INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Weighted Model Fusion for Imbalanced Learning

S Josephine Theresa[1]*

1 Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Trichy, Tamil Nadu, India

OPEN ACCESS

Received: 15/05/2025

Accepted: 25/05/2025

Published: 16/06/2025

Citation: Theresa SJ (2025) Weighted Model Fusion for Imbalanced Learning. Indian Journal of Science and Technology 18(23): 1818-1824. https://doi.org/10.17485/IJST/v18i23.904

* **Corresponding author**.

rschlrjosephine@yahoo.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop a Weighted Multi-model Ensemble (WME) to improve binary and multiclass data predictions, particularly in handling imbalanced datasets. The model aims to achieve high performance metrics, such as precision and recall, while minimizing false positive rates. Additionally, the study seeks to explore better handling mechanisms to enhance prediction accuracy further. **Methods:** The methodology involves a two-phase approach for the Weighted Multi-Model Ensemble (WME). The first phase includes data preprocessing, segregating training and test data, and training models like Decision Tree, Random Forest, and Extra Tree Classifier to determine their accuracy and weights. The second phase applies the test data to these trained models, calculates probability levels, and makes final predictions based on the weighted probabilities. **Results:** The results indicate that the WME model achieved high True Positive Rate (TPR) levels, nearing 100%, and low False Positive Rate (FPR), demonstrating effective performance on the Brazilian bank dataset. Precision and recall values exceeded 90%, confirming the model's capability in identifying minority classes in moderately imbalanced data. **Findings**: The findings reveal that the Weighted Multi-Model Ensemble (WME) effectively handles data imbalance, achieving high precision and recall levels, particularly in the Brazilian bank dataset and KDD Cup 99 data. **Novelty:** The novelty of the study lies in the development of the Weighted Multi-Model Ensemble (WME), which integrates multiple classifiers with a unique weighting mechanism to enhance prediction accuracy on imbalanced datasets. This approach not only improves performance metrics like precision and recall but also addresses the challenges of minority class identification more effectively than existing models.

Keywords: Data imbalance; Ensemble modeling; Weighted prediction; Probability based analysis; Classification

1 Introduction

Data imbalance has become one of the major factors affecting the performance of classifier models[(1)]. Existing machine learning algorithms are susceptible to data imbalance, as they usually consider data to be balanced before processing[(2)]. The

negative impact of data imbalance is further aggravated by noise, data insufficiency, and class overlap issues contained in the data. Data imbalance is a common issue that could be observed in several real-time applications like fraud detection, intrusion detection systems, medical systems used in anomaly detection, and cheminformatics[3]. Hence, it becomes mandatory to develop models that can effectively work on imbalanced data to provide accurate predictions. The issue of data imbalance is generally very apparent in binary classification problems[4]. Data is considered to be imbalanced if one of its classes consists of a large number of instances, while the other existing classes contains a very low number of instances. The class that has the largest number of instances is the majority class, while all the other classes are known as the minority classes[5]. In most of the real-time scenarios, minority classes are the classes of importance. For example, identifying a record with disease is much more significant compared to identifying a normal record. Further, misclassifying a minority instance tends to be costly compared to misclassifying a majority instance[6].

Several techniques are available to handle the data imbalance[7]. The techniques can be classified as data-driven techniques and domain-driven models[8]. Data driven techniques are focused on modifying data to handle the imbalance. Oversampling and under sampling are the techniques used towards handling imbalance[9]. These techniques tend to modify the class distribution to handle the data imbalance. Domain driven models handle imbalance based on the data availability and data quality in the domain[10]. Hybrid techniques are also available in the literature, which uses a combination of oversampling and under sampling techniques[11]. This work presents a domain based solution to data imbalance by providing an integrated framework that can effectively handle data imbalance.

Handling data imbalance is one of the major requirements in several domains. Although several works have been proposed to provide effective results, they are highly data specific. A knowledge distillation-based framework for classifying of imbalanced data has been proposed[12]. This model has been designed as an information extraction technique and performs event detection. It operates on actual data and handles data imbalance to reduce identification errors. It uses a baseline model for initial analysis and a teacher network that learns from the baseline model. A credit card fraud detection model that handles data imbalance has been presented[13]. This work proposes a window-based bagging model to ensure data balancing. The work also uses an incremental learning mechanism that aids in keeping the model up to date. Cost sensitivity is added to the model to ensure that the minority class is provided with increased significance during the training process. Other credit card fraud detection strategies focused on handling data imbalance include works by Zhang et al.[14], Dantas et al.[15], and Alenzi et al.[16].

