Boosting Ensemble Accuracy by Revisiting Ensemble Diversity Metrics (Supplementary Material)

Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka-Ho Chow, Wenqi Wei
School of Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30332
yanzhaowu@gatech.edu, lingliu@cc.gatech.edu, {zhongweixie, khchow, wenqiwei}@gatech.edu

In the main paper, we analyzed the inherent problems of using Q-diversity metrics to select high quality ensembles. In comparison, our FQ-diversity metrics can (1) separately measure and compare the ensemble teams of equal size, (2) leverage the negative samples from the focal model to measure ensemble diversity, and (3) partition the candidate ensemble teams by using binary clustering with strategically selected initial centroids. These optimizations enable FQ-diversity metrics to more accurately capture the failure independence among the member models of ensemble teams, and efficiently select high quality ensemble teams. Furthermore, we further improve the quality of selected ensemble teams by introducing EQ diversity metrics to combine the top performing FQ metrics. The source codes are provided at https://github.com/git-disl/DP-Ensemble.

In this supplementary material, we provide additional materials and technical details. We organize the supplementary material into five sections: (1) Additional examples and visualization to illustrate the strength of FQ diversity metrics over Q diversity metrics in selecting good quality ensembles. (2) Diversity analysis in terms of uncorrelated errors. (3) Definition of three pairwise and three non-pairwise Q-diversity metrics, which is the basis for defining their corresponding FQ-diversity metrics. (4) Algorithm for Q ensemble selection, which is introduced in Section 2 of the main paper. (5) Algorithm for FQ ensemble selection, which supplements Section 3 of the main paper.

1. Additional Experiments for FQ metrics

1.1. Case Study on $EnsSet(\mathbf{F_5}, S = 5)$, ImageNet

Table 1 shows the comparison of selecting high quality ensemble teams from the set of candidate ensembles $EnsSet(F_5, S = 5)$ with team size S and focal model F_5 (ResNet152) by FQ-GD and Q-GD metrics. We list 10 example ensemble teams in $EnsSet(F_5, S = 5)$ in this table. The green check mark indicates that the ensemble team is selected and the red cross implies that the team is pruned out, according to the corresponding diversity metric and the binary partitioning threshold. In this case, FQ-GD uses K-means to perform binary partitioning and to determine the diversity threshold. Q-GD uses the mean GD-diversity value as the threshold for binary partitioning. It clearly shows that our FQ-GD (measured on a set of randomly selected samples from $NegSampSet(F_5)$) can successfully identify high accuracy ensemble teams and avoid these ensemble teams with low accuracy. In comparison, the Q-GD metric (measured on a set of random samples from $NegSampSet = \bigcup_{i=0}^{M-1} NegSampSet(F_i)$ over all M base models) fails to capture these high accuracy ensemble teams. Instead, many low accuracy ensemble teams are selected by Q-GD, such as 02578 and 02567 with the ensemble accuracy below m_max=78.25%, the max accuracy of the member models in the above two ensemble teams, which both have F_5 as a member model. This observation also explains the worse ensemble accuracy lower bound of the ensemble teams selected by Q diversity metrics, comparing to our FQ metrics.

1.2. Case Study on $EnsSet(\mathbf{F_2}, S = 5)$, ImageNet

For a comparison purpose, we replace the focal model F_5 (ResNet152)) with F_2 (EfficentNet-B0) and obtain the set of candidate ensemble teams, $EnsSet(F_2, S = 5)$. Table 2 lists all of the 7 ensemble teams from Table 1, which contain F_2 as their member model, and hence are included in $EnsSet(F_2, S = 5)$. Comparing Table 1 and Table 2, we observe that different focal models may select different ensemble teams, this is mainly due to two reasons: (1) when changing the focal model from F_5 to F_2, some ensemble teams will naturally be removed if they do not have F_2 as a member model, such as the ensemble team 13459. (2) FQ diversity scores are measured based on the negative samples from the chosen focal model. Hence, when changing the focal model from F_5 to F_2, some ensemble teams that may

Ensemble Team	12345	13459	13458	12458	23458	02357	23567	03579	02578	02567
Ensemble Accuracy	80.77	80.63	80.50	80.43	80.44	79.24	79.19	78.91	77.96	77.64
FQ-GD	0.860	0.865	0.866	0.863	0.863	0.884	0.883	0.892	0.893	0.896
(<0.882)	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
Q-GD	0.706	0.705	0.691	0.717	0.729	0.654	0.661	0.615	0.659	0.655
(<0.665)	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓

Table 1: 10 Ensemble Examples on ImageNet (S=5, $focal$=5)

Ensemble Team	12345	12458	23458	02357	23567	02578	02567
Ensemble Accuracy	80.77	80.43	80.44	79.24	79.19	77.96	77.64
FQ-GD	0.834	0.837	0.836	0.857	0.855	0.871	0.875
(<0.858)	✓	✓	✓	✓	✓	✗	✗
Q-GD	0.706	0.717	0.729	0.654	0.661	0.659	0.655
(<0.665)	✗	✗	✗	✓	✓	✓	✓

Table 2: 7 Ensemble Examples on ImageNet (S=5, $focal$=2)

have large diversity scores (poor diversity in terms of failure independence) under focal model F_5 but achieve small diversity scores (good failure independence) under focal model F_2. For example, the two ensemble teams: `02357` and `23567` both are outperforming all member models, and are selected by FQ-GD with the focal model F_2 (Table 2). Interestingly, both of them were not selected when the focal model is F_5 (Table 1). This further demonstrates by example that for each ensemble team of size S, we will compute S number of diversity scores, one for each of the S focal models. Then we combine these S scores to obtain the ensemble diversity for this ensemble team (recall Section 3 Step (4) in the main paper).

