Cross-Validation-Based Hierarchical Decision Tree Framework for Dispersed Data Classification

Benjamin Agyare Addo[1][0000−0003−0618−5221] and Małgorzata Przybyła-Kasperek[1,2][0000−0003−0616−9694]

[1] University of Silesia in Katowice, Institute of Computer Science, Będzińska 39, 41-200 Sosnowiec, Poland
malgorzata.przybyla-kasperek@us.edu.pl, benjamin.addo@us.edu.pl
[2] Constantine the Philosopher University in Nitra, Trieda Andreja Hlinku 1, Nitra, 949 01, Slovakia

Abstract. The proliferation of fragmented, high-dimensional data across independent sources renders centralized classification impractical due to structural and privacy constraints. Existing hierarchical frameworks often worsen these challenges by taking part of the already limited test data for validation, reducing the reliability of final evaluation. This paper introduces a hierarchical decision tree architecture for dispersed sources that removes the need to extract a validation subset from the test data during global training. The method uses a two-level learning strategy. At the local level, decision trees are trained independently on each table using stratified cross-validation, and out-of-fold probability estimates are generated for all training objects to ensure reliable, leakage-free predictions. These vectors represent the predictive behaviour of each local view. At the global level, probability vectors from all sources are concatenated into a unified representation used to train a global decision tree that integrates information across views and produces the final classification. Experimental evaluation on multiclass benchmark datasets with varying levels of dispersion shows that the proposed method achieves performance comparable to other hierarchical and ensemble approaches designed for distributed data. The comparison included methods that train separate local classifiers and combine their outputs at the decision level. The results demonstrate that the hierarchical strategy based on cross–validation makes more effective use of limited and fragmented information while maintaining a strict separation between training and testing data. As a result, the global classifier is trained in a fully leakage-free manner and remains robust even when individual local tables contain only a small or highly uneven set of features.

Keywords: Dispersed Data Classification · Stacking · Local and Global Models · Ensemble Learning.

1 Introduction

Machine learning has become indispensable for high-stakes predictive analytics, particularly in domains such as smart agriculture where complex, multi-modal

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_7

feature sets are used to classify plant health. However, the increasing fragmentation of real-world data across disparate sensor nodes makes traditional centralized learning impractical due to privacy constraints and communication latency [15]. This has necessitated a shift toward decentralized paradigms, most notably Federated Learning (FL) [1, 3, 6] and Ensemble Learning (EL) [2, 12, 13], which allow for collaborative model training or fusion without requiring raw data aggregation [9, 10].

Classification in these dispersed environments can be effectively achieved by constructing a global model that aggregates class probability vectors generated by independent local models. This architecture treats local diagnostic confidences as high-level features, significantly reducing communication overhead while maintaining data privacy and accommodating heterogeneous sensor formats. Prior research has explored this via decision trees with bagging [8, 7] or K-nearest neighbor algorithms [4]. Such approaches are particularly valuable in "Small Data" scenarios where local sampling is essential to derive meaningful results from limited samples [14].

The core contribution of this paper is a structured hierarchical learning framework specifically tailored for fragmented data environments. Unlike existing stacking strategies that often require centralized data access or an explicit validation split of the test data for higher-level training, our approach utilizes a cross-validation-driven strategy to generate out-of-fold probability vectors. This ensures that the global decision tree is trained in a strictly leakage-free manner, preserving the integrity of the test set exclusively for final evaluation. By eliminating the need to partition limited datasets for validation, the proposed framework addresses critical constraints related to data locality and evaluation reliability in multi-source classification.

Building upon the dual-level architectures introduced in [8], this study investigates a methodology that employs cross-validation at the local level to produce robust probability features for a second-level global decision tree. This design aims to enhance classification performance in high-dimensional, multiclass settings while ensuring that every available data point is utilized effectively without compromising experimental validity. The remainder of the paper is organized as follows: Section 2 details the proposed hierarchical model; Section 3 describes the experimental setup and datasets; Section 4 discusses the results; and Section 5 provides concluding remarks and future research directions.

2 Model

This research addresses a classification problem where data is distributed across multiple autonomous sources rather than a centralized repository. Each source provides a partial and potentially heterogeneous perspective on a single underlying phenomenon. We define a collection of dispersed local decision tables $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$, where each source i is represented as $D_i = (U_i, A_i, d)$. Here, U_i denotes the local object set, A_i represents the conditional attributes specific to that source, and d is a shared decision attribute. These tables are

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_7

typically maintained by independent entities—such as distinct medical clinics or financial branches—meaning that while the attribute sets A_i may differ or partially overlap, the objective d remains consistent across the entire system.

