

# Hierarchical Pruning of Deep Ensembles with Focal Diversity

YANZHAO WU\*, Florida International University, USA

KA-HO CHOW, WENQI WEI, and LING LIU, Georgia Institute of Technology, USA

Deep neural network ensembles combine the wisdom of multiple deep neural networks to improve the generalizability and robustness over individual networks. It has gained increasing popularity to study and apply deep ensemble techniques in the deep learning community. Some mission-critical applications utilize a large number of deep neural networks to form deep ensembles to achieve desired accuracy and resilience, which introduces high time and space costs for ensemble execution. However, it still remains a critical challenge whether a small subset of the entire deep ensemble can achieve the same or better generalizability and how to effectively identify these small deep ensembles for improving the space and time efficiency of ensemble execution. This paper presents a novel deep ensemble pruning approach, which can efficiently identify smaller deep ensembles and provide higher ensemble accuracy than the entire deep ensemble of a large number of member networks. Our hierarchical ensemble pruning approach (HQ) leverages three novel ensemble pruning techniques. First, we show that the focal ensemble diversity metrics can accurately capture the complementary capacity of the member networks of an ensemble team, which can guide ensemble pruning. Second, we design a focal ensemble diversity based hierarchical pruning approach, which will iteratively find high quality deep ensembles with low cost and high accuracy. Third, we develop a focal diversity consensus method to integrate multiple focal diversity metrics to refine ensemble pruning results, where smaller deep ensembles can be effectively identified to offer high accuracy, high robustness and high ensemble execution efficiency. Evaluated using popular benchmark datasets, we demonstrate that the proposed hierarchical ensemble pruning approach can effectively identify high quality deep ensembles with better classification generalizability while being more time and space efficient in ensemble decision making. We have released the source codes on GitHub at <https://github.com/git-disl/HQ-Ensemble>.

CCS Concepts: • **Computing methodologies** → **Ensemble methods**.

Additional Key Words and Phrases: Ensemble Pruning, Ensemble Learning, Ensemble Diversity, Deep Learning

## ACM Reference Format:

Yanzhao Wu, Ka-Ho Chow, Wenqi Wei, and Ling Liu. 2023. Hierarchical Pruning of Deep Ensembles with Focal Diversity. *ACM Trans. Intell. Syst. Technol.* 1, 1, Article 1 (November 2023), 24 pages. <https://doi.org/10.1145/3633286>

## 1 INTRODUCTION

It has become an attractive learning technique to leverage deep neural network (DNN) ensembles to improve the overall generalizability and robustness of many deep learning systems. Some mission-critical applications often require a large number of deep neural networks to achieve the target accuracy and robustness, which entails high space and time costs for ensemble execution. Recent

\*Also with Georgia Institute of Technology.

Authors' addresses: Yanzhao Wu, yawu@fiu.edu, Florida International University, 11200 SW 8TH ST, Miami, Florida, USA, 33199; Ka-Ho Chow, khchow@gatech.edu; Wenqi Wei, wenqiwei@gatech.edu; Ling Liu, lingliu@cc.gatech.edu, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, Georgia, USA, 30332.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2157-6904/2023/11-ART1 \$15.00  
<https://doi.org/10.1145/3633286>

studies have revealed that deep neural network ensembles with highly diverse member networks tend to have high failure independence, which is critical for enhancing overall ensemble predictive performance, including ensemble accuracy and robustness under adverse situations [20, 34, 36, 38, 39]. However, the member networks in a large deep ensemble team may not have high ensemble diversity and failure independence to complement each other, which will result in sub-optimal ensemble prediction performance and high ensemble execution cost in practice [10, 20, 35, 37–39]. For a deep ensemble team of a large size, such as 10, it is often not only possible to identify substantially smaller deep ensembles (e.g., 3~5 member networks) with the same or improved ensemble accuracy but also beneficial to reduce the ensemble execution cost [18, 24, 37–39]. This motivates us to propose an efficient hierarchical ensemble pruning approach, coined as HQ. By leveraging our focal diversity metrics, the HQ pruning method can effectively identify high quality deep ensembles with small sizes and high ensemble diversity, which not only improves ensemble accuracy over the large entire ensembles but also significantly reduces the space and time cost of ensemble execution.