An energy system-based framework for load prediction and imbalance handling has been proposed[17]. This work proposes a clustering decision tree model for load prediction. Imbalance is handled by sampling the training data into various levels and testing for efficiency. A clustering-based technique for handling class imbalance has been proposed[18]. An oversampling-based technique to handle data imbalance has been proposed by Fu et al.[19]. This work uses affinity propagation to ensure balanced class distributions. Data uncertainty is handled using evidential reasoning-based classifier models. Lim et al. proposed a cluster-based oversampling technique to handle data imbalance[20]. This work is based on creating an evolutionary ensemble modeling technique to perform synthetic oversampling. A bagging-based evolutionary model for imbalanced datasets has been proposed by Roshan et al.[20]. This work uses evolutionary multi-objective optimization to handle data imbalance.

2 Methodology

2.1 Weighted Multi-model Ensemble

Imbalance is a major issue that occurs in several real-time domains. Although several model-based solutions exist for handling data imbalance, they are strictly data-based. Hence, the model fails even if there is a slight variation in data. However, data imbalance is a general issue in several varied domains and requires a generic rather than a data-specific solution. This work presents a generic model that can be used in any domain experiencing data imbalance. Further, the model has also been designed to handle binary and multiclass data. The proposed Weighted Multi-Model Ensemble (WME) comprises three major sections. The data preprocessing and segregation section processes the data and makes it model-ready; the multiple model creation section creates varied models for training, and the weighted probability-based prediction section makes the final predictions based on the weights determined from models. The algorithm for the WME model is provided below.

2.2 Data Preprocessing and Segregation

This work uses the Brazilian bank data and the KDD cup 99 data. Brazilian bank data is a financial fraud detection data. It is completely anonymized data and is a binary classification data. All features in the Brazilian bank data are represented in numerical format. Hence, the data can be directly used by the machine learning models. The KDD cup 99 data is network intrusion detection data. This data is composed of several categorical features. Since machine learning models cannot directly

use these features, they are converted to numerical format using one hot encoding technique. Both datasets are normalized to ensure data consistency. Brazilian bank data is binary class data, and the KDD cup 99 data is multi-class. The proposed model has been designed to handle both binary and multiclass data. Hence, class-based changes are not performed. The preprocessed data is divided in the ratio 8:2 for training and testing.

2.3 Multiple Model creation

The model creation process follows data preprocessing. This work creates a multi-model-based architecture for handling data imbalance. Three varied models are created for the training process. The architecture uses a decision tree, a random forest, and an extra tree for the model creation. A decision tree is a tree-based classifier model. The model uses entropy values of the features while creating the tree. The root node is composed of the feature exhibiting the highest entropy. Features with lower entropy levels occupy the next level in the tree. The leaf nodes represent the final class. Random forest is an ensemble-based modeling technique. A bagging ensemble uses multiple decision trees to create the prediction model. Each decision tree is provided with a distinct part of the data. Every tree is trained on a different set of data. Hence, every tree creates different rules. The resultant prediction is a combination of rules from all the trees and regulations from all the trees. This final prediction has been observed to be better than a prediction from a single tree. The extra trees classifier is also an ensemble-based model. The difference is that the extra trees classifier follows a sampling without replacement strategy. Node split for trees is performed at random during the tree creation process. All three models are trained using all the training data. Data split for the trees is performed internally. The training data is again passed to the model, and training-based predictions are obtained. Accuracy level is determined for each model. The significance of the model is determined based on the accuracy levels that were obtained. Accuracy levels vary between 0 and 1. The weight of a model is determined based on the accuracy levels. The weight of a model m is given by,

$$W_m = \frac{acc_m}{\sum_{i=1}^{n} acc_i} \tag{1}$$

Where acc_m is the accuracy of the model m, and n is the total number of models considered for analysis.

