SSP-Ensemble: A Sufficient Subspace Projection Ensemble for Multiclass Classification

Jiafeng Chen[1] **Cheng Meng**[2] **Peize Wang**[3] **Jingyi Zhang**[1,4] **Jun Zhu**[3,*]

Abstract

High-dimensional multiclass classification is challenging since the discriminative information is often concentrated in low-dimensional structures, while many observed variables are irrelevant or noisy. In particular, different class pairs may depend on different informative directions, making a single global representation insufficient for capturing all local discriminative structures. Random subspace and random projection ensembles address this issue by aggregating classifiers trained on multiple low-dimensional views, thereby improving stability. However, these views are typically generated without using label information and may discard discriminative structure. To address the above limitations, we propose the sufficient subspace projection ensemble, abbreviated as SSP-Ensemble, for high-dimensional multiclass classification. For each class pair, SSP-Ensemble learns two supervised projection bases and combines them to form a pairwise sufficient representation. Probabilistic weak learners are then trained on these pairwise representations, and their class probability estimates are aggregated to produce the final multiclass prediction. We provide a theoretical analysis showing that, under a pairwise SDR sufficiency condition, the pairwise representations preserve the posterior probabilities used in aggregation. We further establish an excess-risk bound controlled by the pairwise subspace-estimation errors and the weak-learner log-loss regrets, which implies Bayes consistency when the sufficient subspaces and weak learners are consistently estimated. Simulation studies and real-data experiments show that SSP-Ensemble performs favorably against random projection and other popular classifiers.

1 Introduction

Multiclass classification in high-dimensional settings has attracted considerable attention in recent years, with numerous applications in genomics [Yang et al., 2022, Islam and Xing, 2023], image analysis [Dosovitskiy et al., 2021, Hassani et al., 2023, Belhasin et al., 2026], and text mining [Devlin et al., 2019, Minaee et al., 2021]. In such problems, the number of predictors often exceeds the sample size, while the information relevant for classification is often concentrated in a lower-dimensional structure, with many variables being irrelevant or noisy [Shen et al., 2022, Zeng et al., 2024]. These characteristics make classifiers trained in the original feature space unstable and prone to overfitting [Shen et al., 2022, Jia et al., 2022]. A common strategy is therefore to construct low-dimensional representations [Zeng et al., 2024, Cook and Yin, 2001] and combine the resulting classifiers through an ensemble mechanism [Yang et al., 2023, Ho, 1998].

[1] School of Mathematical Sciences, Beijing University of Posts and Telecommunications, Beijing, China.
[2] Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China.
[3] Institute of Statistics and Big Data, Renmin University of China, Beijing, China.
[4] Key Laboratory of Mathematics and Information Networks (Beijing University of Posts and Telecommunications), Ministry of Education, Beijing, China.
* Corresponding author.
† All authors contributed equally, and the authors are listed in alphabetical order by surname.
Emails: cjf2025111691@bupt.edu.cn; chengmeng@ruc.edu.cn; wpz2024@ruc.edu.cn; jyzhang2024@bupt.edu.cn; dfsxzj@ruc.edu.cn*.

STAI-X 2026 (Full Paper Track). Do not distribute.

Random subspace and random projection ensembles have been widely used for high-dimensional classification [Vrahatis et al., 2020, Qian et al., 2024, Mo et al., 2023]. By training base classifiers on multiple low-dimensional views of the data and aggregating their predictions, these methods can improve stability, reduce variance, and alleviate the computational burden of learning in the original feature space. However, the candidate subspaces or projection directions are typically generated randomly and are therefore not explicitly aligned with the label information [Qian et al., 2024]. As a result, some projected views may discard discriminative structure, especially when the informative directions are sparse [Zhang et al., 2025, Jin and Luo, 2026], weak, or class-dependent. To mitigate this issue, more adaptive variants use label information to screen or select informative subspaces before aggregation. For example, the random-projection ensemble classifier proposed by Cannings and Samworth [2017] first generates random projections and then retains projections with small estimated classification error, while RaSE [Tian and Feng, 2021] selects informative random subspaces in a similar screening spirit. Such screening improves the quality of the retained low-dimensional views, but the informative directions are still searched within a randomly generated collection of candidate subspaces.

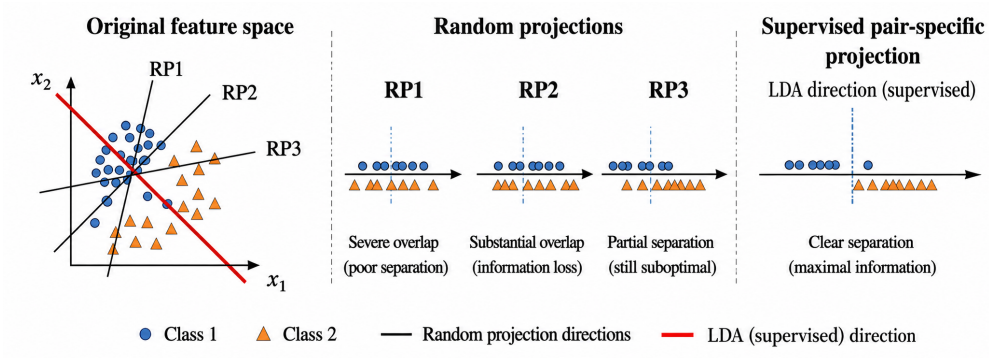

Figure 1: Illustration of random projection versus supervised projection. Random projection may miss the discriminative direction, whereas a supervised projection is aligned with the class structure.

