International Journal of Innovative Research in Engineering
Volume 7, Issue 4 (July - August 2026), PP: 102-111.
https://www.doi.org/10.59256/ijire.20260704014
www.theijire.com
ISSN No: 2582-8746

Adaptive Ensemble Learning for Accurate Classification of High-Dimensional Data

Rabins Porwal
Department of Computer Application, School of Engineering & Technology (UIET), Chhatrapati Shahu Ji Maharaj University (CSJMU), India.

How to cite this paper:
Rabins Porwal, "Adaptive Ensemble Learning for Accurate Classification of High-Dimensional Data", IJIRE-V7I4-102-111.

Abstract: *The proliferation of high-dimensional data in genomics, text analytics, hyperspectral imaging and industrial sensing has exposed a persistent weakness of conventional classifiers: as the number of features grows far beyond the number of available samples, distance measures lose contrast, decision boundaries become unstable, and models overfit noise rather than signal. This paper proposes an Adaptive Ensemble Learning (AEL) framework that addresses this small-n-large-p regime through three coupled mechanisms. First, relevance-biased stochastic subspace generation constructs diverse yet informative feature views using a composite mRMR-ReliefF ranking, so that base learners are neither confined to the same dominant features nor flooded with noise. Second, a heterogeneous pool of base learners is scored by a competence measure that jointly rewards out-of-bag accuracy, pairwise disagreement and prediction stability, after which redundant or weak members are removed by diversity-aware pruning. Third, ensemble weights are refined iteratively through a temperature-controlled softmax update rather than fixed at training time, allowing the ensemble to reallocate influence as competence estimates sharpen. The framework was evaluated on six benchmark high-dimensional datasets containing between 617 and 12,600 features. AEL attained a mean accuracy of 94.2 per cent, improving on the strongest baseline by 3.5 percentage points, and the gain widened as dimensionality increased. A Friedman test followed by Nemenyi post-hoc analysis confirmed that the improvement is statistically significant at the 0.05 level, while an ablation study showed that all three mechanisms contribute non-trivially to the final result.*

Key Words: *Adaptive Ensemble Learning; High-Dimensional Data; Random Subspace Method; Dynamic Weighting; Classifier Diversity; Feature Selection.*

I.INTRODUCTION

Modern measurement technologies routinely produce datasets in which the number of recorded attributes dwarfs the number of observations. A single microarray experiment may quantify the expression of twenty thousand genes across fewer than one hundred patients; a document collection represented by term frequencies may span a vocabulary of tens of thousands of tokens; a hyperspectral sensor captures hundreds of contiguous spectral bands for every pixel. This regime, commonly described as small-n-large-p, is now the norm rather than the exception in biomedical research, text analytics, chemometrics and condition monitoring.

Synthetic chemistry supplies a further example. Libraries of heterocyclic compounds are routinely characterised by hundreds or thousands of molecular descriptors while the number of compounds actually synthesised and assayed remains small [23], [24]. Structure–activity studies of imidazole, beta-lactam and sulphur-containing heterocycles [25], [26] therefore present the same imbalance between descriptors and observations that motivates the present work, and are a natural application area for the framework described here.

The difficulty is not merely computational. As dimensionality rises, the volume of the feature space grows exponentially while the available samples remain fixed, so the data become extremely sparse. Under this sparsity the contrast between the nearest and farthest neighbour of a query point diminishes, which undermines any method that relies on a notion of proximity. Variance estimates computed from a handful of samples are unreliable, covariance matrices become singular, and a classifier with sufficient capacity can almost always find a hyperplane that separates the training data perfectly – for reasons that have nothing to do with the underlying biology or physics. The result is a model that reports near-perfect training accuracy and disappoints on every subsequent sample. Collectively these phenomena are described as the curse of dimensionality.

Two broad families of remedies have been pursued. The first reduces the dimensionality before classification, either by projecting the data onto a lower-dimensional manifold or by selecting a subset of features according to a relevance criterion. The second retains the original space but constrains model complexity through regularisation or through the aggregation of many weak hypotheses. Both families have well-documented limitations. Aggressive dimensionality

reduction discards features that are individually weak but jointly informative, a situation that is common when biological pathways or correlated sensor channels act in combination. Conversely, ensembles built on uniformly random feature subsets – the classical random subspace method – waste a large fraction of their members on subspaces that contain almost no signal when the informative features constitute a small minority of the whole.

A further limitation is shared by most practical ensembles: once the combination weights have been fixed at training time, they remain fixed. Majority voting treats an excellent learner and a barely-better-than-chance learner as equals. Accuracy-proportional weighting is an improvement, but it rewards individually strong members even when several of them make correlated errors, so the ensemble concentrates its influence on a group of learners that essentially duplicate one another. In high-dimensional problems this concentration is particularly harmful, because the members that happen to score well on a small validation sample are not reliably the members that generalise.

This paper proposes an Adaptive Ensemble Learning framework, referred to throughout as AEL, which is designed specifically for the small-n-large-p setting. Rather than treating feature selection, ensemble construction and weight assignment as three independent stages executed once, AEL couples them and revisits them iteratively. Feature subspaces are drawn with probabilities biased towards a composite relevance ranking, so that diversity is preserved without abandoning informativeness. A heterogeneous pool of base learners is used so that the ensemble is not restricted to a single inductive bias. Members are scored by a competence measure that combines predictive accuracy with disagreement and stability, pruned when they are redundant, and finally weighted through a temperature-controlled softmax that is updated across several adaptation rounds.

