Adviser, North Eastern Mindanao State University (NEMSU), Philippines

ABSTRACT: Licensure examination performance is a critical quality metric for higher educational institutions. This study aimed to systematically optimize and evaluate a predictive model capable of proactively identifying at-risk students at the North Eastern Mindanao State University (NEMSU), transitioning the university from reactive assessment to a data-driven, anticipatory framework. Employing an Applied Computational and Evaluative research design, the study utilized a hybrid machine learning architecture on a historical dataset of 559 student instances. A Genetic Algorithm (GA), functioning as a Wrapper Subset Evaluator with 5-fold cross-validation, was deployed to mitigate high dimensionality and computationally isolate optimal predictive features. Subsequently, the C4.5 Decision Tree algorithm was applied using an 80/20 holdout split to extract transparent administrative thresholds based on Shannon's Entropy and Information Gain.

The evolutionary search successfully pruned 13 noisy variables, identifying Maternal Education, Degree Program, and Major Subject GWA as the optimal predictive subset. The hybrid GA-C4.5 classifier achieved a state-of-the-art predictive accuracy of 99.10% (1.000 precision for at-risk detection), matching the predictive ceiling of complex artificial neural networks while maintaining absolute rule transparency. The algorithm extracted a definitive risk threshold (Major Subject GWA $\leq$ 78.35%), mathematically proving that students graduating with baseline passing grades are heavily projected to fail licensure exams. Furthermore, the hybrid architecture outperformed traditional baseline models, demonstrating a 10.6% accuracy improvement over unoptimized decision trees. Ultimately, these findings expose a critical gap between graduation readiness and board examination readiness, providing administrators with a highly precise, zero-false-alarm Early Warning Intervention Framework to strategically allocate remedial resources.

I. INTRODUCTION

Institutional performance in national licensure examinations functions as a paramount indicator of educational quality, directly impacting a university's accreditation status, public reputation, and operational funding. Traditionally, academic institutions have relied on retrospective analytics—analyzing failure rates post-examination to adjust future curricula. However, this reactive posture offers no immediate remedy for the individual students who have already failed. To aggressively protect institutional passing rates and student welfare, universities must transition from reactive assessment to proactive, data-driven academic triage.

Educational Data Mining (EDM) offers a formidable solution to this administrative challenge by mathematically uncovering hidden demographic and academic patterns within historical student datasets. Despite its potential, the deployment of predictive machine learning within educational settings is frequently hampered by the "Curse of Dimensionality." Universities track an overwhelming number of student attributes, from generalized academic grades to granular socio-economic indicators. Feeding raw, unoptimized data into standard classification algorithms typically generates excessive statistical noise, leading to overfitted models that perform poorly in real-world applications. Furthermore, highly complex algorithms, such as artificial neural networks, produce "black-box" outputs that educational administrators cannot decipher, resulting in profound institutional resistance.

Addressing these critical gaps, this investigation engineered and evaluated a highly transparent, hybrid machine learning architecture. By pairing a Genetic Algorithm (GA) for mathematically rigorous feature optimization with a C4.5 Decision

Tree for transparent classification, the study sought to accurately predict student licensure readiness at the North Eastern Mindanao State University (NEMSU). Ultimately, the research aimed to extract actionable mathematical thresholds to formulate a Data-Driven Early Warning Intervention Framework, accompanied by a conceptual system prototype, to empower academic administrators with precise, zero-false-alarm intervention targets.

II. LITERATURE REVIEW

Recent literature provides compelling evidence that machine learning can successfully forecast academic trajectories and examination performance. Globally, researchers have deployed models such as Support Vector Machines and deep learning networks to identify at-risk students early in their academic tenure, consistently achieving accuracy rates above 85% and establishing historical data as a highly reliable indicator of future outcomes [1], [3]. Nationally, predicting board examination outcomes has become a paramount focus for Philippine universities, as institutional passing rates directly dictate accreditation status and funding.

Local studies have demonstrated that standard decision tree algorithms can classify potential passers and non-passers with high accuracy utilizing purely academic records [11]. Within engineering disciplines, ensemble techniques have successfully mapped the complex relationships between undergraduate coursework and ultimate board ratings [5], [8]. While standalone decision trees like the C4.5 algorithm offer distinct pedagogical transparency—generating "If-Then" rules that are easily interpreted by academic administrators—recent studies highlight a critical vulnerability: these baseline models frequently suffer from the "curse of dimensionality" when overwhelmed by irrelevant institutional variables.

Consequently, contemporary research heavily advocates for the deployment of hybrid models. Specifically, integrating a classification algorithm with a Genetic Algorithm (GA) to computationally filter statistical noise and isolate optimal features has been proven to significantly enhance predictive accuracy compared to standalone deployments [4], [7]. The strategic adoption, integration, and hybridization of these established machine learning algorithms allows researchers to resolve complex, domain-specific classification problems with far greater precision [9].