1.3. Case Study on Different FQ metrics, ImageNet

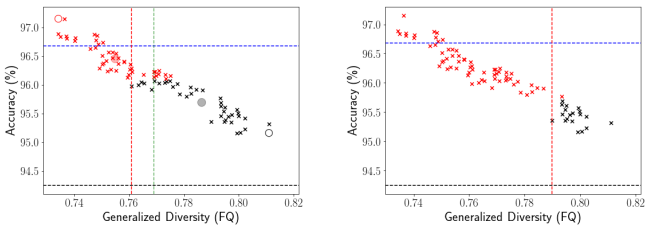
We observe that different FQ diversity metrics can select different ensemble teams. We list the statistics for the ensemble teams selected by all FQ metrics (set union) and uniquely selected by FQ-QS, FQ-KW and FQ-GD in Table 3. Different diversity metrics measure the ensemble diversity from different perspectives with different design principles (see Section 3 in this supplementary material). Therefore, different diversity metrics can capture different ensemble teams. From Table 3, most of these ensemble teams that are uniquely selected by different FQ metrics show high quality and outperform all member models. By combining all six FQ metrics, we obtain a set of 641 good ensemble teams, among which 616 teams (96%) can outperform all member models and 569 teams (89%) can achieve ensemble accuracy higher than p_max=78.25%. This observation shows that combining different FQ metrics can further improve the quality of selected ensemble teams, which motivates our EQ metrics.

1.4. Case Study on the Thresholds, CIFAR-10

We use a binary clustering algorithm to identify the FQ diversity thresholds for selecting high quality ensemble

Methods	#Teams	# (Acc >= m_max)	# (Acc >= 78.25% p_max)
All FQ metrics	641	616	569
Unique by FQ-QS	29	21	12
Unique by FQ-KW	7	7	7
Unique by FQ-GD	24	24	23

Table 3: Ensembles Selected by Different FQ metrics (ImageNet)



(a) K-means v.s. Mean (b) Agglomerative Clustering

Figure 1: Comparison of Different Thresholds (CIFAR-10, S=4, $focal$=2, FQ-GD)

teams. Figure 1 shows the thresholds identified by different methods on CIFAR-10 for the candidate ensemble set $EnsSet(F_2, S = 4)$. First, in Figure 1a, we compare the K-means threshold (0.761, red vertical dashed line), with $K = 2$ and two initial centroids chosen as shown by the two unfilled circles, and the mean threshold (0.769, green vertical dashed line). It is visually clear that the K-means threshold can prune out more low quality ensemble teams than the mean threshold. We further use the agglomerative clustering, a popular bottom-up hierarchical clustering algorithm for binary partitioning, and for identifying the diversity threshold as shown in Figure 1b. The corresponding diversity threshold is 0.790, which is much worse than the K-mean threshold, mainly due to the lack of optimization

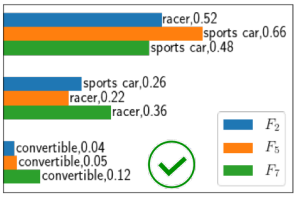
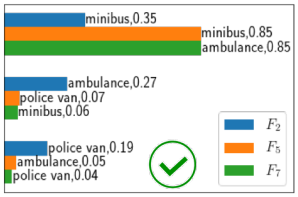
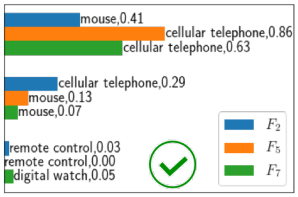
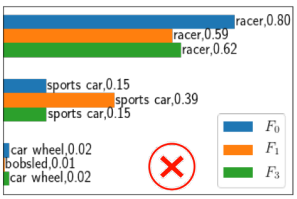
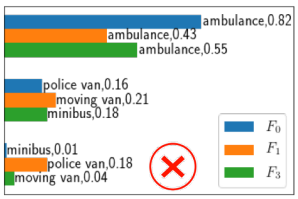
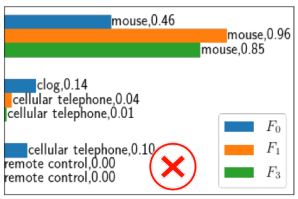
Image			
Ground Truth Label	sports car	minibus	celluar telephone
$F_2F_5F_7$ Acc=78.80% **FQ-GD=0.263** (EfficientNet-B0, ResNet152, SqueezeNet)			
$F_0F_1F_3$ Acc=78.83% FQ-GD=0.461 (AlexNet, DenseNet, ResNeXt50)			

Table 4: Examples on ImageNet and Top-3 Classification Confidence

on the initial centroids as we did in binary K-means. In our first prototype, we use the K-means to perform binary clustering and to determine the proper diversity threshold, as described in Step (3) in Section 3 of the main paper.