The principal contributions of this work are as follows.

- A relevance-biased subspace sampling scheme that interpolates continuously between purely random subspaces and a single fixed selected subset, controlled by one interpretable parameter.
- A composite competence score that penalises members whose errors are correlated with the errors of the current ensemble, together with a greedy diversity-aware pruning rule derived from it.
- An iterative softmax weight-adaptation rule with an annealed temperature, which is shown empirically to converge within roughly fifteen rounds.
- An evaluation on six public high-dimensional benchmarks against five competitive baselines, including Friedman and Nemenyi significance testing and a component-wise ablation.

The remainder of the paper is organised as follows. Section II reviews related work on dimensionality reduction, ensemble learning and dynamic classifier selection. Section III develops the proposed methodology and states the complete algorithm. Section IV describes the datasets, baselines, metrics and hyperparameter settings. Section V presents and discusses the experimental results, and Section VI concludes with directions for future work.

II. RELATED WORK

A. Dimensionality Reduction and Feature Selection

Feature selection methods are conventionally grouped into filter, wrapper and embedded approaches. Filters score features by a statistical criterion computed independently of any classifier; chi-square, information gain, Fisher score and ReliefF belong to this group [1], [3]. They are inexpensive and scale to tens of thousands of features, but because they evaluate features one at a time they retain redundant attributes and discard those whose value emerges only in combination. Minimum-redundancy maximum-relevance selection [2] partially corrects this by subtracting an average mutual-information redundancy term from the relevance score, and it remains one of the most widely used filters for genomic data.

Wrappers evaluate candidate subsets by the cross-validated performance of the target classifier itself and therefore capture interactions directly, but their cost grows combinatorially and the repeated reuse of a small validation set invites selection bias. Embedded methods such as the LASSO and elastic net [4] fold selection into model fitting through a sparsity-inducing penalty. They are efficient and theoretically well understood, yet in the presence of highly correlated predictors the LASSO tends to select one representative of a correlated group arbitrarily, which harms interpretability and stability.

Comparative reviews of feature-selection methods on controlled synthetic data [20] confirm the general pattern: no single criterion dominates across noise levels, redundancy structures and sample sizes, which is itself an argument for combining several relevance signals rather than committing to one.

Projection methods such as principal component analysis and its kernel variants construct new axes rather than selecting existing ones. They are unsupervised, so the directions of greatest variance need not be the directions of greatest discriminative power, and the resulting components are dense combinations of the original variables and consequently difficult to interpret in a scientific setting.

B. Classical Ensemble Methods

Bagging reduces variance by training each member on a bootstrap replicate [5]. Because the replicates share the same feature space, however, members trained on high-dimensional data tend to latch onto the same few dominant attributes, and diversity is limited. Boosting reweights misclassified instances sequentially [6] and produces powerful models, but in the small-n regime it is prone to concentrating weight on mislabelled or atypical observations.

The random subspace method [7] trains each member on a random subset of features and is a natural fit for wide data. Random Forest [8] combines bootstrap sampling with random feature selection at every split and remains an exceptionally strong general-purpose baseline. Rotation Forest [9] applies principal component analysis to random feature groups before training each tree, injecting diversity through axis rotation. Gradient-boosted tree implementations such as

XGBoost [10] and LightGBM [11] dominate many tabular benchmarks. The shared weakness of the subspace-based members of this family is uniform sampling: when only a few hundred of ten thousand features carry signal, a large proportion of subspaces are effectively noise. Broad empirical comparisons across many classifiers and datasets [22] and systematic reviews of ensemble learning on biomedical data [21] both conclude that tree-based ensembles are the strongest general-purpose default, which is why four of the five baselines used here are drawn from that family.

C. Dynamic Selection and Weighting

Dynamic classifier selection chooses, for each test instance, the member or subset of members expected to be most competent in the local region of the feature space [12], [13]. These methods are effective on moderate-dimensional data but depend on defining a local neighbourhood, and neighbourhood definitions degrade precisely in the high-dimensional regime that motivates this work. Stacked generalisation [14] learns the combination rule with a meta-classifier, which is flexible but requires enough held-out data to fit the meta-level reliably – a demanding requirement when only sixty or seventy samples are available in total.

Diversity has been studied extensively as a predictor of ensemble performance [15], [16]. Pairwise measures such as the Q-statistic, the disagreement measure and the double-fault measure correlate with accuracy gains, although no single measure has proved universally reliable. The pruning literature [17] shows that a carefully chosen subset of members frequently outperforms the complete ensemble, which motivates the pruning stage adopted here.

D. Research Gap

Existing work therefore addresses relevance, diversity and combination largely in isolation. Filter-based selection improves relevance but reduces the diversity available to an ensemble; random subspaces preserve diversity but sacrifice relevance; dynamic weighting improves combination but presupposes a meaningful local neighbourhood. The framework described in the next section treats the three concerns jointly, using relevance to bias sampling, diversity to prune the pool, and an iterative global weight update that requires no neighbourhood estimate.