Figure 1 illustrates why random projections may lose discriminatory information. In the original feature space (left), the red line denotes a supervised discriminative direction, whereas the black lines denote randomly generated projection directions. Since random projections are independent of class labels, the projected data may exhibit substantial class overlap when the chosen direction is poorly aligned with the discriminative structure. By contrast, supervised pair-specific projections exploit label information to construct low-dimensional representations that better preserve the relevant discriminative structure.

This limitation is more pronounced in multiclass problems, where discriminative structures are often heterogeneous across classes rather than globally shared [Hastie and Tibshirani, 1997, Belous et al., 2021, Oh and Kwak, 2024]. A direction that separates one pair of classes may be uninformative for another, and the relevant decision boundaries can vary substantially across class combinations [Allwein et al., 2000, Pawara et al., 2020]. Consequently, a single global subspace, or a collection of randomly generated subspaces, may fail to capture the local discriminative information needed for different class pairs.

Motivated by these observations, we propose SSP-Ensemble, a sufficient subspace projection ensemble framework for multiclass classification. Rather than relying on random projections, SSP-Ensemble learns supervised low-dimensional projections that are tailored to local class structures. The framework decomposes the multiclass problem into class-pair-specific views. For each class pair, two sufficient projection bases are estimated for within-pair discrimination and pair-versus-rest separation, and then combined to form a pairwise representation. A probabilistic weak learner is then trained on each pairwise representation, and the resulting class probability estimates are aggregated to produce the final multiclass decision.

This paper makes three main contributions. First, we propose SSP-Ensemble, a supervised sufficient subspace projection ensemble for high-dimensional multiclass classification. Unlike random projection ensembles, SSP-Ensemble learns supervised projections and constructs informative low-dimensional representations. Second, we develop a pairwise sufficient projection construction that

combines within-pair discrimination and pair-versus-rest separation. These projections preserve local discriminative information for each class pair and generate multiple pairwise views for ensemble aggregation. Third, we provide a statistical analysis of SSP-Ensemble. Under a pairwise SDR sufficiency condition, the population pairwise representations preserve the posterior coordinates used in probability aggregation. Furthermore, we establish a risk guarantee showing that SSP-Ensemble is Bayes consistent when the sufficient subspaces and weak learners are consistently estimated.

2 Methodology

We now introduce the proposed Sufficient Subspace Projection Ensemble, abbreviated as SSP-Ensemble. Building on the motivation above, SSP-Ensemble constructs class-pair-specific low-dimensional representations that preserve the information needed for local discrimination, and then aggregates the resulting probabilistic predictions across pairs.

Our construction is based on sufficient dimension reduction (SDR), which seeks a low-dimensional projection that preserves all information about a response. Specifically, for a predictor $X \in \mathbb{R}^p$ and a response Y, SDR aims to find a matrix $B \in \mathbb{R}^{p\times q}$ with $q \leq p$ such that

$$Y \perp\!\!\!\perp X \mid B^\top X.$$

Under this condition, $B^\top X$ preserves the response-relevant information in X. For notational convenience, we call a matrix whose columns span a projection subspace a projection basis. Given a projection basis $B \in \mathbb{R}^{p\times q}$, the corresponding low-dimensional representation of $x \in \mathbb{R}^p$ is $B^\top x$.

The overall framework of SSP-Ensemble consists of three stages. First, we enumerate all class pairs. For each pair (i, j), we learn two SDR-based projection bases: one for within-pair discrimination and the other for pair-versus-rest separation. These two bases are then combined into a pair-specific projection basis, which is applied to all observations to produce a low-dimensional representation. Second, a probabilistic weak learner is trained on each representation. Each learner operates locally on a different view of the full data, rather than on a restricted binary subset. Third, the probability estimates from all representations are aggregated class by class to form the final multiclass prediction.

Figure 2 provides an overview of the proposed framework. Stages 1 and 2 are performed on the training set, where SSP-Ensemble learns the pair-specific projection bases and trains the corresponding weak learners. At test time, Stage 3 applies the learned projection bases and trained weak learners to test observations, and aggregates the resulting probability estimates for prediction.

