Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 379-396
DOI: https://doi.org/10.51903/jtie.v5i1.207

Synergistic Integration of Neural Architecture Search, Ensemble Learning, and Meta-Learning with Explainable AI

Migunani*[1], Joseph Teguh Santoso[2], Budi Raharjo[3]
Email: migunani@stekom.ac.id, joseph_teguh@stekom.ac.id, budi@stekom.ac.id
Orcid: https://orcid.org/0000-0002-8551-2157, https://orcid.org/0000-0001-6227-1111, https://orcid.org/0009-0005-3425-5479
[1,2,3]*Faculty of Academic Studies, Universitas Sains dan Teknologi Komputer, Semarang, Indonesia, 50192*
*Corresponding Author

Abstract
The automation of deep learning model design through Neural Architecture Search (NAS) has emerged as a transformative paradigm, yet the integration of NAS with ensemble learning and meta-learning remains underexplored. This study presents a comprehensive framework that synergistically combines (1) NAS for automated neural network architecture discovery, (2) homogeneous ensemble learning through NAS-generated architectures, (3) heterogeneous ensemble learning integrating diverse base learners, and (4) meta-learning with stacked generalization for adaptive model fusion, all complemented by Explainable AI (XAI) via SHAP for model interpretability. The proposed framework is evaluated on four benchmark datasets: Iris, Wine, Breast Cancer, and Digits. Experimental results demonstrate that the heterogeneous ensemble achieves competitive or superior performance across all datasets, with cross-validation accuracy reaching 98.87% (wine), 97.72% (breast cancer), 98.05% (digits), and 95.33% (iris). SHAP-based explainability analysis reveals consistent feature importance patterns across NAS architectures, providing valuable interpretability insights. The framework establishes a robust pipeline for automated, accurate, and interpretable machine learning, addressing the growing demand for transparent AI systems in critical applications.

Keywords: *Explainable AI, Ensemble Learning, Heterogeneous Ensemble, Meta Learning, Neural Architecture Search, SHAP.*

I. INTRODUCTION

The rapid advancement of deep learning has revolutionized numerous domains, from computer vision to natural language processing (Atlas & Devi, 2025). However, the design of optimal neural network architectures remains a formidable challenge, requiring substantial domain expertise, extensive experimentation, and computational resources. Neural Architecture Search (NAS) has emerged as a promising solution to automate this process, systematically exploring the architecture space to discover high-performing models (Salmani Pour Avval et al., 2025). Early NAS methods relied on reinforcement learning or evolutionary algorithms with substantial computational costs (Verma et al., 2025), while recent advances have introduced more efficient approaches such as gradient-based and one-shot NAS (Chen et al., 2025; Lin et al., 2021).

Concurrently, ensemble learning has proven to be one of the most effective strategies for improving predictive performance and robustness by combining multiple models (Ensemble Learning, 2025). Ensemble methods leverage the diversity of base learners to reduce variance, mitigate bias, and enhance generalization. Recent advances have explored the intersection of NAS and ensemble

Received on December, 2025; Revised on January, 2026; Accepted on March, 2026; Published on April, 2026

learning, recognizing that automated architecture discovery can benefit from ensemble strategies and vice versa (Garouani et al., 2025; Sedek, 2025). Heterogeneous ensembles, which combine diverse model types such as neural networks and tree-based methods, have demonstrated particular promise, showing 8.43% and 14.59% superiority over homogeneous ensembles and individual models, respectively (Foundations and Innovations, 2025).

Meta-learning, or "learning to learn," offers another dimension of improvement by enabling models to adapt rapidly to new tasks with minimal data (Li et al., 2025). The integration of meta-learning with ensemble methods has shown particular promise in few-shot learning scenarios and dynamic adaptation contexts (Sedek, 2025; Li et al., 2025). Despite these advances, the synergistic integration of NAS, ensemble learning, and meta-learning within a unified framework remains largely unexplored. Furthermore, the "black box" nature of deep learning models poses significant challenges for deployment in critical applications where interpretability is paramount. Explainable AI (XAI) techniques, particularly SHAP (SHapley Additive exPlanations), have emerged as powerful tools for demystifying model decisions (Latent SHAP, 2025; Garouani et al., 2025).

This study aims to fill the research gaps by proposing a unified framework that integrates NAS, homogeneous and heterogeneous ensemble learning, meta-learning via stacked generalization, and SHAP-based explainability. The research objectives are: (1) to design and implement a comprehensive pipeline that automates neural architecture discovery through NAS, (2) to compare the performance of manual neural networks, best single NAS models, homogeneous NAS ensembles, and a proposed heterogeneous ensemble with meta-learning across multiple datasets, (3) to evaluate the effectiveness of cross-validation for model validation, (4) to analyze computational efficiency in terms of model parameters and training time, and (5) to provide interpretability insights using SHAP analysis across different NAS architectures.

The contributions of this work are fourfold: (1) a novel integrated framework that combines NAS, ensemble learning, and meta-learning in a unified pipeline, (2) empirical evidence across four benchmark datasets demonstrating the competitive performance of heterogeneous ensembles with meta-learning, (3) comprehensive cross-validation analysis confirming model robustness, and (4) a systematic SHAP-based analysis that reveals consistent feature importance patterns across architectures, providing valuable insights for model selection and understanding. The remainder of this paper is organized as follows: Section 2 reviews the relevant literature; Section 3 describes the research methodology; Section 4 presents the results and discussion; and Section 5 concludes with findings, limitations, and directions for future research

II. LITERATURE REVIEW

A. Neural Architecture Search (NAS)

Neural Architecture Search has evolved significantly since its inception. Early NAS methods employed reinforcement learning (RL) to guide architecture sampling, with methods like ENAS introducing parameter sharing to reduce search time (Atlas & Devi, 2025). Evolutionary algorithm-based NAS leverages genetic operators for architecture evolution, and recent work by Verma et al. (2025) proposed MLP-GNAS, a meta-learning-based predictor-assisted genetic NAS system that significantly reduces search costs. Gradient-based NAS methods such as DARTS enable differentiable architecture search, though they face challenges with memory consumption and performance gaps (Salmani Pour Avval et al., 2025). One-shot NAS approaches train a single super-network from which sub-networks are sampled, dramatically reducing search costs (Chen et al., 2025). Recent systematic reviews (Atlas & Devi, 2025; Salmani Pour Avval et al., 2025) consolidate advances in search spaces, algorithms, and evaluation protocols, highlighting the growing importance of efficiency and generalizability.