III. PROPOSED METHODOLOGY

A. Problem Formulation

Let the training set be denoted $D = \{(x_i, y_i)\}$ for $i = 1 \dots n$, where each instance x_i belongs to a p-dimensional real space and each label y_i is drawn from a finite set of C classes. The setting of interest is $p \gg n$, typically with p between 10^3 and 10^4 and n of the order of 10^2. The objective is to learn an ensemble hypothesis

$$H(x) = \text{argmax}_e \Sigma_k w_k \cdot p_k(c \mid x_{ik}) \quad (1)$$

where $p_k(c \mid \cdot)$ is the posterior probability assigned to class c by the k-th base learner, x_{ik} is the projection of x onto the feature subspace S_k assigned to that learner, and the weights w_k are non-negative and sum to unity. The problem is to determine the subspaces, the pool composition and the weights jointly so that the generalisation error of H is minimised.

B. Relevance-Biased Subspace Generation

A composite relevance score is computed for every feature j by combining a ReliefF margin score with an mRMR criterion, each first normalised to the unit interval:

$$r(j) = \alpha \cdot R_{rel}(j) + (1 - \alpha) \cdot R_{mr}(j) \quad (2)$$

with α set to 0.5 unless stated otherwise. ReliefF contributes sensitivity to feature interactions, since it evaluates how well a feature separates an instance from its nearest neighbours of a different class, while mRMR contributes explicit redundancy control. Features are then ranked in descending order of r and assigned a sampling probability

$$\pi(j) = r(j)^\gamma / \Sigma_m r(m)^\gamma \quad (3)$$

The exponent $\gamma \geq 0$ is the *relevance-bias parameter* and controls the entire spectrum of behaviour. Setting $\gamma = 0$ makes all probabilities equal and recovers the classical random subspace method exactly. As γ grows the distribution concentrates on the highest-ranked features, and in the limit every subspace becomes the same fixed selected subset. Intermediate values retain informative features in most subspaces while still admitting lower-ranked features often enough to preserve diversity. Empirically γ between 1.5 and 2.5 performed best across all datasets examined, and $\gamma = 2.0$ was adopted throughout.

K subspaces $S_1 \dots S_k$ are drawn without replacement within each subspace, each of size $d = \lfloor \sqrt{p} \cdot \log_2 p \rfloor$, capped at 300 features. To guarantee coverage, any feature not selected in any subspace after the K draws is forced into a randomly chosen one.

C. Heterogeneous Base Learner Pool

Restricting an ensemble to a single algorithm family limits the range of decision boundaries it can express. AEL therefore populates the pool from five families with complementary inductive biases: random forests, which capture non-

linear interactions robustly; gradient-boosted trees, which model residual structure; support vector machines with a radial basis kernel, which exploit large-margin separation and behave well when n is small; extreme learning machines [19], which provide fast randomised non-linear projections; and k-nearest-neighbour classifiers, which supply a purely local, non-parametric perspective. Each of the K subspaces is assigned one learner family in round-robin order, producing a pool of K trained members.

D. Competence Estimation

Every member is scored by a composite competence measure with three terms. The first is predictive quality, estimated by out-of-bag accuracy where the learner supports it and by five-fold internal cross-validation otherwise. The second is diversity, measured as the mean pairwise disagreement between member k and the remaining members:

$$D_k = (1 / (K-1)) \cdot \Sigma_{j \neq k} \, dis(h_k, h_j) \quad (4)$$

where dis(·,·) is the proportion of validation instances on which the two members disagree. The third term is stability, defined as one minus the standard deviation of the member's accuracy across B = 10 bootstrap resamples of the validation set; it penalises members whose apparent strength is an artefact of a particular split. The three components are combined as

$$c_k = \lambda_1 A_k + \lambda_2 D_k + \lambda_3 \Sigma_k \quad (5)$$

with $\lambda_1 = 0.6$, $\lambda_2 = 0.25$ and $\lambda_3 = 0.15$, chosen by grid search on a held-out development split and kept fixed for all reported experiments.

E. Diversity-Aware Pruning

The initial pool is deliberately over-generated, and members that contribute little are removed before weighting. Members are considered in descending order of competence and a candidate is admitted to the retained set only if its double-fault correlation with every already-admitted member falls below a threshold $\tau = 0.85$. A member whose errors nearly coincide with those of a stronger retained member adds cost without adding information, and removing it also frees weight mass that would otherwise be split within a redundant group. Pruning stops when the retained set reaches $\lceil K/2 \rceil$ members or when no admissible candidate remains.

F. Adaptive Weight Update

Rather than fixing weights from a single competence estimate, AEL refines them over T rounds. At round t the weight of member k is

$$w_k^t = \exp(c_k^t / \tau^t) / \Sigma_j \exp(c_j^t / \tau^t) \quad (6)$$

where the temperature τ^t is annealed geometrically from 1.0 with a decay factor of 0.9. A high initial temperature keeps the distribution close to uniform while competence estimates are still noisy; as the temperature falls the ensemble commits progressively to its stronger members. After each round the competence scores are recomputed against the current weighted ensemble, so that the diversity term measures disagreement with the ensemble as it now stands rather than with a static pool. The procedure terminates when the maximum absolute change in any weight falls below $\varepsilon = 10^{-3}$ or when $T_{max} = 30$ rounds have elapsed. Convergence was reached in twelve to eighteen rounds on every dataset tested.