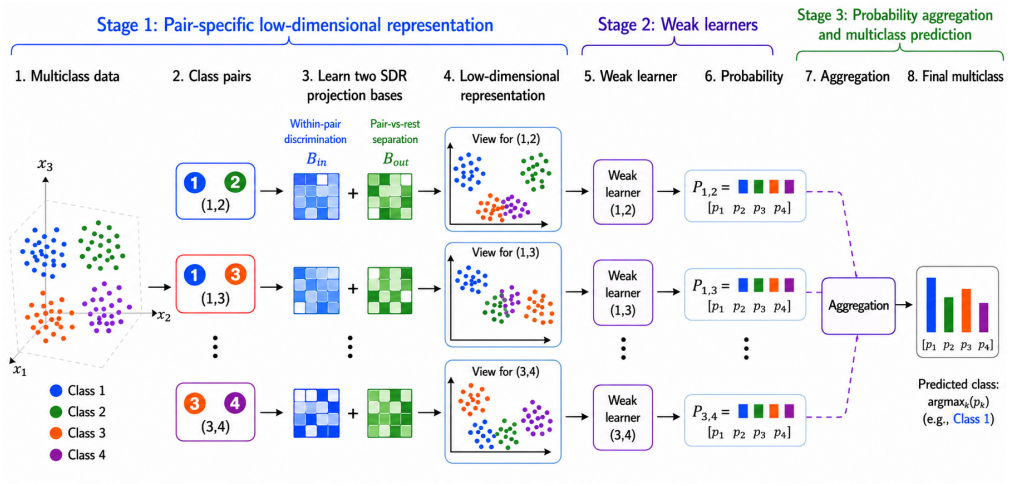

Figure 2: Overview of the proposed SSP-Ensemble framework. Stages 1 and 2 learn pair-specific representations and weak learners, while Stage 3 aggregates the resulting probability estimates for multiclass prediction.

2.1 Pairwise sufficient projection bases and low-dimensional representation

Let $\mathcal{D}_N = \{(X_i, Y_i) : i = 1, \ldots, N\}$ be an i.i.d. training sample from the distribution of a generic pair $(X, Y) \in \mathbb{R}^p \times \{1, \ldots, K\}$. Let $\mathcal{I} = \{(i, j) : 1 \le i < j \le K\}, |\mathcal{I}| = \binom{K}{2}$ denote the set of class pairs. For each pair $(i, j) \in \mathcal{I}$, SSP-Ensemble learns pair-specific projection bases from two local classification tasks.

The first task separates the target pair $\{i, j\}$ from the other classes. We define the pair-membership label $U_{ij} = \mathbf{1}\{Y \in \{i, j\}\}$. The second task distinguishes class i from class j within the target pair. Conditioning on the event $\{U_{ij} = 1\}$, we define the within-pair label $C_{ij} = \mathbf{1}\{Y = i\}$.

We estimate the two projection bases by formulating both local tasks as binary SDR problems. For each pair (i, j), we learn $\widehat{B}_{ij}^{\text{out}} \in \mathbb{R}^{p \times q_1}$ from the pair-versus-rest label U_{ij}, and learn $\widehat{B}_{ij}^{\text{in}} \in \mathbb{R}^{p \times q_2}$ from the within-pair label C_{ij} restricted to samples with $Y \in \{i, j\}$, where $q_1, q_2 < p$. In our implementation, both bases are estimated by POTD, which is designed to estimate sufficient subspaces for classification [Meng et al., 2020, Ouyang et al., 2025]. The final projection basis, which we refer to as the pairwise sufficient projection basis, is obtained by combining and orthonormalizing these two bases:

$$\widehat{W}_{ij} = \operatorname{orth}\left([\widehat{B}_{ij}^{\text{out}}, \widehat{B}_{ij}^{\text{in}}]\right).$$

The corresponding low-dimensional representation is $\widehat{Z}_{ij} = \widehat{W}_{ij}^{\top} X$. This representation is designed to capture both the information needed to identify whether an observation belongs to the target pair and the information needed to distinguish the two classes within the pair.

Apart from POTD, other SDR methods can also be used provided that they provide suitable subspace estimation for binary classification tasks.

2.2 A brief overview of POTD

SSP-Ensemble uses POTD [Meng et al., 2020] as the default SDR estimator for the local binary classification tasks. POTD is a supervised dimension-reduction method designed for classification problems with categorical responses. Its main idea is to identify projection directions along which the conditional distributions of the covariates differ most across response classes.

For a binary classification problem, POTD compares the empirical conditional distributions of X in the two classes through optimal transport. The optimal transport map induces displacement vectors that describe how observations from one class are transported toward the other class. These displacement vectors summarize the class-conditional distributional changes of X. POTD then forms a displacement matrix from these vectors and applies singular value decomposition to this matrix. The leading right singular vectors are used as the estimated supervised projection directions. POTD can also be applied in high-dimensional or large-scale settings, where scalable OT techniques are needed to estimate the underlying transport. Such techniques include projection-based map estimation [Meng et al., 2019, Zhang et al., 2023, Li et al., 2024, 2025], and sampling-based sparsification methods [Li et al., 2023a,b, Hu et al., 2025, Ouyang et al., 2026].

Compared with classical SDR methods such as SIR and SAVE, POTD is particularly suitable for classification tasks because it directly exploits differences between class-conditional distributions. Classical inverse-regression-based SDR methods often rely on moment-based summaries, such as conditional means or conditional variances, which may be insufficient when the class differences are nonlinear, distributional, or not well captured by low-order moments. In contrast, POTD uses optimal transport to compare the empirical distributions of different classes and extracts directions from the resulting displacement structure. Therefore, it can capture more general class-discriminative distributional changes beyond simple mean or variance differences.

This property makes POTD suitable for SSP-Ensemble because both local tasks in our framework are binary SDR problems. The pair-versus-rest task separates the target pair $\{i, j\}$ from the remaining classes, and the within-pair task distinguishes class i from class j. In both cases, POTD identifies low-dimensional directions that preserve class-discriminative distributional changes. The theoretical analysis further uses the subspace convergence property of POTD to control the estimation error of the learned pairwise projection bases. Detailed algorithmic construction and theoretical properties of POTD can be found in [Meng et al., 2020].