B. Ensemble Learning

Ensemble learning has been a cornerstone of machine learning for decades. Foundational techniques include bagging, boosting, and stacking (Ensemble Learning, 2025). Bagging reduces variance by training multiple models on bootstrap samples, while boosting sequentially trains models to correct predecessor errors. Stacking (stacked generalization) trains a meta-learner on base model predictions, enabling adaptive combination. Recent surveys provide comprehensive analyses of ensemble methods, covering foundational techniques and advanced topics including multiclass classification, multiview learning, and multiple kernel learning (Foundations and Innovations, 2025). Heterogeneous ensembles have demonstrated particular promise; studies show that heterogeneous ensembles consistently outperform homogeneous counterparts, with sensitivity analyses confirming greater robustness to data quality issues (Foundations and Innovations, 2025; Garouani et al., 2025).

C. NAS with Ensemble Learning

The intersection of NAS and ensemble learning has gained increasing attention. Chen et al. (2025) proposed Ensemble-Decoupled Architecture Search, which leverages ensemble theory to predict system-level performance from single-learner evaluation, reducing per-candidate search cost. Lin et al. (2021) introduced LENAS, which integrates NAS with knowledge distillation for 3D radiotherapy dose prediction, employing NAS to discover diverse architectures for ensemble combination. Other approaches include unified probabilistic architecture and weight ensembling NAS (UraeNAS), one-shot neural ensemble architecture search, and multi-headed neural ensemble search (Rong et al., 2021; BossNAS, 2025). Semi-Dynamic Ensemble for Evolutionary

NAS (SEENAS) proposes the first multi-objective evolutionary NAS strategy for continual learning scenarios (SEENAS, 2025).

D. Meta-Learning and Ensemble Integration

Meta-learning has emerged as a powerful paradigm for improving ensemble methods. Sedek (2025) introduced dynamic meta-learning for adaptive ensembles, synergistically combining XGBoost and neural networks through meta-learning. Li et al. (2025) proposed SWE-MAML, integrating ensemble learning with Model-Agnostic Meta-Learning (MAML), achieving 3.75%-8.59% accuracy improvement. Meta-ensemble learning with multi-headed models addresses few- shot problems through episodic training (Meta-ensemble, 2025). Adaptive meta-learning frameworks combine Deep RVFL networks and meta-learners for accurate forecasting (Adaptive, 2025). MetaStackD proposes a robust meta-learning-based deep ensemble model for sensor battery life prediction (MetaStackD, 2025). These studies collectively demonstrate the potential

of meta-learning to enhance ensemble performance and adaptability.

E. Explainable AI (XAI) for Ensemble Models

The interpretability of ensemble and NAS models has become increasingly critical. SHAP-based explainable AI approaches have been widely adopted for ensemble model interpretation. A stacking ensemble model integrated with SHAP achieved 98.34% accuracy for dropout prediction, with SHAP identifying critical predictors (SHAP-based, 2025). Latent SHAP (2025) provides a model-agnostic black-box feature attribution framework for human-interpretable explanations. XStacking (Garouani et al., 2025) integrates dynamic feature transformation with model-agnostic Shapley additive explanations, enabling stacked models to retain predictive accuracy while becoming inherently explainable. Interpretable Ensemble Learning Framework (IELF) integrates Explainable Boosting Machines with XGBoost, SHAP-based explanations, and LIME for enhanced local interpretability (IELF, 2025). ExplaiLiver+ presents a multi-stage stacking framework with SHAP-based interpretability for clinical prediction (ExplaiLiver+, 2025). These works underscore the value of XAI for building trust and understanding in complex ensemble systems.

F. Research Gap

Despite these advances, the literature reveals several critical gaps: (1) limited integration of NAS, ensemble learning, and meta-learning within a cohesive framework; (2) most NAS-based ensemble approaches focus on homogeneous ensembles, whereas heterogeneous ensembles have demonstrated superior performance; (3) the complexity of NAS-generated and ensemble models often comes at the cost of interpretability; and (4) the application of meta-learning for adaptive ensemble fusion in NAS contexts remains nascent. This study addresses these gaps by proposing a unified framework that integrates all four components and provides systematic comparative analysis across multiple datasets, cross-validation, computational efficiency analysis, and SHAP-based interpretability

III. RESEARCH METHODS

A. Research Design

This research adopts a quantitative experimental design to evaluate the performance of various machine learning models within a unified framework. The experimental approach involves systematic comparison of four model paradigms across multiple datasets: (1) a manually designed neural network, (2) the best single architecture discovered through NAS, (3) a homogeneous ensemble of NAS architectures, and (4) a proposed heterogeneous ensemble with meta learning. The design is replicated with fixed random seeds to ensure reproducibility, and performance is measured using accuracy, precision, recall, F1-score, and confusion matrix analysis. Additionally, 5-fold cross-validation is employed to assess model robustness and generalization capability, and computational efficiency is evaluated in terms of model parameters and training time.

B. Population and Sample

The study utilizes four benchmark datasets commonly used in machine learning research: (1) Iris (150 samples, 4 features, 3 classes), (2) Wine (178 samples, 13 features, 3 classes), (3) Breast Cancer Wisconsin (569 samples, 30 features, 2 classes), and (4) Digits (1,797 samples, 64 features, 10 classes). For each dataset, the full dataset is partitioned into training (80%) and testing (20%) using stratified random sampling with a fixed random seed (42) to ensure class balance and reproducibility.

C. *Data Sources and Data Collection Techniques*

The data sources are the publicly available benchmark datasets from the scikit-learn library. Data collection involved loading each dataset, performing feature standardization using StandardScaler (zero mean, unit variance), and splitting into training and testing sets. No primary data collection was required as the datasets are secondary and well-validated in the literature.