1.5. Ensemble Diversity and Accuracy

Ensemble diversity metrics are by design to capture the failure independence among member models of an ensemble team. Even though we improved the correlation of ensemble diversity and ensemble accuracy in this study, the proposed FQ diversity metrics can still capture the failure independence and complementary capacity among member models which cannot be directly measured by ensemble accuracy. Table 4 shows two example ensemble teams $F_2F_5F_7$ and $F_0F_1F_3$ with three example images from ImageNet and the top-3 classification confidence of each member model. The two teams, $F_2F_5F_7$ and $F_0F_1F_3$, have similar ensemble accuracy, that is 78.80% for $F_2F_5F_7$ and 78.83% for $F_0F_1F_3$, while their ensemble diversity measurements are significantly different in terms of the unifying FQ-GD scores, i.e., 0.263 for $F_2F_5F_7$ and 0.461 for $F_0F_1F_3$. Based on the FQ-GD scores, $F_2F_5F_7$ is preferred as the member models in this team are more diverse in terms of failure independence in comparison to the team of $F_0F_1F_3$. The visualization for these two ensemble teams in Table 4 illustrates the impact of such FQ-GD diversity on the ensemble prediction results by three example images. For a high diversity and hence high failure independence team $F_2F_5F_7$, when one member model in $F_2F_5F_7$ makes the wrong predictions, the other two member models can help in correcting such mistakes. On the contrary, with the low failure independence (low diversity, a high FQ-GD value), the majority of the member models of $F_0F_1F_3$ tend to make the same prediction errors, and fail to collectively improve the prediction quality. This case study further demonstrates that the design of our FQ diversity metrics are effective in capturing the degree of failure independence and complementary capacity among member models in terms of FQ-GD values.

2. Diversity by Uncorrelated Error

Neural network ensemble uses multiple (say $M > 1$) neural networks to form a committee (team) to collaborate and combine the predictions of individual member models to make the final prediction. A consensus method will be used to combine the individual predictions, such as majority voting, plurality voting, or soft voting (model averaging, the average of prediction vectors). [2, 8] In this study, we use the soft voting to combine individual member model predictions for each ensemble team, which in general performs better than majority voting or plurality voting. [9]

A neural network classifier is typically trained to minimize a cross-entropy loss and output a probability vector to approximate a posteriori probability densities for the corresponding class. For a given input x, the ith element in the output probability vector of model F_k can be modeled as: $f_i^k(x) = p(c_i|x) + \epsilon_i^k(x)$, where $p(c_i|x)$ is the posteri-

ori probability distribution of the ith class (c_i) for the input x, and $\epsilon_i^k(x)$ is the error associated with this output. For making the Bayes optimum decision, x will be predicted as class c_i if $p(c_i|x) > p(c_j|x), \forall j \neq i$. Therefore, the Bayes optimum boundary locates at all points x^* such that $p(c_i|x^*) = p(c_j|x^*)$ where $p(c_j|x^*) = max_{l \neq i} p(c_l|x)$. Given the neural network model will output $f_i^k(x)$ instead of $p(c_i|x)$, the decision boundary of the model, $\bar{x}$, may vary from the optimum boundary by an offset $o = \bar{x} - x^*$. [7] shows that the added error beyond Bayes error is $E_{add} = \frac{d\sigma_o^2}{2}$ where d is the difference between the derivatives of the two posteriors and σ_o^2 is the variance of the boundary offset o, $\sigma_o^2 = \frac{2\sigma_{\epsilon_i^k}^2}{d^2}$. Combining the predictions of S models with model averaging (avg), the ith element in the combined probability vector gives an approximation to $p(c_i|x)$ as $f_i^{avg}(x) = \frac{1}{S}\sum_{k=1}^{S} f_i^k(x) = p(c_i|x) + \bar{\epsilon}_i(x)$, where $\bar{\epsilon}_i(x) = \frac{1}{S}\epsilon_i^k(x)$. We can calculate the variance of $\bar{\epsilon}_i$ with Formula (1) as follows:

$$\begin{aligned}\sigma_{\bar{\epsilon}_i}^2 &= \frac{1}{S^2}\sum_{k=1}^{S}\sum_{l=1}^{S} cov(\epsilon_i^k(x), \epsilon_i^l(x)) \\ &= \frac{1}{S^2}\sum_{k=1}^{S}\sigma_{\epsilon_i^k}^2 + \frac{1}{S^2}\sum_{k=1}^{S}\sum_{l\neq k} cov(\epsilon_i^k(x), \epsilon_i^l(x))\end{aligned} \tag{1}$$

where $cov()$ represents the covariance. With $cov(a,b) = corr(a,b)\sigma_a\sigma_b$, we can replace the covariance with correlation $corr()$ and derive

$$\sigma_{\bar{\epsilon}_i}^2 = \frac{1}{S^2}\sum_{k=1}^{S}\sigma_{\epsilon_i^k}^2 + \frac{1}{S^2}\sum_{k=1}^{S}\sum_{l\neq k} corr(\epsilon_i^k(x), \epsilon_i^l(x))\sigma_{\epsilon_i^k}\sigma_{\epsilon_i^l}$$