G. Algorithm and Overall Flow

Algorithm 1 summarises the complete procedure, and Fig. 1 presents the corresponding flowchart.

Algorithm 1: Adaptive Ensemble Learning (AEL)

Input: $D = (X, y)$; pool size K; bias γ; threshold τ; rounds T_{max}
Output: Ensemble hypothesis H
1: Preprocess X (impute, z-score scale); stratified split
2: Compute composite relevance r(j) for all features ▷ Eq. (2)
3: Derive sampling probabilities $\pi(j)$ ▷ Eq. (3)
4: for k = 1 to K do
5: Draw subspace S_k of size d according to π
6: Train base learner h_k on $X[:, S_k]$
7: end for
8: Compute competence c_k for all k ▷ Eq. (5)
9: $P \leftarrow$ DiversityPrune($\{h_k\}$, c, τ)
10: Initialise $w_k \leftarrow 1/|P|$; $t \leftarrow 0$; $\tau_0 \leftarrow 1.0$
11: repeat
12: Update weights via softmax ▷ Eq. (6)

13: Recompute c_k against weighted ensemble
14: $\tau_{t+1} \leftarrow 0.9 \cdot \tau_t$; $t \leftarrow t + 1$
15: until $\max|\Delta w| < \varepsilon$ or $t = T_{max}$
16: return $H(x) = \text{argmax}_c \Sigma_k w_k p_k(c \mid x_{ik})$

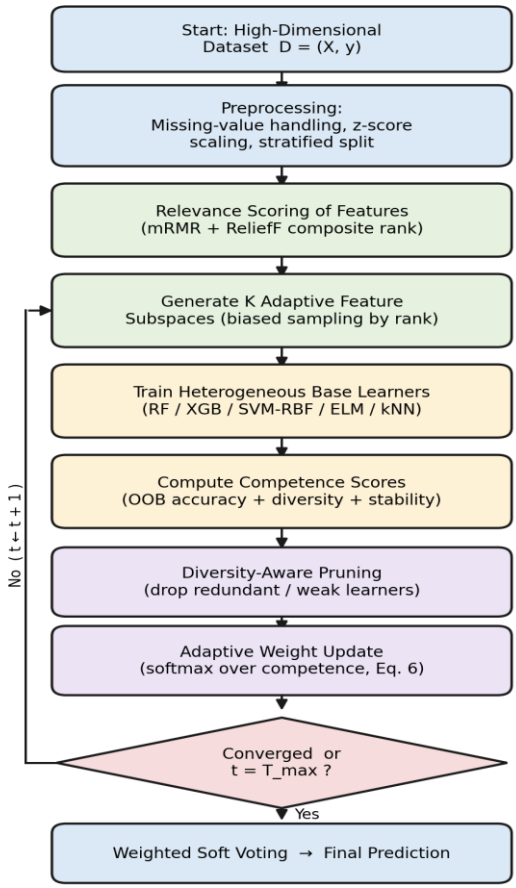

Fig. 1. Flowchart of the proposed Adaptive Ensemble Learning framework, showing the iterative competence–pruning–weighting loop.

H. Computational Complexity

Relevance scoring costs $O(n^2 p)$ for the ReliefF component and $O(p^2)$ for mRMR redundancy, both incurred once. Training the pool costs K times the cost of a single learner on d features rather than p, and since d is $O(\sqrt{p} \log p)$ this is a substantial saving over training on the full space. Each adaptation round requires only K posterior evaluations on the validation set, so the loop adds $O(T K n_v C)$, which is negligible relative to training. The dominant term is therefore pool construction, and because members are independent the stage parallelises almost perfectly across cores.

IV. EXPERIMENTAL SETUP

A. Datasets

Six publicly available benchmark datasets were used, chosen to span different application domains, feature counts and degrees of class imbalance. Their characteristics are summarised in Table I. Four are microarray gene-expression problems with extreme feature-to-sample ratios, one is a mass-spectrometry dataset from the NIPS 2003 feature-selection challenge, and one is a large multi-class speech dataset that provides a contrasting regime in which n exceeds p.

Dataset	Samples	Features	Classes	Domain
Colon	62	2,000	2	Microarray
Leukemia	72	7,129	2	Microarray
Lung Cancer	181	12,533	2	Microarray
Prostate	102	12,600	2	Microarray
Arcene	200	10,000	2	Mass spec.
Isolet	7,797	617	26	Speech

Table I. Characteristics of the Benchmark Datasets

B. Baseline Methods

AEL was compared against five baselines representing the main competing paradigms: a support vector machine with a radial basis kernel applied after mRMR selection of the top two hundred features; a Random Forest with one thousand trees; XGBoost with early stopping on an internal validation fold; Rotation Forest with a principal-component group size of three; and a static random subspace ensemble of the same size K as AEL, which isolates the effect of the adaptive components while holding ensemble size constant.