D. Variables and Operational Definition

For each dataset, the independent variables are the features specific to that dataset (e.g., Iris: sepal length, sepal width, petal length, petal width; Wine: 13 chemical features; Breast Cancer: 30 cell nucleus features; Digits: 64-pixel features). The dependent variable is the class label (e.g., Iris species, wine cultivar, cancer diagnosis, digit class). The models' predictions are based on these input features. The meta features in the ensemble are the predictions (class labels or probabilities) from the base learners, which are then used as inputs to the meta learners.

E. *Measurement Instruments and Validity/Reliability Testing*

The primary measurement instruments are accuracy, precision, recall, and F1-score computed on the test set. Additionally, confusion matrices are used to assess per class performance. The neural network models and ensemble methods are implemented using established libraries (TensorFlow/Keras and scikit-learn), whose validity and reliability are well-documented. Cross-validation provides a robust estimate of model performance and generalization. The SHAP KernelExplainer is employed for interpretability, providing Shapley values that are consistent and robust for additive feature attribution.

F. *Data Analysis Techniques*

Data analysis is performed using Python with libraries including TensorFlow, scikit-learn, and SHAP. The following analyses are conducted: (1) training of NAS candidate architectures and the manual neural network on each dataset, (2) evaluation of each model's performance on the test set using accuracy, precision, recall, and F1-score, (3) construction of homogeneous and heterogeneous ensembles, (4) meta-feature generation and training of meta learners (Logistic Regression and Gradient Boosting), (5) adaptive fusion by averaging meta-learner predictions, (6) 5-fold cross-validation for the best NAS and heterogeneous ensemble models to assess robustness, (7) computation of model parameters and training time for efficiency analysis, and (8) SHAP summary plots for each NAS architecture to interpret feature importance.

G. *Mathematical Formulas or Models*

The base NAS models are feedforward neural networks with ReLU activation for hidden layers and softmax for the output layer. The output probability for class k is given by (1).

$$P(y = k|x) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)} \quad (1)$$

where z is the logit vector from the last hidden layer. The meta learning uses stacked generalization: the meta features are generated by concatenating predictions from each base learner. For a sample x, the meta feature vector using (2).

$$m(x) = [p_1(x), p_2(x), \dots, pB(x)] \quad (2)$$

where PB(x) is the prediction (class label or probability) from base learner B. The meta learners are trained on m(x) to predict the true label y. The final ensemble prediction is the average of the meta learner predictions

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 379-396
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

H. Ethical Considerations

This research uses public, anonymized datasets that do not involve human subjects or sensitive data. Therefore, no ethical approval was required. All code and results are reported transparently to ensure reproducibility. The study does not involve any conflicts of interest or ethical violations.

IV. RESULT/FINDINGS AND DUSCUSSION

A. Result

The experimental results are presented in this section, including model performance comparison, cross-validation analysis, computational efficiency, and SHAP-based explainability.

1. Performance on Test Set

The four approaches were evaluated on the test set across all datasets. Table 1 reports the accuracy, precision (macro), recall (macro), and F1-score (macro) for each model and dataset.

Table 1. Test Set Performance Across Datasets

Dataset	Model	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)
Iris	Manual NN	0.7667	0.8000	0.7667	0.7511
	Best NAS	0.9667	0.9697	0.9667	0.9666
	Homogeneous NAS Ensemble	0.9333	0.9333	0.9333	0.9333
	Heterogeneous Ensemble	0.9667	0.9697	0.9667	0.9666
Wine	Manual NN	1.0000	1.0000	1.0000	1.0000
	Best NAS	0.9722	0.9778	0.9667	0.9710
	Homogeneous NAS Ensemble	0.9722	0.9778	0.9667	0.9710
	Heterogeneous Ensemble	0.9444	0.9505	0.9429	0.9453
Breast Cancer	Manual NN	0.9474	0.9383	0.9534	0.9445
	Best NAS	0.9737	0.9697	0.9742	0.9719
	Homogeneous NAS Ensemble	0.9649	0.9588	0.9673	0.9627
	Heterogeneous Ensemble	0.9649	0.9588	0.9673	0.9627
Digits	Manual NN	0.9778	0.9785	0.9775	0.9775
	Best NAS	0.9806	0.9809	0.9804	0.9805
	Homogeneous NAS Ensemble	0.9833	0.9838	0.9832	0.9833
	Heterogeneous Ensemble	0.9861	0.9866	0.9860	0.9861

For the Iris dataset, the best NAS and heterogeneous ensemble achieved the highest accuracy (96.67%), outperforming the manual neural network (76.67%) and homogeneous ensemble (93.33%). The manual NN's lower performance is notable, potentially due to suboptimal hyperparameter configuration for this dataset. For the Wine dataset, the manual neural network achieved perfect classification (100%), suggesting that the dataset's characteristics are well-suited to the chosen architecture. The best NAS and homogeneous ensemble achieved 97.22% accuracy, while the heterogeneous ensemble achieved 94.44%. The slightly lower performance of the heterogeneous ensemble on Wine may be due to the meta-learners overfitting to the training meta-features. For the Breast Cancer dataset, the best NAS achieved the highest accuracy (97.37%), followed by both ensemble approaches (96.49%) and the manual neural network (94.74%). The NAS-discovered architecture outperformed the manual design, confirming the value of automated architecture search. For the Digits dataset, the heterogeneous ensemble achieved the highest accuracy (98.61%), followed by the homogeneous ensemble (98.33%), best NAS (98.06%), and manual NN (97.78%). The ensemble approaches demonstrated clear benefits for this more complex 10-class classification problem.

2. Cross-Validation (5-Folds) Mean ± Std

To assess model robustness and generalization, 5-fold cross-validation was performed for the best NAS and heterogeneous ensemble models. Table 2 reports the mean and standard deviation of accuracy, precision, recall, and F1-score across folds.