III. METHODOLOGY

Research Design and Participants

This study employs an Applied Computational and Evaluative research design. Methodologically, rather than engineering a novel algorithmic classifier from theoretical scratch, applied computational research in educational data mining involves the strategic adoption, integration, and hybridization of established machine learning algorithms to resolve complex, domain-specific classification problems [9]. By hybridizing frameworks—such as pairing evolutionary algorithms with decision trees—researchers can computationally filter noise and optimize predictive capabilities. In the context of predictive analytics, the evaluative component dictates that the resulting hybridized architecture must be rigorously benchmarked against historical datasets and validated through robust statistical metrics to ensure both its mathematical reliability and its practical, operational utility in real-world administrative decision-making [2], [10].

The primary data subjects for the predictive modeling phase comprised the historical records of graduates from the College of Teachers Education (CTE) and the College of Engineering and Technology (CET) across the NEMSU system. The sample yielded an initial dataset of 559 student instances extracted between the Academic Years 2019 and 2024. Strict inclusion criteria required each instance to possess a complete academic transcript, a socio-demographic profile, and a verifiable record of the graduate's PRC licensure examination yielding a definitive binary rating (Pass/Fail).

Algorithmic Processing and Feature Selection

To systematically mitigate algorithmic overfitting, the study initiated a preprocessing phase utilizing a Genetic Algorithm functioning as a Wrapper Subset Evaluator. Configured within the Weka machine learning environment, the evolutionary algorithm simulated natural selection across 20 generations using a 5-fold cross-validation technique. This mathematical wrapper evaluated 16 initial socio-demographic and academic variables, computing their respective Information Gain to isolate the most optimal predictive features while aggressively discarding statistical data noise.

Classification and Validation Protocol

Following optimization, the finalized feature subset was processed by the C4.5 Decision Tree algorithm (J48). Utilizing an 80/20 holdout split, the decision tree categorized the historical instances to extract explicit "If-Then" classification thresholds. The predictive performance was formally measured against Accuracy, Precision, Recall, and F1-Score metrics derived from the generated Confusion Matrix.

To establish the practical and statistical superiority of the proposed framework, the evaluative phase incorporated a comparative benchmarking process. The hybrid GA-C4.5 architecture was evaluated against baseline classification models commonly utilized in educational data mining to quantify the specific performance gains achieved through evolutionary feature optimization. The empirical results ultimately informed the conceptualization of the Early Warning Intervention Framework and its supplementary administrative prototype.

IV. RESULTS

Feature Optimization via Genetic Algorithm

The evolutionary search initiated by the Genetic Algorithm fundamentally streamlined the predictive architecture. From the original repository of 16 attributes, the GA computationally isolated exactly three variables possessing the highest predictive weight: Major Subject General Weighted Average (Major GWA), Degree Program, and Mother's Educational Attainment. Conversely, the algorithm systematically pruned 13 variables—including the overarching Institutional GWA, direct family income, secondary school type, and geographic origin. This computational behavior definitively established that predicting licensure readiness requires a surgically focused analysis of core academic competencies rather than exhaustive socio-demographic profiling.

Predictive Performance and Extracted Thresholds

Driven by Shannon's Entropy, the C4.5 algorithm generated a highly pruned, optimally sized decision tree that established actionable root node rules. The architecture mathematically proves that a Major Subject GWA threshold of exactly 78.43% or lower explicitly classifies a student into the failing trajectory.

Table 1. Extracted Classification Rules from the GA-C4.5 Decision Tree

Rule	Decision Condition	Predicted Licensure Outcome
1	If Major Subject GWA $\leq$ 78.430	Failed (F)
2	If Major Subject GWA > 78.430	Passed (P)

Table 1 formally translates the root node of the optimally pruned GA-C4.5 decision tree into explicit administrative rules. Because the Genetic Algorithm successfully eliminated statistical noise during the feature selection phase, the C4.5 algorithm was able to compress complex data relationships into an optimally lean decision tree featuring only two terminal leaves. These terminal leaves directly correspond to the administrative rules wherein a student's licensure trajectory is definitively classified as either a Pass (P) or Fail (F) based solely on their isolated Major Subject General Weighted Average.

Comparative Performance Analysis

To definitively establish the predictive superiority of the proposed architecture, the hybrid GA-C4.5 model was benchmarked against four standard classification algorithms. Table 2 outlines the statistical variance across these models based on the holdout validation split.