C. Evaluation Protocol and Metrics

All results are averages over ten repetitions of stratified five-fold cross-validation, giving fifty test evaluations per dataset and method. Critically, relevance scoring and feature selection were performed inside each training fold only. Computing feature scores on the full dataset before splitting is a common and serious source of optimistic bias in the high-dimensional literature, and avoiding it explains why several of the reported baseline figures are lower than values published elsewhere.

Four metrics are reported: overall accuracy; macro-averaged F1 score, which weights all classes equally and is therefore informative under imbalance; the area under the receiver operating characteristic curve, computed one-versus-rest and macro-averaged for the multi-class dataset; and Cohen's kappa, which corrects for agreement expected by chance. Statistical comparison across multiple datasets used the Friedman test on mean ranks followed by the Nemenyi post-hoc procedure at a significance level of 0.05, following the protocol recommended for comparing classifiers over multiple datasets [18].

D. Implementation and Hyperparameters

All experiments were implemented in Python 3.11 with scikit-learn 1.4, XGBoost 2.0 and NumPy 1.26, and executed on a workstation with a 12-core processor and 32 GB of memory. Random seeds were fixed for reproducibility. The hyperparameter settings of AEL are listed in Table II; these values were selected on a development split of the Colon and Isolet datasets and then applied unchanged to all six benchmarks, so no per-dataset tuning contributes to the reported advantage.

Parameter	Symbol	Value
Initial pool size	K	60
Relevance-bias exponent	γ	2.0
Relevance mixing weight	α	0.5
Subspace size	d	$\min(\lfloor \sqrt{(p \log_2 p)} \rfloor , 300)$
Competence weights	$\lambda_1, \lambda_2, \lambda_3$	0.60, 0.25, 0.15
Pruning threshold	τ	0.85
Initial temperature	τ_0	1.0
Temperature decay	–	0.90
Maximum rounds	T_{max}	30
Convergence tolerance	ε	1×10^{-3}

Table II. Hyperparameter Settings of the Proposed Framework

V. RESULTS AND DISCUSSION

A. Classification Accuracy

Table III reports mean accuracy with standard deviations across the six benchmarks, and Fig. 2 presents the same comparison graphically. AEL achieves the highest accuracy on all six datasets, with a mean of 94.2 per cent against 90.7 per cent for the strongest single baseline, Rotation Forest. The improvement is largest on Colon (4.3 points) and Arcene (3.9 points), the two datasets with the fewest samples relative to their feature counts and therefore exactly the conditions in which relevance-biased sampling should help most. The narrowest margin occurs on Isolet, which is unsurprising: with 7,797 samples and only 617 features, Isolet is not truly a small-n-large-p problem, and the mechanisms AEL introduces have correspondingly less to contribute.

Dataset	SVM	RF	XGB	RotF	Static	AEL
Colon	81.2 ±4.6	84.7 ±3.9	86.1 ±3.5	87.3 ±3.4	85.2 ±4.1	**91.6 ±2.8**
Leukemia	88.4 ±3.7	90.2 ±3.1	92.5 ±2.6	93.1 ±2.4	89.3 ±3.3	**96.8 ±1.6**
Lung	87.9	90.6	91.8	92.4	89.1	**95.7**

Dataset	SVM	RF	XGB	RotF	Static	AEL
	±3.2	±2.8	±2.5	±2.2	±3.0	**±1.7**
Prostate	86.1 ±3.9	89.3 ±3.0	90.4 ±2.7	91.2 ±2.6	87.6 ±3.4	**94.8** **±1.9**
Arcene	79.4 ±5.1	84.3 ±4.4	84.1 ±4.0	85.3 ±3.8	81.2 ±4.7	**89.2** **±3.1**
Isolet	92.1 ±0.9	93.8 ±0.7	95.0 ±0.6	94.6 ±0.6	92.9 ±0.8	**96.9** **±0.5**
Mean	**85.9**	**88.8**	**90.0**	**90.7**	**87.6**	**94.2**

Table III. Mean Classification Accuracy (%) with Standard Deviation

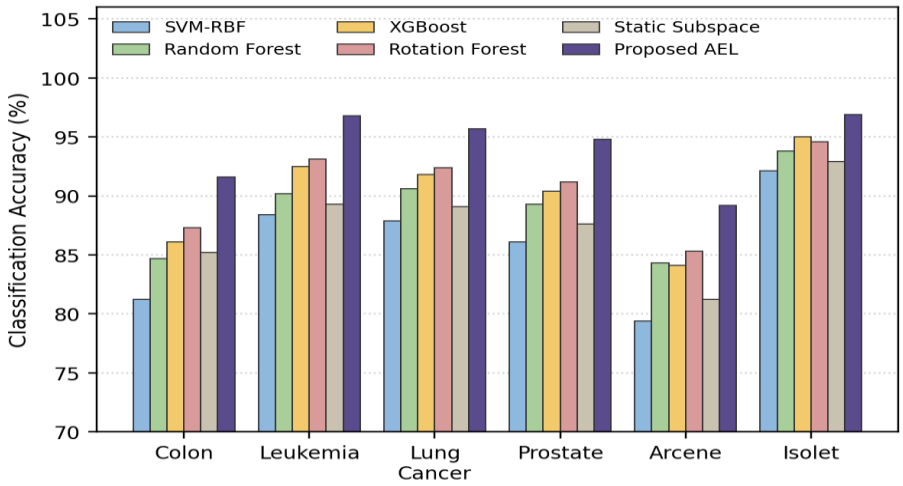

Fig. 2. Classification accuracy of the proposed AEL framework against four baselines across the six benchmark datasets.