2.3 Weak learners on pairwise representations

For each pair $(i, j) \in \mathcal{I}$, we train a probabilistic weak learner on the learned representation $\widehat{Z}_{ij}$. Specifically, the weak learner takes the form $\widehat{f}_{ij} : \mathbb{R}^{d_{ij}} \to \Delta^{K-1}$, where d_{ij} is the dimension of $\widehat{Z}_{ij}$ and Δ^{K-1} is the probability simplex over K classes. In this paper, we use multinomial logistic regression, k-nearest neighbors (KNN), and support vector machines (SVM) with calibrated probabilities as weak learners, because each produces class probability estimates. For a test point x, we write its fitted probability output for class k as

$$\widehat{\pi}_k^{(ij)}(x) = \widehat{f}_{ij,k}(\widehat{W}_{ij}^\top x), \qquad k = 1, \ldots, K.$$

Although the representation is constructed through the specific pair (i, j), the weak learner outputs a K-dimensional probability vector. In the final aggregation step, only the two coordinates corresponding to classes i and j are used. This keeps all pairwise learners on a common multiclass probability scale.

Algorithm 1 SSP-Ensemble

Input: $\mathcal{D}_N = \{(X_i, Y_i) : i = 1, \ldots, N\}$, the projection dimensions q_1, q_2, and a probabilistic weak learner.
Construct the class-pair set $\mathcal{I} = \{(i, j) : 1 \le i < j \le K\}, |\mathcal{I}| = \binom{K}{2}$.
for each pair $(i, j) \in \mathcal{I}$ **do**
  Define $U_{ij} = \mathbf{1}\{Y \in \{i, j\}\}$ and $C_{ij} = \mathbf{1}\{Y = i\}$ on $\{Y \in \{i, j\}\}$.
  $\widehat{B}_{ij}^{\text{out}} = \text{POTD}_{q_1}(X, U_{ij})$.
  $\widehat{B}_{ij}^{\text{in}} = \text{POTD}_{q_2}(X_{\{i,j\}}, C_{ij})$.
  $\widehat{W}_{ij} = \text{orth}([\widehat{B}_{ij}^{\text{out}}, \widehat{B}_{ij}^{\text{in}}])$.
  Train a probabilistic weak learner $\widehat{f}_{ij}$ on $\widehat{W}_{ij}^\top X$.
end for
For a test point x, compute $\widehat{s}_k(x) = \frac{1}{K-1} \sum_{j:j\neq k} \widehat{\pi}_k^{(k,j)}(x)$.
Output $\widehat{g}(x) = \arg\max_k \widehat{s}_k(x)$.

2.4 Probability aggregation

The final multiclass prediction is obtained by aggregating the probability outputs from all pairs involving each class. For class k, define the aggregated score

$$\widehat{s}_k(x) = \frac{1}{K-1} \sum_{j:j\neq k} \widehat{\pi}_k^{(k,j)}(x),$$

where the superscript (k, j) denotes the stored class pair containing k and j. The final classifier is

$$\widehat{g}(x) = \arg \max_{1\le k\le K} \widehat{s}_k(x).$$

The aggregation is performed at the level of estimated class probabilities rather than hard votes. Thus, the score for class k only combines posterior estimates from pairwise representations that contain class k. Algorithm 1 summarizes the full procedure of SSP-Ensemble.

3 Theoretical Analysis

In this section, we provide a statistical analysis of the proposed SSP-Ensemble classifier. We use the notation introduced in Section 2, including the class-pair set $\mathcal{I}$, the local labels U_{ij} and C_{ij}, and the learned pairwise representation $\widehat{W}_{ij}^\top X$. For theoretical analysis, we additionally define the collapsed three-way label

$$T_{ij} = \begin{cases} i, & Y = i, \\ j, & Y = j, \\ o, & Y \notin \{i, j\}, \end{cases}$$

where o denotes all "other" classes.

Let $\eta_k(x) = \mathbb{P}(Y = k \mid X = x)$, $k = 1, \ldots, K$, denote the multiclass posterior probabilities. The Bayes classifier is $g^*(x) = \min\{k : \eta_k(x) = \max_{1 \le j \le K} \eta_j(x)\}$, and its risk is denoted by $R^* = \mathbb{P}\{g^*(X) \neq Y\}$.

3.1 Sufficient Representations and Probability Aggregation

Before studying the finite-sample behavior of SSP-Ensemble, we first examine its population-level representation. We formalize this requirement through the following pairwise SDR sufficiency condition.

The standard SDR condition [Li, 1991, Cook, 2009] assumes that, for a predictor $X \in \mathbb{R}^p$ and a response Y, there exists a projection basis $W \in \mathbb{R}^{p \times q}$, with $q \le p$, such that

$$Y \perp\!\!\!\perp X \mid W^\top X.$$

Under this condition, the conditional distribution of the response is preserved after projecting X onto the subspace spanned by W. We refer to such a matrix W as a sufficient projection basis. The condition includes a trivial full-dimensional case. When $q = p$, any full-rank $W \in \mathbb{R}^{p \times p}$ gives an invertible transformation $W^\top X$, which contains exactly the same information as X.

In SSP-Ensemble, instead of imposing the above condition on the original multiclass response globally, we apply the same SDR principle to the collapsed three-way label T_{ij} for each class pair. This leads to the following pairwise SDR sufficiency condition.