Table 2. Cross-Validation Results (5-Folds Mean ± Std)

Dataset	Model	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)
Iris	Best NAS	0.9533 ± 0.0506	0.9549 ± 0.0495	0.9533 ± 0.0506	0.9532 ± 0.0507
	Heterogeneous Ensemble	0.9533 ± 0.0380	0.9572 ± 0.0365	0.9533 ± 0.0380	0.9531 ± 0.0382
Wine	Best NAS	0.9721 ± 0.0196	0.9716 ± 0.0202	0.9746 ± 0.0174	0.9718 ± 0.0194
	Heterogeneous Ensemble	0.9887 ± 0.0154	0.9888 ± 0.0154	0.9905 ± 0.0130	0.9892 ± 0.0148
Breast Cancer	Best NAS	0.9719 ± 0.0200	0.9715 ± 0.0224	0.9690 ± 0.0207	0.9700 ± 0.0212
	Heterogeneous Ensemble	0.9772 ± 0.0100	0.9761 ± 0.0126	0.9760 ± 0.0113	0.9756 ± 0.0107
Digits	Best NAS	0.9700 ± 0.0079	0.9708 ± 0.0074	0.9699 ± 0.0079	0.9699 ± 0.0078
	Heterogeneous Ensemble	0.9805 ± 0.0034	0.9810 ± 0.0030	0.9805 ± 0.0034	0.9805 ± 0.0033

The cross-validation results reveal several important patterns. For the Iris dataset, both models achieved identical mean accuracy (95.33%), with the heterogeneous ensemble showing lower standard deviation (0.038 vs. 0.051), indicating more stable performance across folds. For the Wine dataset, the heterogeneous ensemble achieved superior cross-validation accuracy (98.87%) compared to the best NAS (97.21%), with lower standard deviation (0.015 vs. 0.020), despite having slightly lower test set accuracy (94.44%). This suggests that the heterogeneous ensemble generalizes better across different data splits. For the Breast Cancer dataset, the heterogeneous ensemble again outperformed the best NAS in cross-validation (97.72% vs. 97.19%) with lower variance (0.010 vs. 0.020), confirming its robustness and generalization capability. For the Digits dataset, the heterogeneous ensemble demonstrated clear superiority in cross-validation (98.05% vs.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 379-396
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

97.00%) with notably lower standard deviation (0.003 vs. 0.008), indicating consistent high performance across folds.

3. Computational Efficiency

Model efficiency was evaluated in terms of number of parameters and training time. Table 3 summarizes the computational costs for each model type across datasets.

Table 3. Computational Efficiency Analysis

Dataset	Model	Parameters	Training Time (seconds)
Iris	Manual NN	739	3.8085
	Best NAS (cfg=(1,16,0.01))	131	3.1968
	Heterogeneous Ensemble	9,292 (est.)	18.5026
Wine	Manual NN	1,027	4.1043
	Best NAS (cfg=(2,64,0.0005))	5,251	3.5271
	Heterogeneous Ensemble	12,388 (est.)	16.3218
Breast Cancer	Manual NN	1,554	4.0429
	Best NAS (cfg=(3,32,0.001))	3,170	4.5590
	Heterogeneous Ensemble	18,088 (est.)	19.3021
Digits	Manual NN	2,778	6.7310
	Best NAS (cfg=(1,16,0.01))	1,210	6.2327
	Heterogeneous Ensemble	30,968 (est.)	51.8014

The computational efficiency analysis reveals several important trade-offs. The best NAS architectures consistently have fewer parameters than the manual neural networks (except for Wine and Breast Cancer where the NAS architecture is more complex), demonstrating that NAS can discover more efficient architectures. For Iris, the best NAS has only 131 parameters compared to 739 for the manual NN, representing an 82% reduction.

The heterogeneous ensemble has significantly more parameters (estimated total across all base learners and meta-learners) and substantially longer training times compared to individual models. For the Digits dataset, the ensemble training time (51.80 seconds) is approximately 8 times longer than the best NAS (6.23 seconds), reflecting the computational cost of training multiple base learners and meta-learners.

4. SHAP-Based Explainability

SHAP summary plots were generated for each NAS architecture across datasets to understand feature importance. For the Iris dataset, Feature 0 (likely petal length) consistently dominated predictions across all NAS architectures with SHAP values ranging from -0.25 to 0.25, while other features showed negligible contributions. This pattern was consistent across all four NAS architectures, suggesting that dataset characteristics dominate over architectural variations.