Let δ_i denote the average correlation factor among these models, we have

$$\delta_i = \frac{1}{S(S-1)}\sum_{k=1}^{S}\sum_{l\neq k} corr(\epsilon_i^k(x), \epsilon_i^l(x))$$

Assuming the common variance $\sigma_{\bar{\epsilon}_i}^2 = \sigma_{\epsilon_i^k}^2$ holds for every model F_k, with δ_i we have

$$\sigma_{\bar{\epsilon}_i}^2 = \frac{1}{S}\sigma_{\epsilon_i}^2 + \frac{S-1}{S}\delta_i\sigma_{\epsilon_i}^2$$

With the variance of the ensemble decision boundary offset $\sigma_{o^{avg}}^2 = \frac{\sigma_{\epsilon_i}^2 + \sigma_{\epsilon_j}^2}{d^2}$ given in [7], we have

$$\sigma_{o^{avg}}^2 = \frac{1}{d^2 S}(\sigma_{\epsilon_i}^2 + (S-1)\delta_i\sigma_{\epsilon_i}^2 + \sigma_{\epsilon_j}^2 + (S-1)\delta_j\sigma_{\epsilon_j}^2)$$

Assume that the error between classes are i.i.d., that is $\sigma_{\epsilon_i}^2 = \sigma_{\epsilon_j}^2$. With $\sigma_{\epsilon_i}^2 = \sigma_{\epsilon_i^k}^2$ (the previous assumption) and $\sigma_o^2 = \frac{2\sigma_{\epsilon_i^k}^2}{d^2}$ given in [7], we have the following Formula (2).

$$\begin{aligned}\sigma_{o^{avg}}^2 &= \frac{1}{d^2 S}(2\sigma_{\epsilon_i}^2 + (S-1)\sigma_{\epsilon_i}^2(\delta_i + \delta_j)) \\ &= \frac{2\sigma_{\epsilon_i}^2}{d^2 S}(1 + (S-1)\frac{(\delta_i + \delta_j)}{2}) \\ &= \frac{2\sigma_{\epsilon_i^k}^2}{d^2 S}(1 + (S-1)\frac{(\delta_i + \delta_j)}{2}) \\ &= \frac{\sigma_o^2}{S}(1 + (S-1)\frac{\delta_i + \delta_j}{2})\end{aligned} \tag{2}$$

To extend the above formula to include all classes, given $\delta = \sum_{i=1}^{C} P_i\delta_i$, where P_i is the prior probability of class c_i and C is the total number of classes. Assuming the prior probability P_i of class c_i is uniformly distributed, we have

$$\sigma_{o^{avg}}^2 = \frac{\sigma_o^2}{S}(1 + (S-1)\delta) = \sigma_o^2(\frac{1 + (S-1)\delta)}{S})$$

So we can derive the added error for the ensemble prediction E_{add}^{avg} as Formula (3) shows:

$$\begin{aligned}E_{add}^{avg} &= \frac{d\sigma_{o^{avg}}^2}{2} \\ &= \frac{d\sigma_o^2}{2}(\frac{1 + (S-1)\delta)}{S}) \\ &= E_{add}(\frac{1 + (S-1)\delta)}{S})\end{aligned} \tag{3}$$

Hence, the ideal scenario corresponds to S diverse member models of an ensemble team with size S, where they can output predictions with uncorrelated errors (failure independence), i.e., $\delta \leq 0$. In this case, the overall prediction error can be dramatically reduced by at least $S\times$ with a simple model averaging method. In the meantime, the worst scenario corresponds to highly correlated errors of individual member models with $\delta = 1$, such as S perfect model duplicates, the error of this ensemble will remain the same as the initial error. In general, the correlation δ lies between 0 and 1, and therefore, it is always beneficial to use ensemble to reduce the prediction errors.

3. Ensemble Diversity Metrics

In this study, we have covered six representative diversity metrics. In the literature, different studies will use one of these diversity metrics to select models and analyze the prediction results. However, there are few studies to provide guidelines for choosing them or to compare and evaluate these diversity metrics in terms of ensemble selection quality with respect to boosting ensemble accuracy. Our paper is one of the first attempts to investigate the impact of ensemble diversity metrics on boosting the overall ensemble accuracy.

In general, diversity metrics can be classified into two major categories based on how the fault independence and uncorrelated errors are computed using a set of negative samples. They are pairwise metrics and non-pairwise metrics. We below describe six representative diversity metrics considered in our study: Cohen's Kappa, Q Statistics and Binary Disagreement for pairwise, and Fleiss' Kappa, Kohavi-Wolpert Variance and Generalized Diversity for non-pairwise.

