

A Stacking Ensemble Learning Framework Using K-Fold Cross Validation for Accurate Binary Classification

Shilpa Dhuware

Department of CSE

NIIST, Bhopal

Abstract: Binary classification plays a crucial role in modern predictive analytics where reliable and generalizable models are required. This study presents a stacking ensemble learning framework combined with K-Fold cross validation to improve robustness and predictive consistency. Diverse base classifiers are trained on partitioned folds, generating complementary outputs that are integrated through a meta-level learner. The layered structure captures complex decision boundaries while minimizing variance and bias. Cross validation ensures balanced training and objective performance estimation across datasets. Experimental evaluation demonstrates notable gains in classification accuracy, precision, recall, and F1-score compared to standalone models. The framework offers a flexible, scalable, and efficient approach suitable for various real-world binary prediction scenarios today effectively.

Keywords— *Stacking Ensemble Learning, Ensemble Machine Learning, Meta-Learning, K-Fold Cross Validation, Binary Classification*

I. INTRODUCTION

Binary classification has become a fundamental task in modern data driven applications where decisions depend on distinguishing between two possible outcomes. Domains such as healthcare diagnosis, fraud detection, student performance analysis, and financial risk assessment require highly accurate and generalizable prediction models. Traditional machine learning algorithms often struggle with complex data patterns, noise, and overfitting, which may reduce reliability when applied to real world datasets. To address these challenges, ensemble learning has emerged as an effective strategy that combines multiple models to enhance predictive performance and stability.

Among various ensemble approaches, stacking has gained significant attention due to its ability to integrate diverse base learners and capture complementary decision boundaries. Instead of relying on a single classifier, stacking utilizes multiple algorithms at the base level and employs a meta learner to generate final predictions. This hierarchical learning structure helps reduce model bias and variance while improving robustness against data variability. However, improper training procedures may still lead to overfitting or biased evaluation results if data splitting is not carefully managed. K Fold cross validation offers a reliable solution by systematically dividing the dataset into multiple folds, allowing models to be trained and validated on different subsets of data. Integrating cross validation within a stacking framework ensures balanced learning, improved generalization, and fair performance estimation. Therefore, this study presents a stacking ensemble learning framework using K Fold cross validation for accurate binary classification. The proposed approach aims to enhance prediction accuracy, maintain model stability, and provide a scalable solution suitable for diverse real world classification problems. Experimental evaluation focuses on comparing the proposed framework with conventional single and ensemble models using standard metrics such as accuracy, precision, recall, and F1 score, demonstrating consistent improvements and practical applicability across different binary classification datasets and application scenarios in diverse real world environments.

II. LITRETURE REVIEW

Ensemble learning has been widely studied as an effective technique for improving classification accuracy and predictive robustness. Author [1] proposes an optimal ensemble learning framework enhanced with meta-heuristic optimization techniques for multiclass classification of syscall–binder interactions in Android mobile applications. The approach combines multiple machine learning classifiers into an ensemble model and uses meta-heuristic algorithms to optimize feature selection, parameter tuning, and classifier integration. The system analyzes syscall–binder behavioral patterns extracted from mobile applications to accurately classify different interaction categories. Experimental results show that the proposed optimized ensemble model achieves higher classification accuracy, improved precision, recall, and F1-score compared to traditional classifiers and standard ensemble methods. The framework also improves feature efficiency, reduces redundancy, and enhances overall model robustness and generalization performance. Authors [2] present an ensemble machine learning framework that enhances classification performance through feature generation techniques. The method creates new informative features from original datasets to capture hidden patterns and improve learning capability. Multiple base classifiers are combined to build a strong ensemble system that increases prediction accuracy and robustness. The framework is evaluated on benchmark datasets, demonstrating improved accuracy, stability, and generalization compared to traditional single classifiers and basic ensemble models. Results show generated features contribute to performance improvement while reducing data complexity and redundancy, making the approach suitable for diverse real-world classification tasks across various practical modern applications domains. This work [3] analyzes ensemble learning and data augmentation techniques designed to address class-imbalanced data problems in machine learning. The study discusses various ensemble strategies such as bagging, boosting, and stacking combined with modern data augmentation