2.4 Weighted Probability Based Final Predictions

Final predictions are based on weighted probability values. Test data is passed to the multiple models. The probability of predictions is obtained from the models rather than receiving the actual predictions. The Brazilian bank data is a binary classification problem. Hence, the resultant probabilities contain two values, each depicting the probability level for a single class. Similarly, the KDD cup 99 data is a multiclass problem that contains five probability values for each instance. The probability values obtained from each model are multiplied with the model weights to get the weighted probability levels, for example, i. The weighted probability of an instance i for a model x is given by,

$$WProb_{ix} = W_x * Prob_i \tag{2}$$

Where W_x is the weight of the model x, and Probi is the probability matrix of the instance i.

The weighted probability levels for each instance obtained from the different models are added together to form the final probability set. Final probability set is given by,

$$FProb_i = \sum_{x=1}^{n} WProb_{ix} \tag{3}$$

Where WProbix is the weighted probability set for instance i and model x.

Algorithm for WME:
Input:Preprocessed training dataset
Output:Final predictions based on weighted probabilities

1. Data Preprocessing to ensure data consistency
2. Segregate training and test data
3. Use the training data to train Decision Tree, Random Forest and Extra Tree Classifier
4. Perform predictions using trained models over the training data
5. Identify the accuracy value for the three models
6. $Weight\ of\ a\ model\ (W_m) = \frac{acc_m}{\sum_{i=1}^{n} acc_i}$
7. Apply test data over the trained models to determine the probability levels
8. Multiply probably levels with the corresponding weights
9. Add all the corresponding probability levels
10. Determine the final prediction based on the probability levels

3 Results and Discussion

The weighted multi-model ensemble has been implemented using Python. The Brazilian bank data set and the Knowledge Discovery Database (KDD) cup 99 data set have been used to analyze the model's performance. The Brazilian bank data is a binary classification data set with moderate imbalance. The KDD cup 99 data is a multi-class data set with high imbalance levels. The datasets are divided 80:20, with 80% allocated for training and 20% for testing. This proportion ensures that the model learns from a sufficient amount of data while maintaining a validation set to evaluate generalization performance. Once preprocessed, the datasets are used to train three distinct models Decision Tree, Random Forest, and Extra Trees Classifier. Each model is trained independently using the 80% training data, and accuracy levels are computed to assign weighting factors in the ensemble approach. The results obtained are used to calculate the standard performance metrics like True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), accuracy, precision, and recall.

- True Positive Rate (TPR) – measures the model's ability to correctly classify positive instances.
- False Positive Rate (FPR) – identifies false alarms in classification.
- True Negative Rate (TNR) – validates proper recognition of majority class instances.
- False Negative Rate (FNR) – ensures rare instances are not overlooked.
- Precision & Recall – critical for assessing minority class identification performance.

The Receiver-operating characteristic curve (ROC) plot representing two positive rates and false positive rates of the model On the Brazilian bank data is shown in Figure 1. High TPR levels and low FPR levels are desirable for effective models. The figure shows that the FPR level of WME is very low, representing low false alarm levels. The TPR level of WME is very high and almost 100%. These two factors represent that the WME model can operate on data with moderate imbalance and provide highly effective results.

The PR plot representing the precision and recall values of WME in Brazilian bank data is shown in Figure 2. High precision and recall values are desirable for effectively performing models. The curve shows high precision and recall levels greater than 90%. This supports the fact that the WME model performs well in identifying minority classes over moderately imbalanced data.

The classification report for KDD Cup 99 data is provided in **Except for class 2, all the other classes depict high precision, recall, and F1-Score. Class 2 exhibits slightly reduced recall levels of 5% and F1-Score of 3%. However, the accuracy of prediction remains at 100%. These metrics show the model performs well on moderate to high imbalance levels. However, on very high imbalances (>1000), the model exhibits a slight reduction in performance.**. Except for class 2, all the other classes depict high precision, recall, and F1-Score. Class 2 exhibits slightly reduced recall levels of 5% and F1-Score of 3%. However, the accuracy of prediction remains at 100%. These metrics show the model performs well on moderate to high imbalance levels. However, on very high imbalances (>1000), the model exhibits a slight reduction in performance.