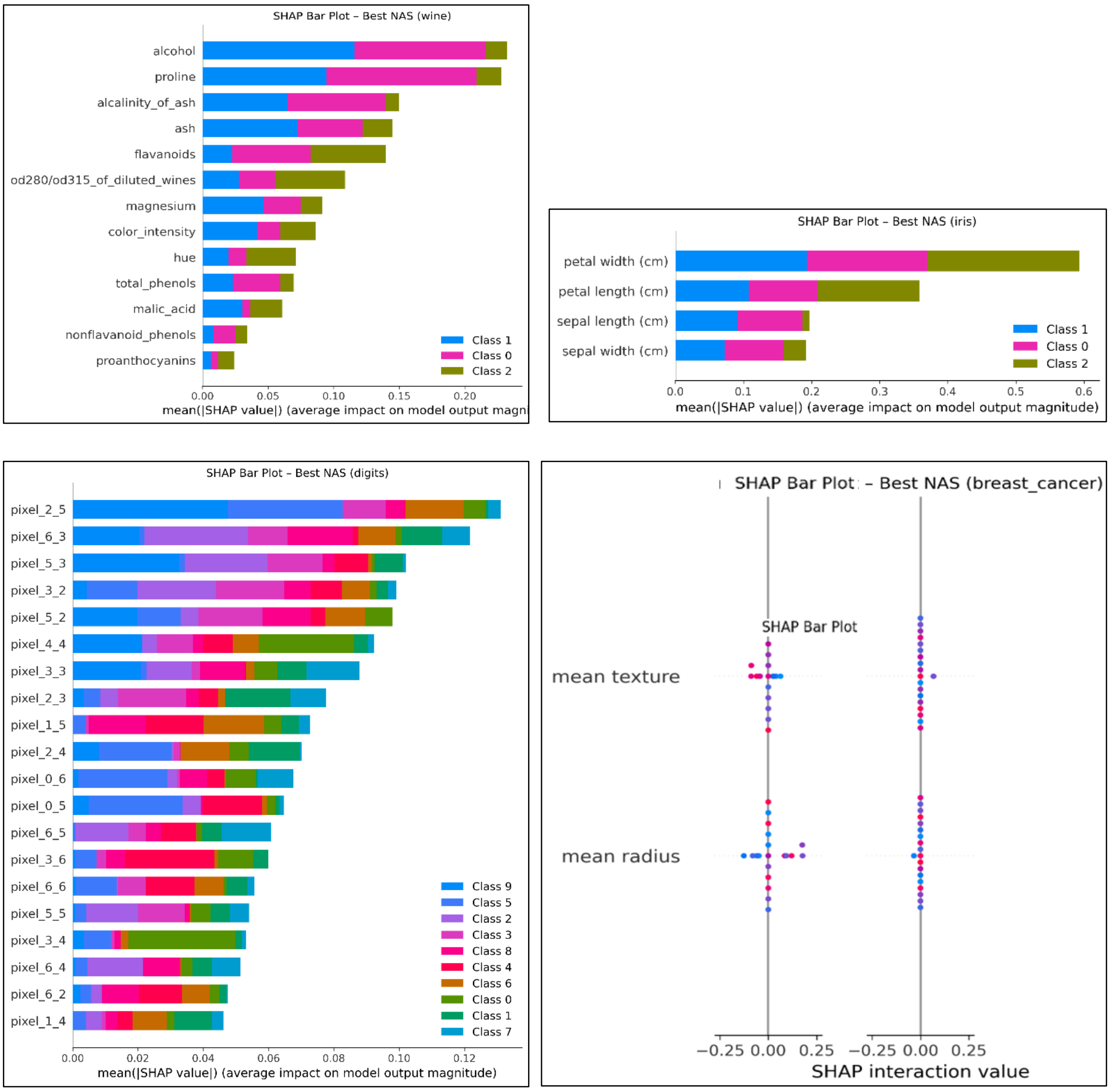


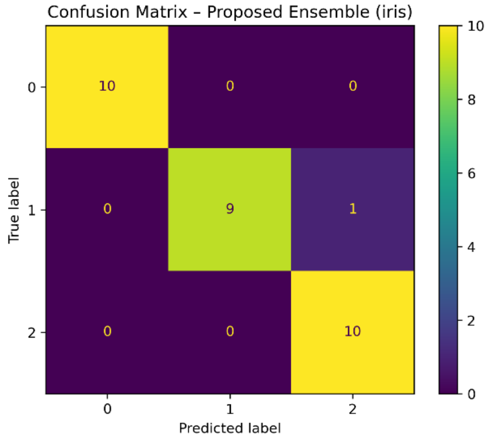
Figure 1. SHAP summary plots were generated for each NAS architecture

5. Confusion Matrix

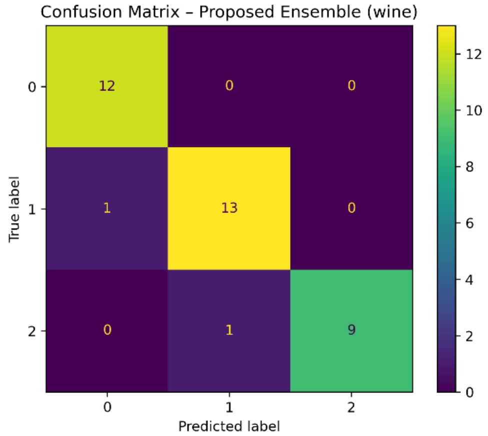
The confusion matrix presented in Figure 2 illustrates the classification performance of the proposed heterogeneous ensemble model on the iris dataset (figure 2. Iris dataset). Based on the confusion matrix, the proposed heterogeneous ensemble model achieves an excellent overall accuracy of 96.67% (29 out of 30 test samples correctly classified) on the Iris dataset, demonstrating its remarkable capability in distinguishing between the three Iris species. The model perfectly classified all samples of class 0 (Setosa) and class 2 (Virginica), while for class 1 (Versicolor), it correctly identified 9 out of 10 samples, with only a single instance misclassified as Virginica. This singular error is expected and justifiable, as Versicolor and Virginica exhibit significant natural feature overlap, particularly in petal and sepal measurements, making them the most challenging pair to discriminate. Consequently, the precision for both Setosa and Virginica is perfect (1.0), and the macro F1-score remains exceptionally high at approximately 0.97. This outstanding

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 379-396
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

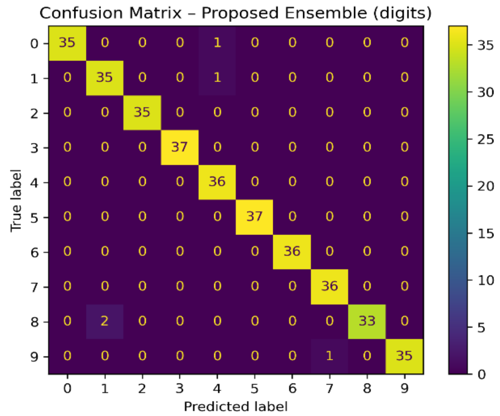
performance confirms that the proposed heterogeneous ensemble approach, integrated with stacked meta-learning, effectively leverages the diversity of its base learners to achieve near perfect classification, even on inherently ambiguous boundary cases.