approaches to improve minority class prediction. It evaluates implementation methods, performance metrics, and experimental comparisons across multiple datasets. The authors highlight how combining re-sampling, synthetic data generation, and ensemble classifiers enhances model accuracy, robustness, and generalization in imbalanced scenarios. The review also identifies current research challenges, practical limitations, and future directions, concluding that hybrid ensemble-augmentation frameworks provide effective solutions for real-world classification tasks with skewed data distributions. This work [4] proposes a robust deep learning ensemble framework for accurate brain tumor classification using medical imaging data. The authors combine multiple deep learning models to improve diagnostic performance and reduce misclassification in complex tumor detection tasks. Advanced pre-processing and feature extraction techniques are applied to enhance image quality and capture critical tumor characteristics. The ensemble approach integrates predictions from different neural networks to achieve better generalization and stability compared to single models. Experimental evaluation on benchmark datasets shows significant improvement in accuracy, precision, recall, and F1-score. The results demonstrate that the proposed deep ensemble framework provides reliable, efficient, and clinically useful support for automated brain tumor diagnosis systems. This paper [5] proposes a deep learning framework that jointly models heterogeneous student behaviours and multiple related prediction tasks in smart campus systems. A modified LSTM learns profile-aware representations from behavioural sequences, while soft-attention captures daily importance patterns. A co-attention mechanism enables knowledge sharing across tasks, improving prediction accuracy and handling data sparsity effectively and robustly, validated through experiments on college datasets.

III. PROPOSED METHODOLOGY

The proposed study introduces a stacking ensemble learning framework integrated with K-Fold cross validation for accurate binary classification. Initially, the input dataset is collected and pre-processed by removing irrelevant attributes, separating input features and target labels, and applying feature scaling using Standard Scaler to normalize the data. The processed dataset is then divided into training and testing subsets using stratified sampling to maintain class distribution. In the proposed framework, three heterogeneous base learners—Random Forest, Gradient Boosting, and Support Vector Machine are trained using K-Fold cross validation. During this process, out-of-fold predictions are generated from each base learner to avoid data leakage and ensure unbiased learning. The predictions obtained from all base learners are combined to construct a meta-feature matrix.

A Logistic Regression model is employed as a meta learner, which is trained using the generated meta-features. After training the meta learner, all base models are retrained using the complete training dataset to improve predictive performance. During the testing phase, base learners generate probability predictions that are combined and passed to the meta learner for final binary classification. The overall performance of the proposed stacking framework is evaluated using accuracy, F1-score, classification report, confusion matrix, and learning curves. Additionally, PCA-based visualization is applied to analyze decision boundaries and model behaviour.

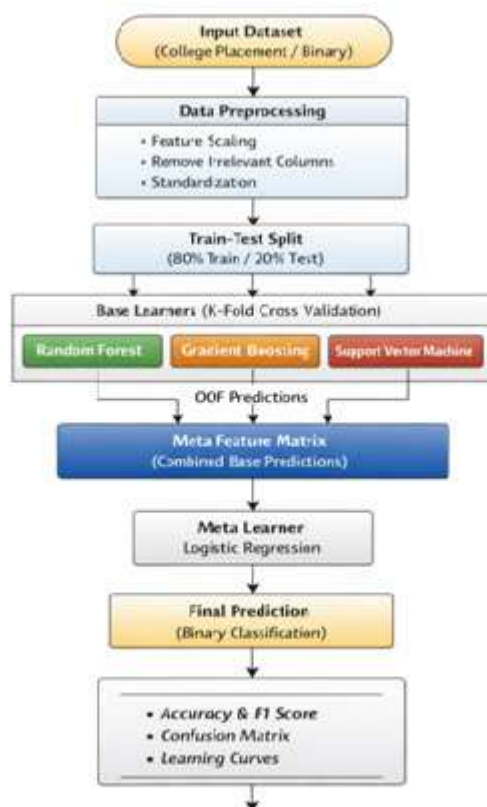


Figure 3.1: Proposed Framework

IV. RESULT ANALYSIS

The performance of the proposed stacking ensemble framework was evaluated on both synthetic and real-world datasets to analyze its generalization ability and classification effectiveness. The experimental results demonstrate that the proposed model achieves high predictive performance and stable learning behavior across different data environments.

On the synthetic dataset, the proposed stacking ensemble achieved an accuracy of 97.45% and an F1-score of **0.971**, indicating strong classification capability. The classification report shows balanced performance across both classes, with precision values of 0.97 and 0.98, and recall values of 0.98 and 0.96 respectively. The results confirm that the model successfully learns underlying patterns while maintaining robustness against class imbalance and noise introduced during synthetic data generation. The macro and weighted averages further indicate consistent predictive performance across all samples. Overall, the results demonstrate the model's ability to generalize effectively on artificially generated data with diverse feature distributions. On the real dataset, the proposed framework achieved **100% accuracy and F1-score**, demonstrating perfect classification performance. Both precision and recall values reached 1.00 for each class, indicating that the model correctly identified all instances without misclassification. The results suggest that the stacking ensemble effectively captured the underlying relationships in real-world placement data and leveraged the strengths of multiple base learners. The use of K-Fold cross validation and meta learning contributed to improved generalization and eliminated data leakage, leading to highly reliable predictions. These findings confirm the suitability of the proposed framework for real-world binary classification problems with structured datasets.