Assumption 1 (Pairwise SDR sufficiency)**.** *For every pair $(i, j) \in \mathcal{I}$, there exists a pairwise projection basis $W_{ij}^* \in \mathbb{R}^{p \times q_{ij}}$, with $q_{ij} \le p$, such that*

$$T_{ij} \perp\!\!\!\perp X \mid (W_{ij}^*)^\top X.$$

Assumption 1 is a direct pairwise analogue of the standard SDR condition. It only requires the existence of a sufficient projection basis for each collapsed three-way response T_{ij}. The case $q_{ij} = p$ is trivially valid, while the meaningful dimension-reduction case is $q_{ij} \ll p$.

Proposition 1 (Posterior preservation by pairwise sufficient representations)**.** *Under Assumption 1, for every pair $(i, j) \in \mathcal{I}$ and $k \in \{i, j\}$,*

$$\mathbb{P}\{Y = k \mid (W_{ij}^*)^\top X\} = \mathbb{P}(Y = k \mid X) = \eta_k(X) \quad \textit{a.s.}$$

Proposition 1 shows that the pairwise sufficient representation preserves the posterior coordinates needed for aggregation and therefore introduces no representation bias for SSP-Ensemble.

Corollary 1 (Probability aggregation recovers the Bayes scores)**.** *Define the aggregated probability score*

$$s_k^*(X) = \frac{1}{K-1} \sum_{j : j \neq k} \mathbb{P}\{Y = k \mid (W_{kj}^*)^\top X\},$$

where (k, j) again denotes the pair containing classes k and j. By Proposition 1, $s_k^*(X) = \eta_k(X)$ for $k = 1, \ldots, K$. Therefore,

$$\arg \max_{1 \le k \le K} s_k^*(X) = g^*(X) \quad \textit{a.s.}$$

Corollary 1 shows that under the pairwise SDR condition, the aggregated score coincides exactly with the original multiclass posterior probability. Hence, the aggregation rule recovers the Bayes scores and does not contribute any additional bias at the population level.

3.2 Excess risk bound

In practice, neither the sufficient projection basis W_{ij}^* nor the true posterior probabilities are available. We therefore work with their estimators. For each pair $(i, j) \in \mathcal{I}$, let $\widehat{W}_{ij}$ denote the estimated projection basis, and define the corresponding projection matrix by $\widehat{P}_{ij} = \widehat{W}_{ij} \widehat{W}_{ij}^\top$.

We first clarify how the final excess risk is controlled. For a matrix $W \in \mathbb{R}^{p\times d_{ij}}$, we define the posterior coordinate induced by the projected representation $W^\top X$ as

$$\widetilde{\eta}^{W}_{k,ij}(X) = \mathbb{P}(Y = k \mid W^\top X).$$

For each pair $(i,j) \in \mathcal{I}$, we define the relevant pairwise posterior error by

$$E_{ij} = \mathbb{E}\left[\left|\widehat{\pi}^{(ij)}_i(X) - \eta_i(X)\right| + \left|\widehat{\pi}^{(ij)}_j(X) - \eta_j(X)\right|\right].$$

We decompose the pairwise posterior error into three components: weak-learner error, subspace-estimation error, and representation error. The representation component is zero by Proposition 1. The full decomposition is provided in the supplementary material.

We next relate the pairwise coordinate errors to the excess risk of the final multiclass classifier. A standard plug-in comparison argument [Devroye et al., 2013] gives

$$R(\widehat{g}) - R^* \le \frac{2}{K-1}\sum_{(i,j)\in\mathcal{I}} E_{ij}.$$

Thus, controlling the pairwise posterior errors E_{ij} is sufficient for controlling the final multiclass excess risk.

Since the projection bases in SSP-Ensemble are estimated by POTD, the required subspace convergence can be inherited from the existing POTD theory [Meng et al., 2020].

Proposition 2 (POTD subspace convergence)**.** *Suppose the regularity conditions required for the POTD subspace convergence theorem hold for both the pair-versus-rest task and the within-pair task. Then, for every pair $(i,j) \in \mathcal{I}$,*

$$\mathbb{E}\|\widehat{P}_{ij} - P^*_{ij}\|_F \le c_{ij} N_{ij}^{-1/p},$$

where N_{ij} is the effective sample size for estimating the pairwise projection basis.

This proposition imports the POTD subspace convergence rate into SSP-Ensemble. To translate subspace convergence into posterior convergence, we impose the following stability condition.

Assumption 2 (Posterior stability)**.** *For every pair $(i,j) \in \mathcal{I}$ and each $k \in \{i,j\}$, there exists a constant $L_{ij} > 0$ such that*

$$\mathbb{E}\left[\left|\widetilde{\eta}^{\widehat{W}_{ij}}_{k,ij}(X) - \widetilde{\eta}^{W^*_{ij}}_{k,ij}(X)\right|\right] \le L_{ij}\,\mathbb{E}\|\widehat{P}_{ij} - P^*_{ij}\|_F.$$

Assumption 2 can be viewed as a Lipschitz-type regularity condition linking subspace perturbation to posterior perturbation. Together with Proposition 2, this condition controls the subspace-estimation term in posterior error decomposition at the rate $N_{ij}^{-1/p}$.