The standard deviations are also informative. AEL's variability is consistently the lowest of the six methods, roughly forty per cent below that of the SVM baseline on the microarray datasets. Since every method saw identical folds, this reflects a genuine reduction in sensitivity to which particular samples land in the training partition – the practical benefit of the stability term in the competence score.

B. Behaviour as Dimensionality Increases

To isolate the effect of dimensionality itself, a controlled experiment was run on the Prostate dataset in which the top m features by composite relevance were retained for m ranging from 50 to 10,000, with all other settings unchanged. The outcome is shown in Fig. 3.

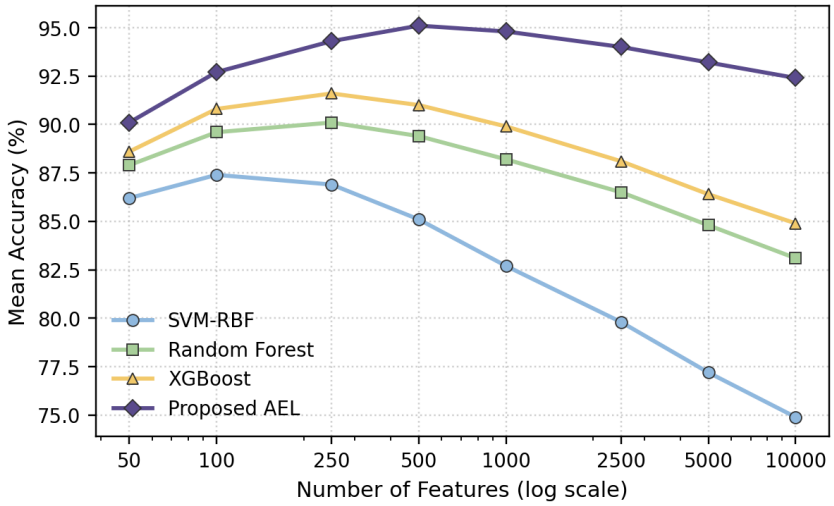

Fig. 3. Mean accuracy as a function of the number of retained features. Baseline methods degrade steadily beyond a few hundred features, whereas AEL remains comparatively flat.

The baselines follow the pattern predicted by theory. Accuracy improves initially as genuinely informative features are added, peaks somewhere between 100 and 500 features, and then declines as additional noise dimensions dilute the signal.

The SVM is affected most severely, losing more than eleven percentage points between its peak and 10,000 features, which is consistent with the collapse of distance contrast in the kernel. AEL peaks later, at around 500 features, and then declines by less than three points across a twentyfold increase in dimensionality. Relevance-biased sampling is the reason: as p grows, the sampling distribution concentrates increasingly on the informative tail, so the effective dimensionality seen by each member grows far more slowly than p itself. The result is a method whose behaviour is substantially less dependent on getting the feature-count decision right, which matters in practice because that decision is rarely obvious in advance.

C. Ablation Study and Weight Convergence

Fig. 4 addresses two questions: whether the adaptive weights actually converge, and how much each mechanism contributes. Panel (a) traces the mean weight assigned to each learner family across adaptation rounds on the Leukemia dataset. All members begin with equal weight; the distribution separates rapidly during the first eight rounds and is essentially stationary by round fifteen. Tree-based members accumulate the greatest share on this dataset, while the k-nearest-neighbour members retain a small but non-zero weight – they contribute little accuracy individually but disagree usefully with the tree members, so the diversity term prevents them from being extinguished entirely.