Consider a base model pool of M base models, all trained on the same dataset for the same learning task. Let $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{N-1}\}$ be the set of randomly selected N labeled negative samples from the training dataset. For a base model F_i and a negative sample set $\mathbf{X}$, F_i will output a vector of binary values on $\mathbf{X}$, denoted as $\boldsymbol{\omega_i} = [\omega_{i,0}, \omega_{i,1}, ..., \omega_{i,N-1}]^T$, and $\omega_{i,k} = 1$ if F_i predicts $\mathbf{x}_k$ correctly, otherwise, $\omega_{i,k} = 0$.

Pairwise Diversity Metrics Pairwise diversity metrics are calculated based on a pair of classifiers. Table 5 lists four different types of prediction results between a pair of classifiers F_i and F_j, such as both F_i and F_j make correct or wrong predictions and either F_i or F_j makes correct predictions. Hence, we can count the number of samples in the four different types, that is N^{ab}, which represents the total number of samples $\mathbf{x}_k \in \mathbf{X}$, such that $\omega_{i,k} = a$ and $\omega_{j,k} = b$.

	F_j correct (1)	F_j wrong (0)
F_i correct (1)	N^{11}	N^{10}
F_i wrong (0)	N^{01}	N^{00}
$N = N^{00} + N^{01} + N^{10} + N^{11}$		

Table 5: The relationship between a pair of classifiers

i. Cohen's Kappa (CK): Cohen's Kappa measures the diversity between a pair of classifiers F_i and F_j from the perspective of agreement [4, 3]. A lower Cohen's kappa value indicates lower agreement and higher diversity. Its definition (κ_{ij}) between a pair of classifiers F_i and F_j is shown in Formula (4). The value of Cohen's kappa ranges from -1 to 1, where 0 represents the amount of agreement by random chance. [4]

$$\kappa_{ij} = \frac{2(N^{11}N^{00} - N^{01}N^{10})}{(N^{11} + N^{10})(N^{01} + N^{00}) + (N^{11} + N^{01})(N^{10} + N^{00})} \tag{4}$$

ii. Q Statistics (QS): The Q statistics [10] is defined as QS_{ij} in Formula (5) for a pair of models F_i and F_j. The value QS_{ij} varies between -1 and 1. When the models F_i and F_j are statistically independent, the expected QS_{ij} value is 0. If both models tend to recognize the same input sample similarly, QS_{ij} will have a positive value. For two diverse models, recognizing the same input sample differently, it will render a small or negative QS_{ij} value.

$$QS_{ij} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{5}$$

iii. Binary Disagreement (BD): The binary disagreement [6, 3] is defined as the ratio of (i) the number of samples on which one model is correct while the other model is wrong to (ii) the total number of samples predicted by the two models F_i, F_j in Formula (6).

$$BD_{ij} = \frac{N^{01} + N^{10}}{N^{11} + N^{10} + N^{01} + N^{00}} \tag{6}$$

For an ensemble team of S member models, we calculate its diversity value as the averaged metric value over all pairs of classifiers in Formula (7), where Q represents a pair-wise diversity metric as recommended by [3].

$$Q = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} Q_{ij} \tag{7}$$

Non-pairwise Diversity Metrics Non-pairwise diversity metrics are widely used for a team of 2 or more models. We focus on three representative non-pairwise diversity metrics to compare with pairwise diversity metrics. For an ensemble team of S classifiers, $l(\mathbf{x}_k)$ denotes the number of classifiers that correctly recognize $\mathbf{x}_k$, i.e., $l(\mathbf{x}_k) = \sum_{i=1}^{S} \omega_{ik}$.

iv. Fleiss' Kappa (FK): Similar to Cohen's Kappa, Fleiss' Kappa [1] also measures the diversity from the perspective of agreement. The difference is that it can be directly calculated for a team of 2 or more models as Formula (8) shows, where $\bar{p}$ is the average classification accuracy for the ensemble team and κ is not simply obtained by averaging the Cohen's kappa (κ_{ij}).

$$\begin{aligned} \bar{p} &= \frac{1}{NS} \sum_{k=1}^{N} \sum_{i=1}^{S} \omega_{i,k} \\ \kappa &= 1 - \frac{\frac{1}{S} \sum_{k=1}^{N} l(\mathbf{x}_k)(S - l(\mathbf{x}_k))}{N(S-1)\bar{p}(1-\bar{p})} \end{aligned} \tag{8}$$

v. Kohavi-Wolpert Variance (KW): Kohavi-Wolpert Variance [3] measures the variability of the predicted class labels for the sample $\mathbf{x}_k$ within the team of models $F_1, F_2, ..., F_S$ as Formula (9) shows. A higher value of KW variance implies higher ensemble diversity of the team.

$$KW = \frac{1}{NS^2} \sum_{k=1}^{N} l(\mathbf{x}_k)(S - l(\mathbf{x}_k)) \tag{9}$$

vi. Generalized Diversity (GD): The generalized diversity was proposed by [5] as Formula (10) shows. Y is a random variable, representing the proportion of classifiers (out of S) that fail to recognize a random sample $\mathbf{x}_k$. p_i denotes the probability of $Y = \frac{i}{S}$, that is the probability of i (out

of S) classifiers recognizing a randomly chosen sample $\mathbf{x}_k$ incorrectly. $p(1)$ represents the expected probability of one randomly picked model failing while $p(2)$ is the expected probability of both two randomly picked models failing. GD varies between 0 and 1. The maximum diversity, 1, can be reached when the failure of one model is accompanied by the correct recognition by the other model for two randomly picked models, which corresponds to $p(2) = 0$. When both two randomly picked models fail, we have $p(1) = p(2)$, corresponding to the minimum diversity, 0.