For the weak-learner term, let $r_{ij,N}$ denote the excess population log-loss regret of $\widehat{f}_{ij}$ on the learned representation $\widehat{W}_{ij}^\top X$, relative to the optimal posterior predictor based on this representation, and assume $r_{ij,N} \to 0$. The standard conversion from log-loss regret to posterior ℓ_1 error gives

$$\mathbb{E}\left[\left|\widehat{\pi}^{(ij)}_i(X) - \widetilde{\eta}^{\widehat{W}_{ij}}_{i,ij}(X)\right| + \left|\widehat{\pi}^{(ij)}_j(X) - \widetilde{\eta}^{\widehat{W}_{ij}}_{j,ij}(X)\right|\right] \le \sqrt{2r_{ij,N}}.$$

The formal statement and proof are deferred to the supplementary material.

Theorem 1 (Excess risk of SSP-Ensemble)**.** *Under the pairwise SDR sufficiency condition, the POTD subspace convergence bound in Proposition 2, the posterior stability condition in Assumption 2, and the weak-learner log-loss regret condition, we have*

$$R(\widehat{g}) - R^* \le \frac{2}{K-1}\sum_{(i,j)\in\mathcal{I}}\left(2L_{ij}c_{ij}N_{ij}^{-1/p} + \sqrt{2r_{ij,N}}\right).$$

When L_{ij}, c_{ij}, and $r_{ij,N}$ are uniformly bounded over pairs and $N_{ij} \asymp N$, the bound reduces to

$$R(\widehat{g}) - R^* = O\left(KN^{-1/p} + K\sqrt{\bar{r}_N}\right),$$

which implies Bayes consistency whenever $N^{-1/p} \to 0$ and $\bar{r}_N \to 0$.

Theorem 1 shows that the excess risk is controlled by the POTD subspace-estimation rate and the weak-learner log-loss regret. No additional representation term appears, because Proposition 1 shows that the population pairwise representation preserves the posterior coordinates used in aggregation.

4 Simulation Studies

We conduct simulation studies under high-dimensional classification settings to evaluate the performance of SSP-Ensemble. The two settings cover weak signals with strong noise, nonlinear multiclass structures, class imbalance, and heterogeneous latent discriminative directions. An additional setting comparing random projection with SSP-Ensemble is reported in the supplementary material. For each setting, we randomly split the data into 80% training and 20% test samples, and repeat the experiment over ten independent runs. We report the mean classification accuracy with standard deviation.

In Setting 1, we consider a five-class problem with imbalanced class sizes $(300, 300, 300, 100, 100)$ and dimension $p = 100$. Write $X = (X_{\text{sig}}^\top, X_{\text{noise}}^\top)^\top$, $X_{\text{sig}} \in \mathbb{R}^{10}$, $X_{\text{noise}} \in \mathbb{R}^{90}$. The signal part is generated from a two-dimensional latent variable $Z \in \mathbb{R}^2$: the first three classes form concentric rings, while the last two classes are Gaussian clusters. The latent variable Z is expanded into ten signal coordinates, and the remaining coordinates are generated as independent Gaussian noise.

In Setting 2, we consider a three-class problem with $K = 3$, $n = 1500$, and $p = 10$. The labels are determined by two latent discriminative directions, $Z_1 = 0.8X_1 + 0.6X_2, Z_2 = -0.4X_3 + 0.2X_4 + 0.9X_5$. The class label is assigned according to the sign pattern of (Z_1, Z_2):

$$Y = \begin{cases} 0, & Z_1 < 0,\ Z_2 < 0, \\ 1, & Z_1 \geq 0,\ Z_2 \geq 0, \\ 2, & Z_1 Z_2 < 0. \end{cases}$$

We compare SSP-Ensemble with standard classifiers trained in the original feature space, including logistic regression (LR), KNN, SVM, random forest (RF), and XGBoost. For SSP-SVM, we use SVM with calibrated class probability outputs, and these probabilities are used in the final probability aggregation step. We also include projection- and subspace-based competitors, including RaSE [Tian and Feng, 2021], and the random-projection ensemble classifier of Cannings and Samworth [2017], denoted by JRSSB-RP. For SSP-Ensemble we examine different choices of weak learners, which we refer to as SSP-KNN, SSP-LR, and SSP-SVM. Details are provided in the supplementary material.

Table 1: Classification accuracy in simulation studies. Entries are mean accuracy with standard deviation in parentheses.

Setting	SSP-LR	SSP-KNN	SSP-SVM	JRSSB-RP	RaSE	LR	KNN	SVM	RF	XGBoost
Setting 1	40.59(3.72)	90.77(1.70)	**92.27(1.63)**	64.45(3.12)	91.64(1.53)	40.18(1.62)	51.36(2.94)	59.41(3.32)	91.59(2.26)	91.91(1.17)
Setting 2	72.67(2.85)	91.63(1.39)	**91.73(1.86)**	76.70(2.74)	90.17(1.91)	72.57(2.15)	71.77(3.01)	85.80(1.53)	89.03(2.31)	89.07(2.11)

Table 1 reports the classification results. Across both settings, SSP-SVM achieves the highest accuracy, supporting the effectiveness of the proposed SSP-Ensemble framework when combined with a nonlinear weak learner. Compared with JRSSB-RP, both SSP-KNN and SSP-SVM obtain substantially higher accuracy. This suggests that SSP can preserve discriminative information more effectively than random projection in these settings. In particular, SSP-KNN and SSP-SVM substantially outperform the corresponding KNN and SVM classifiers trained in the original feature space.