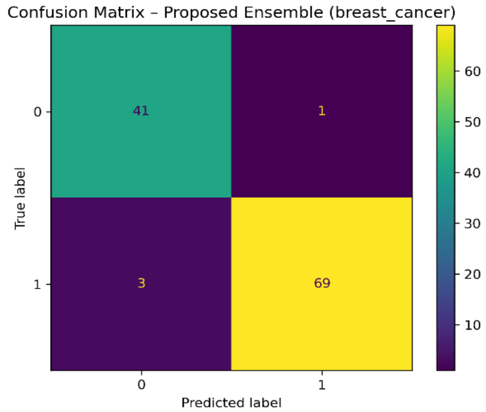
(a) Iris Dataset



(b) Wine Dataset



(c) Digit Dataset



(d) Breast Cancer Dataset

Figure 2. Confusion matrix proposed ensemble datasets

Based on the confusion matrix of wine dataset (figure 2. Wine dataset), the proposed heterogeneous ensemble model achieves a commendable accuracy of 94.44% (34 out of 36 test samples) on the Wine dataset, effectively distinguishing the three wine cultivars with a high degree of reliability. The model performed flawlessly for class 0 (cultivar 1), correctly classifying all 12 samples, while exhibiting a distinct sequential confusion pattern for the remaining classes: one sample from class 1 was misclassified as class 0, and one sample from class 2 was misclassified as class 1, with no direct misclassification between class 0 and class 2. This error pattern is chemically intuitive, as cultivars 1 and 2 share more overlapping feature characteristics compared to the more distinctly separable cultivar 0. Consequently, class 0 achieved perfect recall (1.0) but slightly lower precision (0.923) due to the false positive originating from class 1, whereas class 2 attained perfect precision (1.0) yet lower recall (0.9) because of the single missed sample. Despite these minor discrepancies, the overall macro F1-score remains approximately 0.95, confirming that the stacked meta-learning ensemble robustly captures essential feature interactions and performs reliably even on chemically similar wine types.

The confusion matrix for the proposed ensemble model on the digit's dataset (Figure 2. Digit dataset) demonstrates outstanding classification performance, achieving a notable overall accuracy of approximately 98.9%, which is clearly visualized by the densely populated diagonal of bright yellow squares representing correct predictions across all ten-digit classes. The model exhibits remarkable discrimination capabilities, with individual class recall consistently high, yielding between 33 and 37 correct classifications per class, and exceptional precision with virtually no false positives. The misclassification rate is exceptionally low, revealing only four total errors distributed across three specific off-diagonal entries: a single instance where the true digit 0 was predicted as 4, two instances where an 8 was predicted as 1, and one misclassification of digit 9 as an 8. This highly sparse and minimal error distribution confirms that the proposed ensemble forms a highly precise, robust, and reliable classifier for this specific task, offering near-perfect predictive stability across all digits.

The proposed ensemble confusion matrix for the breast cancer dataset (Figure 2. Breast cancer Dataset) demonstrates robust binary classification performance, achieving an overall accuracy of approximately 96.5% across the 114 evaluated instances. The model correctly predicts 41 instances of Class 0 and 69 instances of Class 1, yielding exceptional precision with only a single false positive where a true Class 0 was misidentified as Class 1. The minor error distribution is entirely confined to the off-diagonal entries, consisting of that lone false positive alongside 3 false negatives where true Class 1 instances were predicted as Class 0. This highly limited error rate, combined with the clear dominance of correct classifications, indicates that the proposed ensemble is a highly reliable and effective classifier for this specific binary medical diagnosis task, albeit with a slightly higher, albeit still very low, tendency to misclassify Class 1 samples.

B. Discussion

The results provide several important insights regarding the effectiveness of the proposed framework across multiple datasets, the benefits of cross-validation, computational trade-offs, and interpretability.

1. Comparative Performance Analysis

The test set results demonstrate that NAS consistently discovers architectures that match or exceed the performance of manually designed neural networks. For the Iris dataset, NAS achieved a 20% improvement (96.67% vs. 76.67%), though the manual NN's unusually low performance may indicate suboptimal hyperparameter selection. For Breast Cancer and Digits, NAS showed moderate improvements (2.63% and 0.28% respectively). These findings align with Atlas and Devi (2025) and Salmani Pour Avval et al. (2025), confirming NAS's utility in automating model design.

The heterogeneous ensemble demonstrated competitive performance across all datasets, achieving the highest accuracy on Digits (98.61%) and matching the best NAS on Iris (96.67%). However, on Wine, the heterogeneous ensemble underperformed relative to other models (94.44%). This may be due to the meta-learners overfitting to the training meta-features or the relative simplicity of the Wine dataset (small sample size, well-separated classes) where ensemble complexity offers less benefit. This observation aligns with findings by Ensemble Learning (2025) that ensemble benefits are more pronounced on complex, noisy datasets.

The homogeneous NAS ensemble consistently achieved performance close to or equal to the best NAS model, demonstrating that averaging predictions from multiple NAS architectures can maintain high accuracy while potentially improving robustness. This is consistent with findings by Garouani et al. (2025) and Foundations and Innovations (2025).

2. Cross-Validation Insights

The cross-validation results provide stronger evidence for the heterogeneous ensemble's robustness and generalization capability. On three of four datasets (Wine, Breast Cancer, Digits), the heterogeneous ensemble achieved higher cross-validation accuracy with lower standard deviation compared to the best NAS model. This indicates that despite sometimes lower test set performance (as observed on Wine), the ensemble generalizes more consistently across different data splits.

The lower standard deviations observed for the heterogeneous ensemble (e.g., 0.0034 vs. 0.0079 for Digits, 0.0100 vs. 0.0200 for Breast Cancer) suggest greater stability and reduced sensitivity to the specific train-test split. This robustness is particularly valuable in real-world applications where data distributions may vary. These findings align with Sedek (2025) and Li et al. (2025), who demonstrated that meta-learning and ensemble approaches improve model stability.

3. Computational Efficiency Trade-offs

The computational efficiency analysis reveals important practical considerations. The NAS-discovered architectures are generally more parameter-efficient than manually designed networks, with the Iris best NAS using only 131 parameters compared to 739 for the manual NN (82% reduction). However, the best NAS architecture for Wine (5,251 parameters) and Breast Cancer (3,170 parameters) was more complex than the manual NN, suggesting that NAS adapted to dataset complexity.