Learning in such a fragmented environment presents significant challenges, primarily due to the potential for inter-source inconsistencies. Because each agent only observes a subset of the global attribute space, the same object might be associated with conflicting decision values across different tables. This necessitates a robust integration strategy that can resolve these contradictions and extract a coherent global signal without requiring the movement of raw, sensitive data to a central location.

To resolve these complexities, we propose a two-level hierarchical decision tree framework. The first level operates locally and independently on each decision table. For every source, we construct an ensemble of decision trees using a localized training strategy. These models do not merely output a hard classification; instead, they generate a probability vector for every object. The dimensionality of this vector corresponds to the number of decision classes, where each component reflects the estimated likelihood of an object belonging to a specific class based on the evidence available at that specific local site.

The second level of the framework acts as an integrator for the information synthesized by the local agents. For each object x, the probability vectors generated by all n sources are concatenated into a unified global feature representation. This resultant vector captures the collective, multi-perspective evidence provided by the distributed system. A global decision tree is then trained on these concatenated vectors using the true labels. By treating local probabilistic outputs as high-level features, the global model learns the relative reliability of each source and produces the final classification decision.

2.1 Local Level and Out-of-Fold Strategy

A critical requirement for the global model is the ability to generalize to unseen data without falling victim to information leakage. If the global model were trained on the same local predictions used to evaluate the local models, it would likely overfit to local training errors. To prevent this, we implement an out-of-fold (OOF) prediction strategy grounded in K-fold cross-validation. For each local table D_i, the object set U_i is partitioned into K disjoint, stratified folds.

During the training phase, for each fold $k \in \{1, \ldots, K\}$, a local tree $Tree_{i,k}$ is trained on the data contained in the remaining $K-1$ folds. This model is then applied to the held-out fold to generate posterior probabilities $P_{i,l}^{(k)}(x) = \text{predict\_proba}(Tree_{i,k}, x)$. By iterating this process across all K folds, we obtain an out-of-fold probability vector $\hat{P}_i(x)$ for every training object in U_i. This ensures that every feature provided to the global level was generated by a model that did not have access to that specific object during its own training phase.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_7

2.2 Global Level Integration

Once the local processing is complete, the framework transitions to the global stage. The individual out-of-fold probability vectors from each source are concatenated to form a single global representation $V(x) = [\hat{P}_1(x)|\hat{P}_2(x)|\ldots|\hat{P}_n(x)]$. This creates a global feature matrix $\mathbf{S} = \{V(x) \mid x \in \bigcup_{i=1}^{n} U_i\}$ which serves as the input for the global stage of the hierarchy.

The global decision tree, $Tree_{\text{global}}$, is trained using this matrix $\mathbf{S}$ and the corresponding decision labels. The role of the global model is to learn decision rules that effectively combine probabilistic evidence from multiple dispersed sources. During the inference phase, any new object to be classified follows the same pipeline: local models generate probability scores, which are then fused and processed by the global tree to yield the final predicted label $\hat{y}$.

3 Experimental Methodology

The proposed framework was evaluated using three multiclass benchmark datasets from the UC Irvine Machine Learning Repository: Vehicle Silhouettes, Soybean Large, and Lymphography. These datasets were selected for their high-dimensional feature spaces and varying levels of class complexity. While these datasets are originally centralized, we adapted them to simulate the dispersed data scenarios common in real-world distributed systems.

For the Vehicle Silhouettes and Lymphography datasets, we employed a stratified random split, allocating 70% of the instances for training and 30% for testing. The Soybean dataset was used with its original training and test partitions. Table 1 summarizes the structural characteristics of these data sets, including the number of conditional attributes and decision classes.

Table 1. Characteristics of the benchmark datasets

Dataset	Training set	Test set	Attributes	Classes
Vehicle Silhouettes	592	254	18	4
Soybean	307	376	35	19
Lymphography	104	44	18	4

To simulate dispersion, the training attributes were partitioned into five different configurations consisting of 3, 5, 7, 9, and 11 local tables. Each table contained a unique subset of conditional attributes but maintained the full set of training objects. We ensured that while some attributes were shared between tables to simulate overlapping views, the majority were distributed to create fragmented perspectives. Notably, object identifiers were not shared across tables to prevent direct row-matching, forcing the models to rely on the shared decision logic.

The complexity of each local table varied inversely with the degree of dispersion. In configurations with only 3 tables, each agent possessed a significant

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_7

portion of the total attribute space (approximately 6 to 12 attributes). Conversely, in the 11-table configuration, the feature space was highly fragmented, with some agents possessing as few as three attributes. This variation allowed us to test the framework's ability to maintain predictive power even when individual sources provided extremely limited information.