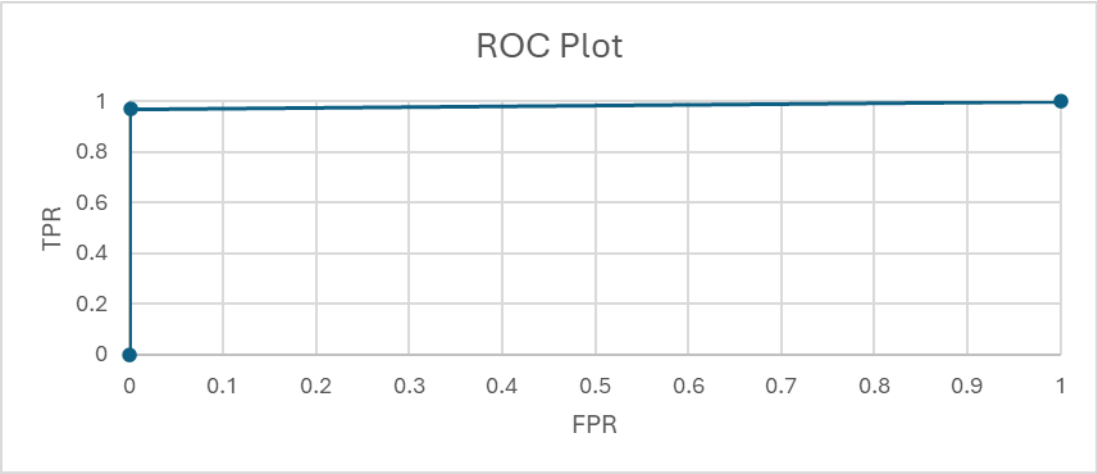

Fig 1. ROC Plot for Brazilian Bank Data

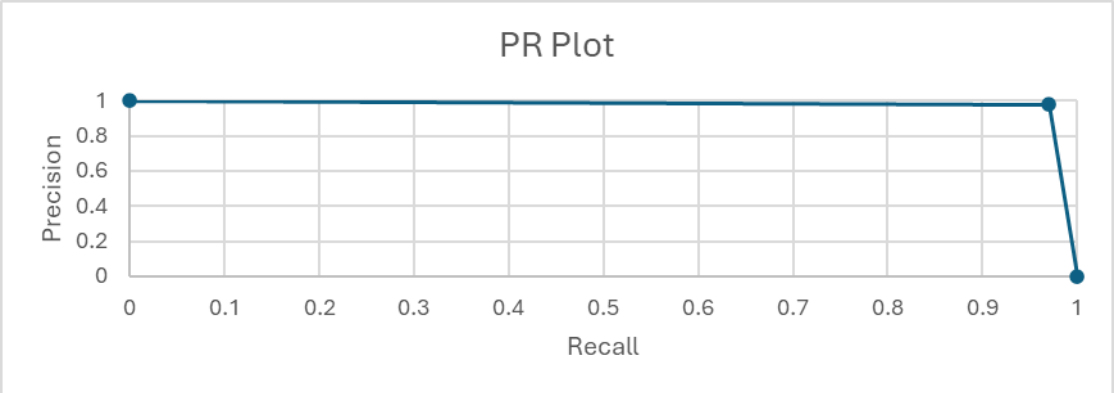

Fig 2. PR Plot for Brazilian Bank Data

Table 1. Classification Report for KDD CUP 99 Data

Class	Precision	Recall	F1-Score	Support
1	1	1	1	13795
2	1	0.95	0.97	19
3	1	1	1	299
4	1	1	1	3490
5	1	1	1	20189
Accuracy			1	37792
Macro Avg	1	0.99	0.99	37792
Weighted Avg	1	1	1	3779

Comparisons are performed by analyzing the prediction levels of the WME model with Traditional Window Bagging (TWB) models, oversampling-based models, and Cost-Sensitive models.