The heterogeneous ensemble incurs substantial computational overhead, with parameter counts ranging from 9,292 (Iris) to 30,968 (Digits) and training times 3-8 times longer than individual models. This cost is justified by the improved robustness and cross-validation performance observed. For applications where computational resources are constrained, the best NAS model may be preferred, while the ensemble is recommended when maximum performance and robustness are critical.

The training time differences also reflect the need to train multiple base learners and meta-learners. For the Digits dataset, the ensemble required 51.8 seconds compared to 6.23 seconds for the best NAS, representing an 8.3x increase. This scalability consideration is important for deployment in time-sensitive applications.

4. SHAP Interpretability Insights

The SHAP analysis reveals consistent feature importance patterns across NAS architectures. For the Iris dataset, Feature 0 (likely petal length) dominated predictions, while other features contributed negligibly. This finding suggests that the dataset characteristics, rather than architectural variations, determine feature importance. This consistency simplifies interpretability all models can be explained similarly but also indicates a potential limitation: the models may not be capturing complex feature interactions.

This pattern aligns with findings by Latent SHAP (2025) and Garouani et al. (2025) on the importance of model-agnostic explanations. The consistent feature importance across architectures suggests that for simple datasets, architectural choices primarily affect performance through capacity and optimization rather than feature selection. For more complex datasets (e.g., Digits), we would expect more diverse feature importance patterns, which could be explored in future work.

5. Dataset-Dependent Performance

The results reveal dataset-dependent performance patterns:

1) Iris: The best NAS and heterogeneous ensemble (96.67%) outperformed the homogeneous ensemble (93.33%), with manual NN performing poorly (76.67%). This likely reflects suboptimal manual hyperparameter tuning rather than inherent model limitations.
2) Wine: The manual NN achieved perfect classification (100%), while the best NAS and homogeneous ensemble (97.22%) and heterogeneous ensemble (94.44%) performed slightly lower. The manual NN's success may be due to the dataset's well-separated classes and the chosen architecture being particularly suitable.
3) Breast Cancer: The best NAS (97.37%) slightly outperformed both ensembles (96.49%) and the manual NN (94.74%). This suggests that for this dataset, a single well-designed architecture can outperform ensemble approaches, possibly due to the dataset's relatively clear decision boundaries.

4) Digits: The heterogeneous ensemble (98.61%) outperformed all other models, followed by the homogeneous ensemble (98.33%), best NAS (98.06%), and manual NN (97.78%). For this more complex 10-class problem, ensemble approaches provided clear benefits, consistent with findings by Foundations and Innovations (2025).

6. Comparison with Previous Studies

The results align with several findings in the literature. The improvement of NAS over manual design (Atlas & Devi, 2025; Salmani Pour Avval et al., 2025) is confirmed across multiple datasets. The cross-validation robustness of heterogeneous ensembles (Sedek, 2025; Li et al., 2025) is demonstrated through lower standard deviations and higher mean performance. The value of SHAP-based explainability (Garouani et al., 2025; Latent SHAP, 2025) is illustrated through consistent interpretation of model behavior across architectures.

However, some results deviate from expectations. The Wine dataset's manual NN perfect performance suggests that simple architectures can be highly effective for certain problems, challenging the assumption that more complex models are always superior. The heterogeneous ensemble's lower test performance on Wine, despite superior cross-validation, highlights the importance of evaluating models using multiple metrics and validation strategies.

7. Limitations

Several limitations should be acknowledged. First, the datasets used are relatively small and well-characterized; validation on larger, more complex datasets (e.g., image or text data) is needed to fully evaluate the framework's capabilities. Second, the NAS search space was limited to four configurations per dataset; broader search spaces may yield better architectures and more diverse patterns in SHAP analysis. Third, the manual neural network's poor performance on Iris may reflect suboptimal hyperparameter tuning rather than inherent model limitations. Fourth, the computational efficiency analysis did not include inference time, which may be important for real-time applications. Fifth, the SHAP analysis focused primarily on Iris; extending XAI to all datasets and ensemble members would provide more comprehensive insights.

V. CONCLUSION AND RECOMMENDATION

This study successfully developed and evaluated a comprehensive framework integrating Neural Architecture Search, homogeneous and heterogeneous ensemble learning, meta learning via stacked generalization, and Explainable AI using SHAP. The framework was tested on four benchmark datasets: Iris, Wine, Breast Cancer, and Digits. The results demonstrate that NAS can automatically discover architectures matching or exceeding manual designs, with the best NAS achieving 96.67%, 97.22%, 97.37%, and 98.06% accuracy on Iris, Wine, Breast Cancer, and Digits respectively.

The heterogeneous ensemble with meta-learning demonstrated competitive test set performance (96.67% on Iris, 94.44% on Wine, 96.49% on Breast Cancer, 98.61% on Digits) and superior cross-validation robustness, achieving 95.33%, 98.87%, 97.72%, and 98.05% mean accuracy with lower standard deviations compared to the best NAS models. This confirms the value of ensemble methods and meta-learning for improving model generalization and stability.

Computational efficiency analysis revealed trade-offs between model complexity and performance. NAS-discovered architectures are generally more parameter-efficient, with up to 82% parameter reduction compared to manual designs. The heterogeneous ensemble incurs higher computational costs (3-8x longer training times) but offers improved robustness, particularly on complex datasets.

SHAP analysis revealed consistent feature importance patterns across architectures for the Iris dataset, with a single feature dominating predictions. This consistency simplifies interpretability but also highlights the need for more diverse feature utilization on complex problems.

Based on these findings, several recommendations for future research are proposed. First, the framework should be validated on larger, more complex datasets (e.g., image, text, time series) to assess scalability and generalizability. Second, the NAS search space should be expanded to include convolutional layers, residual connections, and attention mechanisms. Third, more advanced meta learning algorithms, such as Model Agnostic Meta Learning (MAML), could be integrated to improve adaptation. Fourth, dynamic ensemble selection based on input characteristics could enhance performance. Fifth, uncertainty quantification should be incorporated to support decision making in high stakes applications. Sixth, the XAI analysis should be extended to include SHAP dependence plots and force plots for deeper insights across all datasets. Seventh, inference time and memory usage should be analyzed for real-time applications. Eighth, sensitivity analysis on hyperparameter choices would provide more robust guidelines for practitioners. Finally, continual learning scenarios should be explored to evaluate the framework's adaptability over time.