$$
\begin{aligned}
p(1) &= \sum_{i=1}^{S} \frac{i}{S} p_i \\
p(2) &= \sum_{i=1}^{S} \frac{i(i-1)}{S(S-1)} p_i \\
GD &= 1 - \frac{p(2)}{p(1)}
\end{aligned}
\tag{10}
$$

Table 6 lists the six Q-diversity metrics with three pairwise and three non-pairwise representatives. The arrow column ($\uparrow$ | $\downarrow$) specifies the relationship between the Q-value and the ensemble diversity. The $\downarrow$ indicates that the low Q-value refers to high ensemble diversity and $\uparrow$ indicates that the high Q-value refers to high ensemble diversity. To present a consistent view of all six Q-diversity metrics such that the low Q-value corresponds to high ensemble diversity, we apply $(1 - Q$-value) when calculating the diversity scores using BD, KW and GD.

Type	Name	Notation	$\uparrow$ | $\downarrow$
Pairwise	Cohen's Kappa	CK	$\downarrow$
	Q Statistics	QS	$\downarrow$
	Binary Disagreement	BD	$\uparrow$
Non-pawise	Fleiss' Kappa	FK	$\downarrow$
	Kohavi-Wolpert variance	KW	$\uparrow$
	Generalized Diversity	GD	$\uparrow$

Table 6: A summary of 6 Q-diversity metrics

4. Algorithm for Q Ensemble Selection

In Section 2 of the main paper, we have provided the analysis of the potential problems for the Q-diversity ensemble selection, which motivated the development of FQ-diversity ensemble selection. In this supplementary material, we include the pseudo code of Algorithm 1 for Q-diversity in this section and Algorithm 2 for FQ-diversity in the next section.

Algorithm 1 provides a sketch of the pseudo code, describing the Q-diversity ensemble team selection process. We denote a diversity threshold calculation function by $f_{threshold}$, such as the mean function. For Q-diversity ensemble selection, we compute the mean value of all diversity values computed for all candidate ensemble teams

Algorithm 1 Q Ensemble Team Selection

1: **procedure** QENSTEAMS($NegSampSet, Q, f_{threshold}$)
2: **Input**: $NegSampSet$: negative samples; Q the diversity metric; $f_{threshold}$: the diversity threshold calculation function.
3: **Output**: $GEnsSet$: the set of good ensemble teams.
4: Obtain $EnsSet$ ▷ all possible ensemble teams in $EnsSet$
5: Initialize $GEnsSet = \{\}$, $D = \{\}$
6: **for** $i = 1$ to $|EnsSet|$ **do**
7: ▷ calculate the diversity metric Q for $T_i \in EnsSet$
8: $q_i = DiversityMetric(Q, T_i, NegSampSet)$
9: D.append(q_i) ▷ Store q_i in the diversity measures D
10: **end for**
11: $\theta(Q) = f_{threshold}(D)$ ▷ Calculate the diversity threshold
12: **for** $i = 1$ to $|EnsSet|$ **do**
13: **if** $q_i < \theta(Q)$ **then**
14: $GEnsSet$.add(T_i) ▷ add qualified T_i into $GEnsSet$
15: **end if**
16: **end for**
17: **return** $GEnsSet$
18: **end procedure**

in $EnsSet$ as the Q-diversity threshold. Based on the diversity threshold $\theta(Q)$ (Line 11), we will select teams with the diversity measure $q < \theta(Q)$ and place them into the set of good ensemble teams $GEnsSet$ (Line 12∼16). With Q-diversity metrics, we have high probability to boost the overall ensemble accuracy of the selected teams in $GEnsSet$.

5. Algorithm for FQ Ensemble Selection

Algorithm 2 gives a sketch of the process of using FQ-diversity to select good ensemble teams from a single partition of the ensemble teams with a fixed size and focal model, i.e., $EnsSet(F_{focal}, S)$.

First, consider all the possible ensemble teams of size S including the focal model F_{focal}, i.e., $EnsSet(F_{focal}, S)$. We can calculate the size of $EnsSet(F_{focal}, S)$ via counting all the combinations of $(S-1)$ models from the remaining $(F-1)$ base models in the base model pool, that is $|EnsSet(F_{focal}, S)| = \binom{M-1}{S-1} = \frac{(M-1)!}{(S-1)!(M-S)!}$. For example, when $M = 10$, and $S = 5$, we have $|EnsSet(F_{focal}, S)| = 126$.

Second, with a certain number, such as 100, of negative samples randomly selected from the focal model, F_{focal}, denoted as $NegSampSet(F_{focal})$, we calculate the set of diversity score and ensemble accuracy pairs, (q_i, acc_i), $DA(Q) = \{(q_i, acc_i) | T_i \in EnsSet(F_{focal}, S)\}$, each corresponding to the FQ-diversity measure of the team T_i and its ensemble accuracy as Line 5∼10 in Algorithm 2 shows.