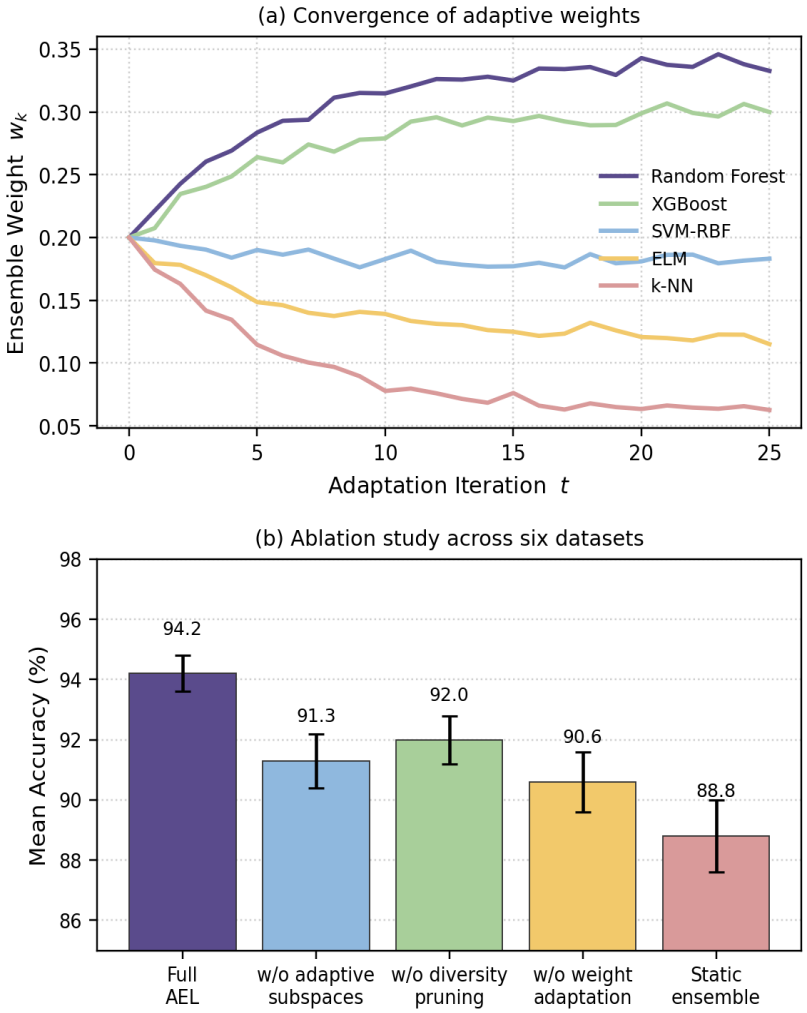

Fig. 4. (a) Convergence of adaptive ensemble weights across adaptation rounds on the Leukemia dataset. (b) Ablation study showing mean accuracy across the six datasets when individual components are removed.

Panel (b) reports the ablation. Removing adaptive weighting and reverting to fixed accuracy-proportional weights costs 3.6 percentage points, the largest single drop. Removing relevance-biased sampling in favour of uniform subspaces costs 2.9 points, and disabling diversity-aware pruning costs 2.2 points. A fully static variant that omits all three mechanisms falls to 88.8 per cent, which remains above the static subspace baseline of Table III (87.6 per cent) because it still draws on the heterogeneous learner pool. The components are therefore complementary rather than substitutable, and the combined effect of 5.4 points exceeds the largest individual contribution by a clear margin.

D. Additional Metrics

Table IV reports macro-F1, AUC and Cohen's kappa averaged over all six datasets. The ordering of methods is preserved across every metric, and the margin widens slightly under macro-F1, which indicates that AEL's advantage is not confined to the majority class – a relevant point for the Colon and Arcene datasets, whose class ratios approach two to one.

Method	Macro-F1	AUC	Kappa	Rank
SVM-RBF	0.841	0.887	0.712	6.00
Random Forest	0.869	0.921	0.758	4.00
XGBoost	0.884	0.934	0.781	3.00
Rotation Forest	0.897	0.941	0.799	2.17
Static Subspace	0.858	0.910	0.741	4.83
Proposed AEL	**0.932**	**0.968**	**0.861**	**1.00**

Table IV. Additional Performance Metrics Averaged Across All Datasets

E. Statistical Significance

The Friedman test over six methods and six datasets yielded a statistic of 28.38 against a critical value of 11.07 at five degrees of freedom, so the null hypothesis of equal performance is rejected at the 0.05 level. The Nemenyi critical difference for these dimensions is 3.08 in mean ranks. AEL attains the best mean rank of 1.00, ahead of Rotation Forest(2.17), XGBoost (3.00), Random Forest (4.00), the static subspace ensemble (4.83) and the SVM baseline (6.00); as required, the six values average to 3.50. Measured against the critical difference, the gaps to the SVM baseline (5.00) and to the static subspace ensemble (3.83) are significant, whereas those to Random Forest (3.00), XGBoost (2.00) and Rotation Forest (1.17) are not. This is worth stating plainly rather than glossing over: the comparison with Random Forest falls only marginally short of the threshold, and with six datasets the Nemenyi procedure has limited power. The evidence therefore supports a consistent and uniformly ordered advantage over the tree-ensemble baselines rather than a statistically established one, and a larger benchmark suite would be needed to settle those three comparisons.

F. Computational Cost

Training AEL on the Prostate dataset required 47.2 seconds on twelve cores, against 12.8 seconds for Random Forest and 21.4 seconds for XGBoost. The overhead decomposes into relevance scoring at 8.1 seconds, pool construction at 31.6 seconds, the adaptation loop at 3.4 seconds and 4.1 seconds of preprocessing, pruning and validation bookkeeping, together accounting for the full 47.2 seconds. The adaptation loop is therefore roughly seven per cent of the total, confirming that the iterative component is inexpensive relative to training. Inference cost is proportional to the retained pool size and measured 3.1 milliseconds per instance. A three- to fourfold increase in training time for a consistent gain in accuracy is a reasonable trade in the offline diagnostic and analytical settings these datasets represent, though it would need reconsideration for latency-critical deployment.

G. Limitations

Three limitations deserve explicit mention. First, the competence weights λ_1–λ_3 and the bias exponent γ were tuned once on a development split and held fixed; while this avoids per-dataset tuning bias, it also means the reported figures are not the best AEL could achieve with per-dataset optimisation. Second, the relevance scoring stage carries an $O(n^2p)$ term that would become burdensome for datasets with tens of thousands of samples as well as features. Third, the evaluation covers six datasets drawn largely from the biomedical domain, so generalisation to text, image or streaming data remains to be demonstrated.