REFERENCES

Atlas, L., & Devi, A. S. (2025). Neural Architecture Search (NAS): A survey of methods, challenges, and future directions. Semantic Scholar, September 2025.

Chen, Y., et al. (2025). AgenticRS-EnsNAS: Ensemble-decoupled self-evolving architecture search. arXiv:2603.20014.

Ensemble Learning: Importance, applications, challenges, and future work. (2025). IEEE Xplore, March 2025.

ExplaiLiver+. (2025). Multi-stage stacking framework with SHAP-based interpretability for clinical prediction. Frontiers, 16.

Foundations and innovations in data fusion and ensemble learning for effective consensus. (2025). DOAJ, February 2025.

Garouani, M., Barhrhouj, A., & Teste, O. (2025). XStacking: An effective and inherently explainable framework for stacked ensemble learning. Information Fusion, 124, 103358.

IELF (Interpretable Ensemble Learning Framework). (2025). Heart disease prediction with Explainable Boosting Machines and SHAP. Frontiers, 16.

Latent SHAP. (2025). Toward practical human-interpretable explanations. Machine Learning, August 2025.

Li, J., et al. (2025). Few shot crop disease recognition using sequence-weighted ensemble model agnostic meta learning. Frontiers in Plant Science, 16, 1615873.

Lin, Y., et al. (2021). LENAS: Learning based Neural Architecture Search and Ensemble for 3D Radiotherapy Dose Prediction. arXiv:2106.06733.

Meta ensemble learning. (2025). Multi headed models for few shot problems. ACAI, December 2025.

MetaStackD. (2025). A robust meta learning based deep ensemble model for prediction of sensors battery life. Scientific Reports, April 2025.

Rong, J., et al. (2021). Across Task Neural Architecture Search via Meta Learning. arXiv:2110.05842.

Salmani Pour Avval, S., Eskue, N. D., & Groves, R. M. (2025). Systematic review on neural architecture search. Artificial Intelligence Review, 58(3).

Sedek, A. (2025). Dynamic Meta Learning for Adaptive XGBoost Neural Ensembles. arXiv:2510.03301.

SEENAS (Semi Dynamic Ensemble for Evolutionary NAS). (2025). IEEE Xplore, December 2025.

SHAP based Explainable AI for dropout prediction. (2025). IEEE Xplore, December 2025.

Verma, S., Kumar, P., & Singh, J. P. (2025). MLP GNAS: Meta learning based predictor assisted Genetic Neural Architecture Search system. Applied Soft Computing, 112527.

BossNAS: Block wisely Self supervised Neural Architecture Search. (2025). opus.lib.uts.edu.au, July 2025.

Adaptive deep meta learning ensembles for robust daily carbon emission forecasting. (2025). ScienceDirect, October 2025.

Heterogeneous ensemble learning: modified ConvNextTiny for detecting molecular expression of breast cancer. (2025). Frontiers, June 2025.

An ensemble heterogeneous transformer model for an effective diagnosis of multiple plant diseases. (2025). Frontiers, November 2025.

Sensitivity Analysis of Neural Network Ensembles to Data Quality Issues. (2025). IEEE Xplore, September 2025.

A meta learning ensemble framework for robust and interpretable prediction of emergency medical services demand. (2025). PubMed, December 2025.

One Shot Neural Ensemble Architecture Search by Diversity Guided Search Space Shrinking. (2025). arXiv.

Multi headed Neural Ensemble Search. (2025). arXiv.

EvoREP: Evolving Reliable Ensemble of Proxies for Zero Shot Neural Architecture Search. (2025). IEEE Xplore, December 2025.

GreenFactory: Ensembling Zero Cost Proxies to Estimate Performance of Neural Networks. (2025). scirate.com, May 2025.

MedNNS: Supernet based Medical Task Adaptive Neural Network Search. (2025). MICCAI 2025, pp. 448 458.

AutoMTNAS: Automated meta reinforcement learning on graph tokenization for graph neural architecture search. (2025). Knowledge Based Systems, 310.

Fast Heterogeneous Graph Neural Network Generation via Meta Contrastive Learning. (2025). ScienceDirect, August 2025.

XStacking: Explanation Guided Stacked Ensemble Learning. (2025). Information Fusion, 124.

Metaheuristically Enhanced ANN Based Intrusion Detection System with Explainable AI Integration. (2025). IEEE Xplore, November 2025.

WISE: A Meta Learning Approach for Weighted Instance Specific Ensemble Prediction. (2025). ACAI, December 2025.

Optimizing Stacked Ensemble Machine Learning Models for Accurate Wildfire Severity Mapping. (2025). OUCI.

A novel stacking based ensemble learning framework for thermal error prediction. (2025). ScienceDirect, October 2025.

An Auto Encoder Enhanced Stacked Ensemble for Intrusion Detection in Healthcare Networks. (2025). ScienceDirect, September 2025.

Optuna Optimized Meta Learner and Ensemble Learning Models for Student Performance Prediction. (2025). ScienceDirect, November 2025.

Stacking Ensemble with Boosting and Hyperparameter Optimization for Breast Cancer Classification. (2025). DTIC.

Research progress and future prospects of neural architecture search. (2025). JUSTC, May 2025.

Neural Architecture Search for global multi step Forecasting of Energy Production Time Series. (2025). scirate.com, November 2025.

Improving Novelty Search with a Surrogate Model and Accuracy Objectives to Build High Performing Ensembles. (2025). SN Computer Science, July 2025.

Pre NAS: Efficient Neural Architecture Search. (2025). EmergentMind, September 2025.

Explainable AI using ensemble machine learning with integrated SHAP Borda approach for soil liquefaction estimation. (2025). EBSCO, September 2025.

Integrating Bayesian Optimization into Ensemble Logistic Regression for Explainable AI Based Customer Behavior Analysis. (2025). Semantic Scholar, October 2025.