Third, for the given diversity metric Q, we propose to use the K-means clustering algorithm with $K = 2$ and two initial centroids, $initCentroid_1, initCentroid_2$, to identify the good diversity threshold from $DA(Q)$. $initCentroid_1$ is defined as the highest diversity centroid, denoted as (q^1_{min}, acc^1_{max}) such that for any $(q_i, acc_i) \in DA(Q)$,

Algorithm 2 FQ Ensemble Team Selection

1: **procedure** FQEnsTeams($F_{focal}, NegSampSet(F_{focal}), S, Q$)
2: **Input**: F_{focal} the focal model; $NegSampSet(F_{focal})$: negative samples from F_{focal}; S the team size; Q the diversity metric.
3: **Output**: $GEnsSet(F_{focal}, S, Q)$: the set of selected diverse ensemble teams with team size S and focal model F_{focal}.
4: Obtain $EnsSet(F_{focal}, S)$
5: Initialize $DA(Q) = \{\}$
6: **for** $i = 1$ to $|EnsSet(F_{focal}, S)|$ **do**
7: $q_i = DiversityMetric(Q, T_i, NegSampSet(F_{focal}))$
8: $acc_i = Accuracy(T_i)$
9: $DA(Q)$.append((q_i, acc_i))
10: **end for**
11: Initialize $GEnsSet(F_{focal}, S, Q) = \{\}$
12: ▷ obtain the initial centroids for clustering
13: $initCentroid_1 = (q^1_{min}, acc^1_{max})$
14: $initCentroid_2 = (q^2_{max}, acc^2_{min})$
15: ▷ obtain the 2 clusters via running K-Means on $DA(Q)$.
16: $Cluster_1, (q^1, acc^1), Cluster_2, (q^2, acc^2) = KMeans(2, DA, initCentroid_1, initCentroid_2)$
17: ▷ get the threshold s.t. $acc^1 \geq acc^2$
18: $\theta_{FQ}(F_{focal}, S, Q) = min(min_{div}(Cluster_2), mean_{div}(DA(Q)))$
19: ▷ add qualified T_i into $GEnsSet(F_{focal}, S, Q)$
20: **for** $i = 1$ to $|EnsSet(F_{focal}, S)|$ **do**
21: **if** $q_i < \theta_{FQ}(F_{focal}, S, Q)$ **then**
22: $GEnsSet(F_{focal}, S, Q)$.add(T_i)
23: **end if**
24: **end for**
25: **return** $GEnsSet(F_{focal}, S, Q)$
26: **end procedure**

we have $q^1_{min} \leq q_i$ and $acc^1_{max} \geq acc_i$ and $\exists j, k \in \{1, 2, ..., |EnsSet(F_{focal}, S)|\}, q^1_{min} = q_j, acc^1_{max} = acc_k$. Similarly, we define $initCentroid_2$ as the lowest diversity centroid, denoted as (q^2_{max}, acc^2_{min}) such that for any $(q_i, acc_i) \in DA(Q)$, we have $q^2_{max} \geq q_i$ and $acc^2_{min} \leq acc_i$ and $\exists j, k \in \{1, 2, ..., |EnsSet(F_{focal}, S)|\}, q^2_{max} = q_j, acc^2_{min} = acc_k$. The K-means clustering will partition $DA(Q)$ into two clusters, $Cluster_1$ with the centroid (q^1, acc^1) and $Cluster_2$ with the centroid (q^2, acc^2) such that the accuracy on the centroid of $Cluster_1$ is higher than $Cluster_2$, that is $acc^1 \geq acc^2$. Leveraging these two clusters, we will focus on those pairs (q_i, acc_i) in $Cluster_1$ that have high acc_i and low q_i (high FQ-diversity). Let $min_{div}(Cluster_2)$ be the lowest FQ-value in $Cluster_2$ and $mean_{div}$(DA(Q)) be the mean value of all FQ-diversity values in DA(Q). We compute the FQ-diversity threshold $\theta_{FQ}(F_{focal}, S, Q)$ as follows: $\theta_{FQ}(F_{focal}, S, Q) = min(min_{div}(Cluster_2), mean_{div}(DA(Q)))$. The team $T_i \in EnsSet(F_{focal}, S)$ will be selected into $GEnsSet(F_{focal}, S, Q)$ if $q_i < \theta_{FQ}(F_{focal}, S, Q)$ (Line 18∼24).

Given a diversity metric Q, a set of candidate ensemble teams of a fixed team size S, and a focal model F_{focal}, we select the set of ensemble teams $GEnsSet(F_{focal}, S, Q)$ with FQ-diversity. For an ensemble team in $GEnsSet(F_{focal}, S, Q)$, it has S different FQ-diversity scores, one for each of its S focal models as Figure 2 shows. For example, the team `012` is included in $EnsSet(F_0, S = 3)$, $EnsSet(F_1, S = 3)$, and $EnsSet(F_2, S = 3)$, then it has three FQ-CK scores ($S = 3$) for $focal = 0, 1, 2$ respectively. These S FQ scores will be combined to produce a final unifying FQ diversity score for each ensemble team.