3.1 Evaluation Methodology

To ensure the scientific validity of our findings, model training and evaluation followed a strictly separated procedure designed to prevent any form of information leakage. The local-level models were trained exclusively on the training data using an out-of-fold (OOF) prediction strategy, where each local table was partitioned into K folds to iteratively train decision trees on $K-1$ subsets while generating class probability predictions for the held-out fold. This exhaustive process resulted in a complete set of OOF probability vectors for all training objects, which served as the foundational features for the global training matrix. The global decision tree was subsequently trained using only these out-of-fold representations and their corresponding true class labels, ensuring that no portion of the test set was accessed during any stage of model construction. Final performance evaluation was conducted by applying the ensemble of local models and the trained global decision tree to the entirely independent test set, with each experimental configuration repeated five times to mitigate variability from random data partitioning and report stable average results.

Classification performance was rigorously assessed using a suite of standard evaluation metrics to capture different facets of model effectiveness. Overall classification accuracy (*acc*) was utilized to measure the total proportion of correctly identified test instances, while precision and recall were employed to evaluate prediction reliability and the model's ability to identify all relevant instances of specific classes, respectively. To provide a single metric that balances these two often-competing objectives, the F-measure was calculated as the harmonic mean: $\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Furthermore, given the potential for class imbalance in the benchmark datasets, we prioritized balanced accuracy (*bacc*), which computes the average recall across all decision classes. This ensures that the model's performance on minority classes is not overshadowed by the majority class, providing a more equitable assessment of the hierarchical framework's predictive power.

4 Results and Discussion

This section presents the experimental results and comparative analysis of the proposed hierarchical framework. We first evaluate the sensitivity of the model to local tree depth before conducting a global comparison against established baseline methods.

The performance of the proposed architecture was evaluated across various local tree depths ($d \in \{4, 6, 8, 10\}$) using balanced accuracy (*bacc*) as the primary

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_7

metric. Table 2 provides a detailed breakdown of these results across the Vehicle Silhouettes, Soybean, and Lymphography datasets for different levels of data dispersion.

The Friedman test indicated no globally significant difference in results across the four investigated depths ($\chi^2(3, 15) = 2.84, p = 0.41$), suggesting that the ensemble method maintains structural resilience regardless of minor depth variations. However, comparative analysis of the performance distributions in shows that $d = 6$ and $d = 8$ exhibit greater stability and higher median values than the shallower $d = 4$ configuration. Post-hoc Wilcoxon each-pair signed-rank tests confirmed these differences as statistically significant ($p = 0.031$ and $p = 0.027$, respectively). Conversely, the transition to $d = 10$ yielded a p-value of 0.18, indicating diminishing returns. Thus, a maximum depth of 6 or 8 is optimal for balancing predictive power while avoiding overfitting.

Table 2. Comparative results of precision, recall, F-measure, balanced accuracy (*bacc*), and accuracy (*acc*) across Vehicle, Soybean, and Lymphography datasets.