Table 2. Performance Comparison of WME with TWB and Other Existing Approaches

Feature/Metrics	WME(Proposed)	TWB [20]	Oversampling-based Models [19,20]	Cost-Sensitive Learning [14–16]
Handling Imbalanced data	Strong (Works on both Binary & Multiclass)	Moderate (Limited to Window-based balancing)	Good (Relies on class redistribution)	Good (Weighted cost assignments)
True Positive Rate (TPR)	99.98%	69.32%	85-90%	88-94%
FPR	0.4%	14.55%	8-12%	5-10%
TNR	99.60%	85.44%	88-94%	90-95%
FNR	0.02%	30.67%	10-15%	8-12%
Precision	99.98%	95.3%	92-96%	97%
Recall	99.98%	69.32%	87-93%	90-95%
Accuracy	99%	85-90%	85-90%	88-94%
Novelty	Weighted probability-based model using ensembles	Window based bagging approach	Synthetic sampling based	Weighted Cost-Sensitive Learning
Adaptability	High (Generic for all domains)	Moderate (Finance-specific)	Limited	Limited

The WME model achieves 99.98% precision and recall (Table 1, Table 2), surpassing traditional approaches that struggle with minority class identification. Prior Studies' Bagging and Cost-Sensitive Models typically yield 93-97% precision and 88-95% recall, lower than WME.

The Weighted Multi-Model Ensemble (WME) demonstrates significant superiority in handling imbalanced learning compared to existing approaches like Traditional Window Bagging (TWB) [20], Oversampling-based models [19,20], and Cost-Sensitive Learning [14–16]. With 99.98% precision and recall, WME substantially improves minority class identification, outperforming TWB and oversampling models, which typically achieve lower recall rates of 69.32% and 87-93%, respectively. Furthermore, WME exhibits exceptionally low false prediction rates (FPR: 0.004%, FNR: 0.0002%), ensuring better classification confidence than prior models that struggle with false alarms. The model's 99.60% True Negative Rate (TNR) further validates its superior reliability over TWB (85.44% TNR), which suffers from classification errors. Unlike domain-restricted models, WME adapts efficiently across binary (Brazilian bank dataset) and multiclass (KDD Cup 99 dataset) problems, proving highly versatile in real-world applications such as fraud detection and intrusion prevention. Unlike oversampling techniques, its novel weighted probability-based mechanism, which dynamically adjusts classifier weights based on accuracy levels, enhances predictive performance without relying on artificial data modification. This innovation ensures improved imbalance handling while avoiding bias toward majority classes. WME's general applicability across multiple domains sets it apart from previous models that often fail in high-imbalance scenarios, making it a robust and efficient solution for diverse classification problems.

4 Conclusion

The Weighted Multi-Model Ensemble (WME) presents a novel and efficient approach to handling imbalanced learning, achieving superior classification accuracy compared to traditional methods. The model comprises two phases; the initial phase uses multiple models and determines their accuracy levels and weights, and the second phase uses these weights and performs probability-based predictions. Experiments and comparisons indicate high performance at 99% precision levels. The WME model could be observed to exhibit better forecasts in all the metrics than the previous models. The findings show that the WME model exhibits better imbalance handling. The significant advantage of this model is that it can handle any level of imbalance. Further, the model can operate on both binary and multiclass data. The limitation of this model is that it exhibits a slight reduction in the class with a very high imbalance. Further experiments will concentrate on integrating better handling mechanisms to improve the predictions. By combining adaptive weighting mechanisms, hybrid sampling techniques, dynamic threshold adjustments, and meta-learning approaches, WME could further enhance its ability to handle extreme imbalance scenarios. These refinements would increase precision, recall, and robustness, making the model more generalizable across diverse datasets.