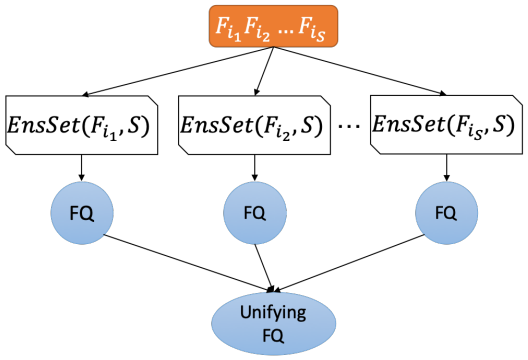


Figure 2: Unifying FQ scores

We follow the steps below to obtain the final unifying FQ-score for each ensemble team: (1) We first scale the FQ-scores obtained in Algorithm 2 for all teams in one partition, i.e., $EnsSet(F_{focal}, S)$, into $[0, 1]$. That is to scale $D = \{q_i | (q_i, acc_i) \in DA(Q)\}$ into $[0, 1]$. The set of scaled FQ-scores is $\overline{D} = \{\overline{q}_i | \overline{q}_i = \frac{q_i - min(D)}{max(D) - min(D)}, q_i \in D\}$ corresponding to the original set of FQ-scores, D, where $min(D)$ and $max(D)$ represent the minimum and maximum FQ-scores in D, and we have $\overline{q}_i \in [0, 1]$. (2) Then for an ensemble team T of size S, we can obtain S scaled FQ-scores, corresponding to the S focal models. Following the previous example, the team `012` also has three scaled FQ-scores, denoted as $\overline{q_{a_0}}, \overline{q_{a_1}}, \overline{q_{a_2}}$ for $focal = 0, 1, 2$ respectively, where a_i is the index of the team `012` in $EnsSet(F_i, S = 3)$. (3) Next, we perform a weighted average of the S scaled FQ-scores to obtain the final unifying FQ-score. The weights are calculated using the rank of focal model accuracy. For example, for the team `012` on the CIFAR-10 dataset, the first focal model F_0 has the highest model accuracy and rank 3 within the team in ascending order as Table 1 in the main paper shows. Its corresponding weight is $3 - 1 = 2$. The reason to deduct 1 from all the ranks for calculating the weights is to avoid the worst FQ-score when the lowest accuracy model serves as the focal model, i.e., the weight for the focal model F_1 with the lowest model accuracy and rank 1 is $1 - 1 = 0$. Therefore, the final unifying FQ-score for the team `012` is $\frac{2 \times \overline{q_{a_0}} + 0 \times \overline{q_{a_1}} + 1 \times \overline{q_{a_2}}}{2+0+1}$. The final unifying FQ-scores enable a fair diversity comparison among ensemble teams of different team sizes and focal models.

To combine the ensemble teams selected for a fixed size team of S and focal model F_{focal}. We follow a two-level aggregation. First, we combine the ensemble teams of the

same size S. Here, a set of (q_i, acc_i) pairs of the final unifying FQ-scores (q) and accuracy (acc) for all teams in $\bigcup_{focal=0}^{M-1} GEnsSet(F_{focal}, S, Q)$ with the same team size S will be formed. We then follow the same steps as Line 11∼24 in Algorithm 2 with the threshold more loosely set as $\theta_{FQ}(S, Q) = q^2$ (corresponding to the centroid of $Cluster_2$) in Line 18 to further prune out bad ensemble teams. Hence, we can obtain a better set of ensemble teams of team size S, denoted as $GEnsSet(S, Q)$. For all teams with different team sizes S, we simply perform a union operation across $GEnsSet(S, Q)$ to obtain all the good ensemble teams, that is $\bigcup_{S=2}^{M-1} GEnsSet(S, Q)$. The team with all base models ($S = M$) is a special case, which will be considered separately. Through the above steps, we can obtain a set of all good ensemble teams selected by an FQ-diversity metric as well as the final unifying FQ diversity scores for each ensemble team.

References

[1] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013. 5

[2] Cheng Ju, Aurélien Bibaut, and Mark Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45, 04 2017. 3

[3] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207, May 2003. 5

[4] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276—282, 2012. 5

[5] D. Partridge and W. Krzanowski. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology*, 39(10):707–717, 1997. 5

[6] David B. Skalak. The sources of increased accuracy for two proposed boosting algorithms. In *In Proc. American Association for Arti Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*, pages 120–125, 1996. 5

[7] KAGAN TUMER and JOYDEEP GHOSH. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3-4):385–404, 1996. 4

[8] M. van Erp, L. Vuurpijl, and L. Schomaker. An overview and comparison of voting methods for pattern recognition. In *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 195–200, 2002. 3

[9] Y. Wu, L. Liu, Z. Xie, J. Bae, K. H. Chow, and W. Wei. Promoting high diversity ensemble learning with ensemblebench. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, pages 208–217, 2020. 3

[10] G. Udny Yule. On the association of attributes in statistics: With illustrations from the material of the childhood society, c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194:257–319, 1900. 5