Agent		**Vehicle Silhouettes**				**Soybean**				**Lymphography**			
No.	Metric	4	6	8	10	4	6	8	10	4	6	8	10
3	Prec.	0.693	0.687	0.652	0.690	0.748	0.715	0.704	0.707	0.689	0.653	0.620	0.625
	Recall	0.665	0.676	0.644	0.672	0.718	0.693	0.701	0.672	0.707	0.665	0.651	0.637
	F-m.	0.670	0.674	0.641	0.672	0.683	0.662	0.671	0.642	0.679	0.629	0.625	0.601
	bacc	0.655	0.664	0.632	0.659	0.673	0.620	0.626	0.614	0.511	0.478	0.469	0.459
	acc	0.665	0.676	0.644	0.672	0.718	0.693	0.701	0.672	0.707	0.665	0.651	0.637
5	Prec.	0.665	0.696	0.681	0.668	0.822	0.806	0.765	0.769	0.694	0.704	0.711	0.680
	Recall	0.656	0.695	0.665	0.669	0.829	0.799	0.772	0.773	0.726	0.740	0.730	0.716
	F-m.	0.654	0.694	0.666	0.664	0.815	0.789	0.754	0.755	0.705	0.717	0.713	0.693
	bacc	0.639	0.669	0.649	0.649	0.776	0.760	0.730	0.710	0.520	0.530	0.523	0.514
	acc	0.656	0.695	0.665	0.669	0.829	0.799	0.772	0.773	0.726	0.740	0.730	0.716
7	Prec.	0.673	0.666	0.680	0.701	0.796	0.779	0.791	0.755	0.685	0.632	0.594	0.551
	Recall	0.677	0.659	0.669	0.691	0.797	0.796	0.809	0.776	0.702	0.651	0.628	0.581
	F-m.	0.671	0.659	0.672	0.694	0.781	0.779	0.792	0.756	0.684	0.630	0.609	0.561
	bacc	0.655	0.639	0.649	0.666	0.773	0.785	0.792	0.766	0.505	0.469	0.449	0.416
	acc	0.677	0.659	0.669	0.691	0.797	0.796	0.809	0.776	0.702	0.651	0.628	0.581
9	Prec.	0.681	0.690	0.676	0.682	0.766	0.803	0.753	0.743	0.691	0.670	0.688	0.655
	Recall	0.674	0.680	0.662	0.669	0.744	0.773	0.720	0.722	0.712	0.712	0.698	0.670
	F-m.	0.670	0.682	0.665	0.674	0.735	0.771	0.712	0.713	0.685	0.688	0.688	0.658
	bacc	0.659	0.662	0.646	0.651	0.678	0.714	0.674	0.664	0.512	0.508	0.500	0.481
	acc	0.674	0.680	0.662	0.669	0.744	0.773	0.720	0.722	0.712	0.712	0.698	0.670
11	Prec.	0.657	0.645	0.649	0.651	0.718	0.751	0.731	0.734	0.707	0.714	0.716	0.715
	Recall	0.654	0.635	0.641	0.639	0.694	0.710	0.720	0.710	0.726	0.730	0.740	0.749
	F-m.	0.649	0.633	0.643	0.640	0.685	0.705	0.704	0.701	0.699	0.706	0.717	0.725
	bacc	0.638	0.622	0.622	0.622	0.641	0.650	0.653	0.648	0.523	0.527	0.532	0.539
	acc	0.654	0.635	0.641	0.639	0.694	0.710	0.720	0.710	0.726	0.730	0.740	0.749

To evaluate effectiveness, the Cross-Validation-Based (CVB) model was compared against centralized models (AdaBoost, Decision Tree, Naive Bayes) using

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_7

majority voting, alongside a dual-level bagging (DLB) method [8]. Table 3 illustrates that the CVB model consistently achieves superior performance, frequently outperforming baselines by substantial margins.

This architectural advantage is most evident in the Soybean dataset, where standard models suffer performance collapse as fragmentation increases. In contrast, CVB effectively integrates sparse local features. Statistical validation using the Friedman test ($\chi^2(74, 4) = 138.22, p < 0.000001$) and post-hoc Wilcoxon tests confirm that CVB provides a significantly more robust and stable classification framework than traditional voting mechanisms.

Table 3. Comprehensive performance comparison across datasets. Best results in blue; second best in red.

No. of Tables	Metric	Vehicle Silhouettes					Soybean Dataset					Lymphography Dataset				
		AB	DT	NB	DLB	CVB	AB	DT	NB	DLB	CVB	AB	DT	NB	DLB	CVB
3	*bacc*	0.628	0.669	0.513	0.696	0.664	0.158	0.878	0.864	0.668	0.673	0.596	0.532	0.467	0.855	0.511
	acc	0.646	0.677	0.520	0.717	0.676	0.202	0.823	0.699	0.793	0.718	0.386	0.773	0.682	0.791	0.707
5	*bacc*	0.619	0.678	0.500	0.695	0.669	0.114	0.050	0.071	0.739	0.776	0.682	0.563	0.457	0.842	0.530
	acc	0.630	0.693	0.504	0.721	0.695	0.177	0.086	0.083	0.830	0.829	0.545	0.818	0.659	0.774	0.740
7	*bacc*	0.518	0.699	0.484	0.655	0.666	0.064	0.056	0.056	0.664	0.792	0.268	0.520	0.424	0.679	0.505
	acc	0.520	0.717	0.484	0.683	0.691	0.135	0.105	0.041	0.803	0.809	0.386	0.750	0.614	0.748	0.702
9	*bacc*	0.441	0.665	0.459	0.665	0.662	0.082	0.046	0.099	0.588	0.714	0.374	0.478	0.472	0.570	0.512
	acc	0.441	0.681	0.457	0.682	0.680	0.110	0.124	0.135	0.745	0.773	0.545	0.682	0.682	0.591	0.712
11	*bacc*	0.551	0.656	0.450	0.636	0.638	0.077	0.104	0.066	0.669	0.653	0.368	0.478	0.683	0.848	0.539
	acc	0.547	0.673	0.441	0.649	0.654	0.133	0.133	0.086	0.779	0.720	0.523	0.682	0.523	0.783	0.749