References

1) Nayak SM, Rout M. Impact of Different Outlier Handling Techniques on GAN Based Hybrid Bankruptcy Prediction Models. *Indian Journal of Science and Technology*. 2024;17(4):373–385. https://doi.org/10.17485/IJST/v17i4.649.
2) Li Y, Wang Y, Jin J, Chen CLP. Feature fusion-based weighted broad learning system for imbalanced data. 2025;p. 708–712. Available from: https://ieeexplore.ieee.org/document/10968350.
3) Li Y, Wang S, Jin J, Zhu F, Zhao L, Liang J, et al. Imbalanced complemented subspace representation with adaptive weight learning. *Expert Systems with Applications*. 2024;249(Part A). https://doi.org/10.1016/j.eswa.2024.123555.
4) Ye-Bin M, Hyeon-Woo N, Choi W, Kim N, Kwak S, Oh TH. SYNAuG: Exploiting synthetic data for data imbalance problems. *Pattern Recognition Letters*. 2025;192:115–121. https://doi.org/10.1016/j.patrec.2025.04.013.
5) Goswami R, Garai A, Sadhukhan P, Ghosh P, Chakraborty T. Shape Penalized Decision Forests for Imbalanced Data Classification. *IEEE Access*. 2025;14:1–17. Available from: https://ieeexplore.ieee.org/document/11003049.
6) Liu J. A minority oversampling approach for fault detection with heterogeneous imbalanced data. *Expert Systems With Applications*. 2021 07;184:115492. https://doi.org/10.1016/j.eswa.2021.115492.
7) Govindrajan V. Machine Learning Based Approach for Handling Imbalanced Data for Intrusion Detection in the Cloud Environment. 2025;p. 810–815. Available from: https://ieeexplore.ieee.org/document/10986614.
8) Leng Q, Yi G, Jiao E. Adversarial feature augmentation via transformer-level feature fusion for imbalanced data classification. *Int J Mach Learn & Cyber*. 2025. https://doi.org/10.1007/s13042-025-02619-8.
9) Zhai J, Qi J, Shen C. Binary imbalanced data classification based on diversity oversampling by generative models. *Information Sciences*. 2021;585:313–343. https://doi.org/10.1016/j.ins.2021.11.058.
10) Song D, Xu J, Pang J, Huang H. Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data. *Information Sciences*. 2021;573:222–238. https://doi.org/10.1016/j.ins.2021.05.045.
11) Wu Y, Chen J, Hu L, Xu H, Liang H, Wu J. Omni Fuse: A general modality fusion framework for multi-modality learning on low-quality medical data. *Information Fusion*. 2025;117. https://doi.org/10.1016/j.inffus.2024.102890.
12) Fragkoulis M, Carbone P, Kalavri V, Katsifodimos A. A survey on the evolution of stream processing systems. *The VLDB Journal*. 2023;33(2):507–541. https://doi.org/10.1007/s00778-023-00819-8.
13) Bernardo A, Gomes HM, Montiel J, Pfahringer B, Bifet A, Valle ED. C-SMOTE: continuous synthetic minority oversampling for evolving data streams. 2020;p. 483–492. https://doi.org/10.1109/bigdata50022.2020.9377768.
14) Zhang C, Li J, Zhao Y, Li T, Chen Q, Zhang X. Problem of data imbalance in building energy load prediction: Concept, influence, and solution. *Applied Energy*. 2021;297:117139. https://doi.org/10.1016/j.apenergy.2021.117139.
15) Dantas RM, Firdaus R, Jaleel F, Mata PN, Mata MN, Li G. Systemic Acquired Critique of Credit Card Deception Exposure through Machine Learning. *Journal of Open Innovation Technology Market and Complexity*. 2022;8(4):192. https://doi.org/10.3390/joitmc8040192.
16) Alenzi HZ, O N. Fraud Detection in Credit Cards using Logistic Regression. *International Journal of Advanced Computer Science and Applications*. 2020;11(12). https://doi.org/10.14569/ijacsa.2020.0111265.
17) Wong TT, Tsai HC. Multinomial naïve bayesian classifier with generalized Dirichlet priors for high-dimensional imbalanced data. *Knowledge-Based Systems*. 2021;228:107288. https://doi.org/10.1016/j.knosys.2021.107288.
18) Koziarski M. Potential anchoring for Imbalanced Data Classification. *Pattern Recognition*. 2021;120:108114. https://doi.org/10.1016/j.patcog.2021.108114.
19) Fu C, Zhan Q, Liu W. Evidential reasoning based ensemble classifier for uncertain imbalanced data. *Information Sciences*. 2021;578:378–400. https://doi.org/10.1016/j.ins.2021.07.027.
20) Ruiz-Villafranca S, Roldán-Gómez J, Carrillo-Mondéjar J, Martinez JL, Gañán CH. WFE-Tab: Overcoming limitations of TabPFN in IIoT-MEC environments with a weighted fusion ensemble-TabPFN model for improved IDS performance. *Future Generation Computer Systems*. 2025;166. https://doi.org/10.1016/j.future.2025.107707.