The relatively poor performance of SSP-LR and LR is also consistent with the data-generating mechanisms. Both settings involve nonlinear class structures, so linear classifiers are not sufficiently flexible to capture the underlying decision boundaries even after dimension reduction. By contrast, SSP-SVM benefits from both the supervised pairwise sufficient projections and the nonlinear classification ability of SVM.

5 Real Data Applications

We further evaluate SSP-Ensemble on real-world datasets from the UCI Machine Learning Repository and OpenML [Asuncion and Newman, 2007, Vanschoren et al., 2014]. The datasets include high-dimensional sparse text data and heterogeneous numerical feature representations. Table 2 summarizes the sample size, dimension, and number of classes for each dataset.

CNAE-9 and 20Newsgroups are high-dimensional sparse text datasets. The 20Newsgroups dataset originally contains documents from 20 topic categories. To construct a moderately sized multiclass

Table 2: Summary of real-data datasets.

Dataset	Number of samples	Dimension	Number of classes
CNAE-9	1080	856	9
MFEAT	2000	649	10
20Newsgroups	9351	2000	10

high-dimensional text classification task, we randomly select 10 categories. After vectorization, the resulting dataset contains 9,351 observations and 2,000 features. For reproducibility, the selected categories and the additional results on the full 20-class dataset are reported in the supplementary material. MFEAT contains handwritten digit samples represented by multiple feature groups.

We compare SSP-Ensemble with the same set of baseline methods as in the simulation studies, and all methods are evaluated using stratified five-fold cross-validation.

Table 3: Real data results: classification accuracy. Entries are mean accuracy with standard deviation in parentheses.

Dataset	SSP-LR	SSP-KNN	SSP-SVM	JRSSB-RP	RaSE	LR	KNN	SVM	RF	XGBoost
CNAE-9	**95.45(0.83)**	90.37(1.44)	94.54(1.44)	90.74(2.36)	77.22(3.33)	90.83(2.69)	84.81(3.81)	85.83(2.75)	91.48(1.93)	90.56(1.45)
MFEAT	**98.25(1.08)**	96.55(1.08)	98.15(0.80)	97.15(1.17)	97.30(0.29)	97.85(0.93)	96.65(0.78)	97.75(0.47)	96.50(0.83)	96.35(0.99)
20Newsgroups	73.71(1.32)	66.96(1.12)	**74.55(0.95)**	71.80(1.58)	50.99(2.00)	63.73(2.01)	20.37(1.11)	72.76(1.26)	68.96(1.24)	70.88(1.24)

The results are summarized in Table 3. SSP-Ensemble achieves the highest accuracy on all three datasets, with SSP-LR performing best on CNAE-9 and MFEAT, and SSP-SVM performing best on 20Newsgroups. Compared with JRSSB-RP and RaSE, SSP-Ensemble consistently yields higher accuracy on these datasets. This suggests that supervised sufficient projections can preserve more discriminative information than randomly generated or screened subspaces on these datasets.

The comparison with the original-space classifiers further supports the benefit of the SSP representation. For each weak learner, the corresponding SSP variant improves over its direct counterpart, indicating that the learned pairwise representations provide more informative inputs for classification than the original features. This is consistent with the motivation discussed in the introduction. In multiclass problems, discriminative structures may vary across class pairs, and a single classifier trained on the full feature space may not effectively capture these local structures. By constructing pairwise sufficient projections, SSP-Ensemble better preserves the relevant local discriminative information in these datasets.

SSP-Ensemble also outperforms RF and XGBoost on all three datasets. These results indicate that the proposed framework benefits not only from ensemble aggregation, but also from constructing supervised pairwise representations. These experiments support the effectiveness of SSP-Ensemble for high-dimensional multiclass classification.

6 Conclusion

In this paper, we proposed SSP-Ensemble, a sufficient subspace projection ensemble framework for high-dimensional multiclass classification. The method learns supervised class-pair-specific projections by combining within-pair discrimination and pair-versus-rest separation, and aggregates probabilistic weak learners trained on the resulting pairwise representations.

We showed that, under a pairwise SDR sufficiency condition, the population pairwise representations preserve the posterior coordinates used for aggregation. An excess-risk bound further implies Bayes consistency when the sufficient subspaces and weak learners are consistently estimated.

Simulation and real-data results show that SSP-Ensemble performs favorably against random projection, RaSE, and standard classifiers, particularly when discriminative structures are low-dimensional and class-dependent. Future work includes incorporating other supervised SDR estimators, adaptively selecting projection dimensions and weak learners, and extending the framework to structured data such as time series or functional observations.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, No. 12301381, and the National Key R&D Program of China, No. 2021YFA1001300.

References

Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of machine learning research*, 1(Dec):113–141, 2000.

A. Asuncion and D.H. Newman. UCI machine learning repository, 2007. URL `http://archive.ics.uci.edu/ml`.

Omer Belhasin, Shelly Golan, Ran El-Yaniv, and Michael Elad. Advancing image classification with discrete diffusion classification modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 124–134, 2026.

Gregg Belous, Andrew Busch, and Yongsheng Gao. Dual subspace discriminative projection learning. *Pattern Recognition*, 111:107581, 2021. doi: 10.1016/j.patcog.2020.107581.

Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):959–1035, 2017.

R Dennis Cook. *Regression graphics: Ideas for studying regressions through graphics*. John Wiley & Sons, 2009.