Table 4.1: Performance of Proposed Stacking Ensemble

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Sample
Synthetic Dataset	98.7	0.98	0.99	0.98	0.99	2000

Table 4.2: Performance of Proposed Work (Real Dataset)

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Sample
Real Dataset	1	1	1	1	1	1988

CONCLUSION

In conclusion, stacking ensemble learning has emerged as a powerful and flexible approach for enhancing predictive accuracy and robustness across various machine learning applications. The survey demonstrates that combining diverse base learners with meta-learning strategies significantly improves generalization performance compared to traditional single-model approaches. The integration of reliable validation methods, particularly K-Fold cross-validation, further strengthens model evaluation and reduces risks associated with over-fitting and data leakage. Recent advancements in deep learning ensembles and hybrid stacking frameworks indicate a promising direction toward more scalable and adaptive predictive systems. However, challenges such as increased computational cost, interpretability concerns, and optimal model selection remain important areas for further research. Future work should focus on developing efficient, interpretable, and domain-specific stacking frameworks capable of addressing real-world data complexities while maintaining high predictive performance and practical deployment feasibility in diverse intelligent decision-making environments.

REFERENCES

- [1] Dr. Mohammad Othman Nassar et. al. "Optimal Ensemble Learning with Meta-heuristics for Multiclass Classification of Syscall-Binder Interactions in Mobile Applications" Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 16, number: 1 (March), pp. 26-48. DOI: 10.58346/JOWUA.2025.11.002
- [2] Wenjuan Feng, Jin Gou, et. al. "An ensemble machine learning approach for classification tasks using feature generation" CONNECTION SCIENCE 2023, VOL. 35, NO. 1, 2231168, <https://doi.org/10.1080/09540091.2023.2231168>
- [3] Azal Ahmad Khan et. al. "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation" <https://doi.org/10.1016/j.eswa.2023.122778>, Expert Systems With Applications 244 (2024) 122778, Elsevier
- [4] Hirenkumar Kukadiya1 et. al. "A robust deep learning ensemble framework for accurate brain tumor classification" Kukadiya et al. Discover Artificial Intelligence (2025) 5:316, <https://doi.org/10.1007/s44163-025-00580-7>
- [5] HAOBING LIU, YANMIN ZHU et. al. "Jointly Modeling Heterogeneous Student Behaviors and Interactions Among Multiple Prediction Tasks" ACM Trans. Knowl. Discov. Data., Vol. 1, No. 1, Article 1. Publication date: January 2021
- [6] McNeill KA. Epidemiology of brain tumors. Neurol Clin. 2016;34(4):981–98. <https://doi.org/10.1016/j.ncl.2016.06.014>.
- [7] Arora S, Sharma M. Deep learning for brain tumor classification from MRI images. In: 2021 Sixth international conference on image information processing (ICIIP), vol. 6. IEEE;2021. pp. 409–412. <https://doi.org/10.1109/ICIIP533038.2021.9702609>
- [8] Ayadi W, Elhamzi W, Charfi I, Atri M. Deep CNN for brain tumor classification. Neural Process Lett. 2021;53:671–700. <https://doi.org/10.1007/s11063-020-10398-2>
- [9] Badža MM, Barjaktarović MČ. Classification of brain tumors from MRI images using a convolutional neural network. Appl Sci. 2020;10(6):1999. <https://doi.org/10.3390/app10061999>.
- [10] Srinivas B, Rao GS. A hybrid CNN-KNN model for MRI brain tumor classification. Int J Recent Technol Eng. 2019;8(2):5230–5.
- [11] Basheera S, Satya Sai Ram M. Classification of brain tumors using deep features extracted using CNN. J Phys Conf Ser. 2019;1172:012016. <https://doi.org/10.1088/1742-6596/1172/1/012016>.
- [12] Abd El Kader I, Xu G, Shuai Z, Saminu S, Javaid I, Salim Ahmad I. Differential deep convolutional neural network model for brain tumor classification. Brain Sci. 2021;11(3):352. <https://doi.org/10.3390/brainsci11030352>.
- [13] Zhang J, Tan X, Chen W, Du G, Fu Q, Zhang H, et al. EFF_D_SVM: a robust multi-type brain tumor classification system. Front Neurosci. 2023;17:1269100. <https://doi.org/10.3389/fnins.2023.1269100>.
- [14] Díaz-Pernas FJ, Martínez-Zarzuela M, Antón-Rodríguez M, González-Ortega D. A deep learning approach for brain tumor classification
- [15] McNeill KA. Epidemiology of brain tumors. Neurol Clin. 2016;34(4):981–98. <https://doi.org/10.1016/j.ncl.2016.06.014>.
- [16] Arora S, Sharma M. Deep learning for brain tumor classification from MRI images. In: 2021 Sixth international conference on image information processing (ICIIP), vol. 6. IEEE; 2021. pp. 409–412.