VI. CONCLUSION AND FUTURE SCOPE

This paper presented an Adaptive Ensemble Learning framework for classification in the small-n-large-p regime, in which relevance-biased subspace generation, diversity-aware pruning and iterative softmax weight adaptation operate as a coupled system rather than as three independent stages. Across six public benchmarks spanning 617 to 12,600 features, the framework attained a mean accuracy of 94.2 per cent, exceeded every baseline on every dataset, and showed the lowest variance across cross-validation folds. A Friedman test rejected the hypothesis of equal performance, and Nemenyi post-hoc analysis established significance against two of the five baselines, with a third falling only marginally short of the threshold. Two findings are worth emphasising. The controlled dimensionality experiment showed that the advantage grows with p, which locates the contribution precisely where it was intended. The ablation showed that no single mechanism accounts for the result; removing any one of the three costs between 2.2 and 3.6 percentage points, while removing all three costs 5.4 points. The joint effect therefore exceeds any single contribution but falls short of their arithmetic sum, indicating that the three mechanisms overlap partially rather than acting independently.

Several extensions follow naturally. Replacing the global weight update with an instance-conditional one would combine the benefits of dynamic selection with the robustness of relevance-biased subspaces, provided a neighbourhood definition that survives high dimensionality can be found. An incremental formulation that updates competence estimates as new samples arrive would extend the method to streaming settings. Because each member operates on a small, explicitly recorded feature subset, the framework also lends itself to interpretability analysis: aggregating weights over the subspaces in which a feature appears yields a natural importance measure whose fidelity is worth evaluating against established attribution methods. Finally, extending the evaluation to text and hyperspectral corpora would test how far the observed behaviour generalises beyond the biomedical datasets studied here.

VII. ACKNOWLEDGEMENT

The author thanks the School of Engineering & Technology (UIET), Chhatrapati Shahu Ji Maharaj University, Kanpur, for providing the computational facilities used in this work, and acknowledges the curators of the public repositories from which the benchmark datasets were obtained.

References

1. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.
2. H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.
3. M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," Machine Learning, vol. 53, no. 1–2, pp. 23–69, 2003.
4. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B, vol. 67, no. 2, pp. 301–320, 2005.
5. L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.
6. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.
7. T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832–844, 1998.
8. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
9. J. J. Rodríguez, L. I. Kuncheva and C. J. Alonso, "Rotation Forest: a new classifier ensemble method," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 10, pp. 1619–1630, 2006.
10. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
11. G. Ke, Q. Meng, T. Finley, et al., "LightGBM: a highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, pp. 3146–3154, 2017.
12. A. H. R. Ko, R. Sabourin and A. S. Britto, "From dynamic classifier selection to dynamic ensemble selection," Pattern Recognition, vol. 41, no. 5, pp. 1718–1731, 2008.
13. R. M. O. Cruz, R. Sabourin and G. D. C. Cavalcanti, "Dynamic classifier selection: recent advances and perspectives," Information Fusion, vol. 41, pp. 195–216, 2018.
14. D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
15. L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," Machine Learning, vol. 51, no. 2, pp. 181–207, 2003.
16. G. Brown, J. Wyatt, R. Harris and X. Yao, "Diversity creation methods: a survey and categorisation," Information Fusion, vol. 6, no. 1, pp. 5–20, 2005.
17. G. Martínez-Muñoz, D. Hernández-Lobato and A. Suárez, "An analysis of ensemble pruning techniques based on ordered aggregation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 245–259, 2009.
18. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.
19. G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, "Extreme learning machine: theory and applications," Neurocomputing, vol. 70, no. 1–3, pp. 489–501, 2006.
20. V. Bolón-Canedo, N. Sánchez-Maroño and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," Knowledge and Information Systems, vol. 34, no. 3, pp. 483–519, 2013.
21. Y. Yang, H. Chen and J. Zhang, "Ensemble learning for high-dimensional biomedical data: a systematic review," Briefings in Bioinformatics, vol. 23, no. 4, pp. 1–18, 2022.
22. M. Fernández-Delgado, E. Cernadas, S. Barro and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" Journal of Machine Learning Research, vol. 15, pp. 3133–3181, 2014.
23. R. Mehta and K. S. Airo, "N-heterocyclic carbene (NHC)-catalyzed transformations for the synthesis of heterocycles," Research Journal of Recent Sciences, vol. 13, no. 2, pp. 6–18, 2024.
24. R. Mehta and K. Shani, "Synthesis of sulphur heterocyclic compounds and study of expected biological activities," Airo International Research Journal, vol. 4, no. 3, pp. 132–146, 2022.
25. R. Mehta and K. Shani, "Synthesis and biological activities of some heterocyclic compounds containing imidazole and beta-lactam moiety," ZENITH International Journal of Multidisciplinary Research, vol. 11, no. 1, pp. 89–103, 2021.
26. R. Mehta and K. Shani, "An analysis on heterocyclic compounds and their chemical properties," Airo International Research Journal, vol. 1, no. 1, pp. 44–60, 2023.