R. Dennis Cook and Xiangrong Yin. Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, 43(2):147–199, 2001. doi: 10.1111/1467-842X.00164.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6185–6194, 2023.

Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *Advances in neural information processing systems*, 10, 1997.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

Yukuan Hu, Mengyu Li, Xin Liu, and Cheng Meng. Sampling-based methods for multi-block optimization problems over transport polytopes. *Mathematics of Computation*, 94(353):1281–1322, 2025. doi: 10.1090/mcom/3989.

Md Tauhidul Islam and Lei Xing. Cartography of genomic interactions enables deep analysis of single-cell expression data. *Nature Communications*, 14(1):679, 2023.

Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduction: A review. *Complex & Intelligent Systems*, 8(3):2663–2693, 2022. doi: 10.1007/s40747-021-00637-x.

Yin Jin and Wei Luo. A unified generalization of the inverse regression methods via column selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 88(1):23–42, 2026. doi: 10.1093/jrsssb/qkaf038.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Mengyu Li, Jun Yu, Tao Li, and Cheng Meng. Importance sparsification for sinkhorn algorithm. *Journal of Machine Learning Research*, 24(247):1–44, 2023a.

Mengyu Li, Jun Yu, Hongteng Xu, and Cheng Meng. Efficient approximation of gromov-wasserstein distance using importance sparsification. *Journal of Computational and Graphical Statistics*, 32 (4):1512–1523, 2023b. doi: 10.1080/10618600.2023.2165500.

Tao Li, Cheng Meng, Hongteng Xu, and Jun Yu. Hilbert curve projection distance for distribution comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4993–5007, 2024. doi: 10.1109/TPAMI.2024.3363780.

Tao Li, Cheng Meng, Hongteng Xu, and Jun Zhu. Efficient variants of wasserstein distance in hyperbolic space via space-filling curve projection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):16814–16824, 2025.

Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. Large-scale optimal transport map estimation using projection pursuit. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Cheng Meng, Jun Yu, Jingyi Zhang, Ping Ma, and Wenxuan Zhong. Sufficient dimension reduction for classification using principal optimal transport direction. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4015–4028. Curran Associates, Inc., 2020.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys*, 54(3):1–40, 2021. doi: 10.1145/3439726.

Tianlan Mo, Linjing Wang, Yuliang Wu, Junrong Huang, Weikun Liu, Ruimeng Yang, and Xin Zhen. Classifier ensemble with evolutionary optimisation enforced random projections. *Expert Systems with Applications*, 222:119845, 2023.

Jiyong Oh and Nojun Kwak. Discriminative subspace learning using generalized mean. *Signal Processing*, 219:109421, 2024. doi: 10.1016/j.sigpro.2024.109421.

Wenbo Ouyang, Ruiyang Wu, Ning Hao, and Hao Helen Zhang. Dynamic supervised principal component analysis for classification. *Journal of Computational and Graphical Statistics*, 34(4): 1446–1455, 2025. doi: 10.1080/10618600.2025.2452935.

Xiaxue Ouyang, Hao Zheng, Haoxian Liang, Jingyi Zhang, Yixuan Qiu, Cheng Meng, and Mengyu Li. Sparsification techniques for large-scale optimal transport problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 18(1):e70056, 2026. doi: 10.1002/wics.70056.

Pornntiwa Pawara, Emmanuel Okafor, Marc Groefsema, Sheng He, Lambert R. B. Schomaker, and Marco A. Wiering. One-vs-one classification for deep neural networks. *Pattern Recognition*, 108: 107528, 2020. doi: 10.1016/j.patcog.2020.107528.

Xiaoyu Qian, Jinru Wu, Ligong Wei, and Youwu Lin. Random projection ensemble conformal prediction for high-dimensional classification. *Chemometrics and Intelligent Laboratory Systems*, 253:105225, 2024.

Liran Shen, Meng Joo Er, and Qingbo Yin. Classification for high-dimension low-sample size data. *Pattern Recognition*, 130:108828, 2022. doi: 10.1016/j.patcog.2022.108828.

Ye Tian and Yang Feng. Rase: Random subspace ensemble classification. *Journal of Machine Learning Research*, 22(45):1–93, 2021.

Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Aristidis G Vrahatis, Sotiris K Tasoulis, Spiros V Georgakopoulos, and Vassilis P Plagianakos. Ensemble classification through random projections for single-cell rna-seq data. *Information*, 11 (11):502, 2020.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4:852–866, 2022. doi: 10.1038/s42256-022-00534-z.

Yongquan Yang, Haijun Lv, and Ning Chen. A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, 56(6):5545–5589, 2023. doi: 10.1007/s10462-022-10283-5.

Jing Zeng, Qing Mai, and Xin Zhang. Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction. *Journal of the American Statistical Association*, 119 (545):343–355, 2024. doi: 10.1080/01621459.2022.2118601.

Jia Zhang, Runxiong Wu, and Xin Chen. Optimal sparse sliced inverse regression via random projection. *Journal of Computational and Graphical Statistics*, 35(2):782–794, 2025. doi: 10.1080/10618600.2025.2571161.

Jingyi Zhang, Ping Ma, Wenxuan Zhong, and Cheng Meng. Projection-based techniques for high-dimensional optimal transport problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(2):e1587, 2023. doi: 10.1002/wics.1587.