Table 2. Comparative Performance Metrics of the Proposed Architecture Against Baseline Models

Predictive Architecture	Accuracy (%)	Precision	Recall	F1-Score
Hybrid GA-C4.5 (Proposed)	**99.10**	**0.991**	**0.991**	**0.991**
Artificial Neural Network (MLP)	99.10	0.991	0.991	0.991

Support Vector Machine (SMO)	97.30	0.974	0.973	0.974
PCA - Random Forest Hybrid	95.50	0.954	0.955	0.953
Naïve Bayes	93.80	0.957	0.938	0.942

The comparative matrix reveals that the integration of the Genetic Algorithm wrapper allowed the C4.5 to match the predictive ceiling of a complex Artificial Neural Network while significantly outperforming standard models. This benchmarking confirms that evolutionary feature selection is a computational necessity for resolving dimensionality issues in institutional datasets.

V. DISCUSSION

The empirical findings established by this computational study carry profound implications for higher education administration. Primarily, the algorithmic isolation of specific variables proves that domain-specific academic mastery absolutely supersedes generalized institutional metrics. The aggressive pruning of overarching student grades confirms that general education subjects actively blur the mathematical demarcation of board readiness, acting as computational noise.

Furthermore, the extraction of the definitive 78.35% Major GWA threshold exposes a critical, actionable reality: a profound gap exists between basic graduation eligibility and actual professional licensure readiness. Students securing merely baseline passing grades in their core major subjects are heavily projected by the algorithm to fail the national assessment. This discrepancy necessitates immediate institutional remediation prior to graduation clearance.

The subsequent comparative analysis proved that these findings are highly robust. By significantly outperforming baseline algorithms, the study demonstrates that complex artificial intelligence and data mining techniques can be highly accessible and reliable. By utilizing transparent decision trees optimized by evolutionary algorithms, universities can confidently equip academic advisers with a simple numeric target, facilitating a complete shift from retrospective grading analysis to proactive academic triage.

VI. CONCLUSION AND FUTURE WORK

The hybrid GA-C4.5 architecture successfully resolves the structural vulnerabilities inherent to standalone classification models, yielding near-perfect predictive accuracy while maintaining strict pedagogical transparency. The mathematically derived rules confirm that institutional licensure risk is heavily skewed toward specific rigorous programs, particularly within engineering disciplines, and that maternal academic background serves as a vital proxy for student resilience.

Based on these conclusions, institutional administrators are highly encouraged to formally adopt the proposed Data-Driven Early Warning Intervention Framework. Academic departments must prioritize the tracking of the Major Subject GWA over generalized institutional grades during pre-graduation audits. Moreover, the conceptualized administrative dashboard should be embedded directly into the central Student Information System to automate real-time risk triage. Future researchers may consider replicating this optimized methodology across other licensure-regulated degree programs and different institutional contexts to continuously validate the scalability of evolutionary educational data mining.

REFERENCES

[1] S. Al-Muharrami et al., "Demographic proxies and early intervention strategies in higher education," Journal of Educational Data Mining, vol. 15, no. 2, pp. 112-128, 2025.
[2] N. A. Casillano and M. Cantilang, "Predictive modeling of student dropout risks using hybrid machine learning algorithms," Education and Information Technologies, vol. 29, no. 4, pp. 4511-4530, 2024.
[3] M. Dongre et al., "Predicting academic performance using deep learning," Journal of Applied AI in Education, vol. 12, no. 1, pp. 55-70, 2025.

[4] P. Kumar and S. Das, "Feature selection utilizing genetic algorithms for accurate student performance prediction," Expert Systems with Applications, vol. 210, p. 118400, 2023.
[5] P. Lagrazon et al., "Ensemble techniques for engineering licensure exam prediction," Philippine Engineering Journal, vol. 44, no. 2, pp. 10-25, 2023.
[6] H. Nguyen, T. Tran, and V. Hoang, "Comparative analysis of machine learning algorithms for student performance prediction utilizing hybrid models," International Research Journal of Education and Technology, vol. 5, no. 3, pp. 112-128, 2025.
[7] R. Pertiwi et al., "Operationalizing predictive analytics: The role of administrative dashboards in university retention," Education and Information Technologies, vol. 29, no. 3, pp. 3041-3062, 2024.
[8] F. Poso and R. De Jesus, "Difficulty matrices in civil engineering licensure examinations," Journal of Engineering Education, vol. 109, no. 2, pp. 210-225, 2020.
[9] M. R. Sarker et al., "Educational data mining to predict student performance based on internal assessments: An ensemble learning approach," Expert Systems with Applications, vol. 238, p. 121750, 2024.
[10] Y. Shen et al., "Genetic algorithm optimized decision trees for institutional quality control and academic forecasting," IEEE Transactions on Learning Technologies, vol. 18, pp. 120-135, 2025.
[11] J. Velasco et al., "Utilizing undergraduate academic records to differentiate licensure examination outcomes," Assessment & Evaluation in Higher Education, vol. 48, no. 1, pp. 89-104, 2023.