5 Summary

This paper investigated the problem of classification in dispersed data environments, where data are distributed across multiple local sources with heterogeneous feature spaces and a shared decision attribute. To address the limitations of existing hierarchical approaches that rely on explicit validation splits of the test set, a two level hierarchical decision tree framework was proposed. The method leverages stratified cross validation at the local level to generate out-of-fold probability predictions, which serve as reliable and leakage free inputs for training a global decision tree. By constructing the global model solely from training data derived probability vectors, the framework preserves the integrity of the test set for unbiased evaluation while maximizing the use of available data.

Experimental results on several benchmark multiclass datasets with varying degrees of dispersion demonstrate that the proposed approach is effective in integrating information from fragmented views and performs competitively across different performance measures. The findings suggest that hierarchical decision trees combined with cross validation based stacking provide a robust and interpretable solution for learning from dispersed data without requiring data centralization or additional validation partitions.

Future work will focus on exploring alternative global learners, analyzing the impact of different cross validation strategies, and extending the framework to more complex real world scenarios with stronger class imbalance and higher dimensional feature spaces.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_7

References

1. Firouzi, R., Rahmani, R., Kanter, T. (2021). Federated learning for distributed reasoning on edge computing. Procedia Computer Science, 184, 419–427.
2. Ksieniewicz, P., Zyblewski, P., Burduk, R. (2021). Fusion of linear base classifiers in geometric space. Knowledge-Based Systems, 227, 107231.
3. Lewy, D., Mańdziuk, J., Ganzha, M., Paprzycki, M. (2022). StatMix: Data augmentation method that relies on image statistics in federated learning. In International Conference on Neural Information Processing (pp. 574–585). Springer Nature Singapore.
4. Marfo, K. F., Przybyła-Kasperek, M. (2022). Radial basis function network for aggregating predictions of k-nearest neighbors local models generated based on independent data sets. Procedia Computer Science, 207, 3234–3243.
5. Michalski, R. S., Chilausky, R. L. (1999). Knowledge acquisition by encoding expert rules versus computer induction from examples: A case study involving soybean pathology. International Journal of Human-Computer Studies, 51(2), 239–263.
6. Pedrycz, W. (2023). Advancing federated learning with granular computing. Fuzzy Information and Engineering, 15(1), 1–13.
7. Przybyła-Kasperek, M., Addo, B. A. (2025). Novel hierarchical decision tree frameworks introducing tree method bagging stump integration and height optimization. In International Conference on Computational Science (pp. 3–11). Springer Nature Switzerland.
8. Przybyła-Kasperek, M., Addo, B. A., Kusztal, K. (2024). Dual-level decision tree-based model for dispersed data classification. In B. Marcinkowski et al. (Eds.), Harnessing Opportunities: Reshaping ISD in the Post-COVID-19 and Generative AI Era (ISD2024 Proceedings). University of Gdańsk. https://doi.org/10.62036/ISD.2024.44
9. Rahim, N., El-Sappagh, S., Rizk, H., El-Serafy, O. A., Abuhmed, T. (2024). Information fusion-based Bayesian optimized heterogeneous deep ensemble model based on longitudinal neuroimaging data. Applied Soft Computing, 162, 111749.
10. Seydi, S. T., Saeidi, V., Kalantar, B., Ueda, N., van Genderen, J. L., Maskouni, F. H., Aria, F. A. (2022). Fusion of the multisource datasets for flood extent mapping based on ensemble convolutional neural network model. Journal of Sensors, 2022(1), 2887502.
11. Siebert, J. P. (1987). Vehicle recognition using rule based methods.
12. Trajdos, P., Burduk, R. (2024). Ensemble of classifiers based on score function defined by clusters and decision boundary of linear base learners. Knowledge-Based Systems, 303, 112411.
13. Węgier, W., Koziarski, M., Woźniak, M. (2022). Multicriteria classifier ensemble learning for imbalanced data. IEEE Access, 10, 16807–16818.
14. Yu, L., Li, M. (2023). A case-based reasoning driven ensemble learning paradigm for financial distress prediction with missing data. Applied Soft Computing, 137, 110163.
15. Yurochkin, M., Agarwal, M., Ghosh, S., Redewald, K., Hoang, N., Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks. In International Conference on Machine Learning (pp. 7252–7261). PMLR.
16. Zwitter, M., Soklic, M. (1988). Lymphography domain. University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_7