### 1.1 Related Work and Problem Statement

Most of the existing ensemble learning studies can be summarized into three broad categories. The first category builds ensemble teams by training a set of member models, represented by bagging [4], boosting [6] and random forests [5]. In practice, it has led to large ensemble teams of tens to hundreds of member models, such as the commonly-used random forests. The second category is to leverage diversity measurements to compare and select high quality ensemble teams whose member models can complement each other to improve the ensemble predictive performance without requiring model training. Ensemble diversity can be evaluated using pairwise or non-pairwise diversity measures, represented by Cohen’s Kappa (CK) [25] and Binary Disagreement (BD) [30] for pairwise metrics and Kohavi-Wolpert Variance (KW) [15, 17] and Generalized Diversity (GD) [27] for non-pairwise metrics. Early studies [17, 31] have reported that it is challenging to use these diversity metrics for evaluating the performance quality of ensemble teams. Some recent studies [10, 20, 38, 39] discussed inherent problems of using these diversity metrics to measure ensemble diversity in terms of failure independence and provided guidelines for improving the ensemble diversity measurement. This paper contributes to this second category by leveraging focal ensemble diversity metrics for effective pruning of deep ensembles. The third category covers the ensemble consensus methods for aggregating member model predictions and producing the ensemble prediction, such as soft voting (model averaging), majority voting and plurality voting [13] or learn to combine algorithms [21, 38].

Table 1. Summary of Three Categories of Related Studies

Category	Description	Representative methods
1. Ensemble training	Training a set of member models to build ensemble teams	Bagging [4], boosting [6], random forests [5], etc.
2. Ensemble diversity powered ensemble pruning	Leveraging ensemble diversity measurements to compare and select high-quality ensembles	CK pruning [25], BD pruning [30], GD pruning [27], etc.
3. Ensemble consensus methods	Aggregating member model predictions to produce the ensemble prediction	Soft voting, majority voting, plurality voting, etc.

We summarize the three categories of related studies in Table 1. These three categories are complementary. To develop an ensemble team for improving prediction performance, we can employ one of the two ways to obtain an ensemble of member models: (1) training multiple models together using an ensemble learning algorithm, such as bagging, boosting, or random forest; and (2) training multiple models independently in parallel and employ a voting method to produce

Table 2. Comparison of Hierarchical Pruning and Existing Studies

Method	Model Types	Diversity Measurement Principles		
		Diversity Comparison	Samples for Measurements	Diversity Calculation
Early studies on diversity based ensemble pruning before 2015 [2, 7, 18, 24, 32]	Ensembles of traditional ML models	Compare all ensembles of different sizes	Random samples from the validation set	Directly calculated on random samples
Focal diversity powered hierarchical ensemble pruning	Ensembles of (1) DNNs and (2) Traditional ML models	Compare the ensembles of the same size $S$	Random negative samples from a focal model $F_f$	Obtain focal negative correlations for each focal (member) model $F_f$ in an ensemble team of size $S$ and then perform an average of $S$ focal negative correlation scores to calculate the focal diversity score

ensemble consensus based predictions. Most of the early studies before 2015 in the first two categories [2, 7, 18, 24, 32] focused on ensemble pruning for traditional machine learning models. Until recently, we have observed several research endeavors on deep neural network ensembles, most of which centered on training multiple networks jointly, such as diversity based weighted kernels [3, 8, 40] and leveraging deep ensembles to strengthen the robustness and resilience of a single deep neural network under adverse situations [8, 9, 20, 34–36, 39]. However, to the best of our knowledge, very few studies have brought forward solutions to efficient deep ensemble pruning to improve prediction performance and reduce ensemble execution cost.

We compare our focal diversity powered hierarchical ensemble pruning approach and early studies using ensemble diversity for ensemble pruning in Table 2. Our focal diversity promotes fair comparison of ensemble diversity among the ensembles of the same size  $S$  and leverages the focal model concept, focal negative correlation, and averaging of multiple focal negative correlation scores to obtain accurate ensemble diversity measurements. We will provide a detailed description of focal diversity powered hierarchical pruning in Section 3 and demonstrate that our hierarchical pruning approach can be effectively applied to both deep neural network ensembles and ensembles of traditional machine learning models in Section 5.

## 1.2 Scope and Contribution

This paper presents a holistic approach to efficient deep ensemble pruning. Given a large deep ensemble of  $M$  member networks, we propose a hierarchical ensemble pruning framework, denoted as HQ, to efficiently identify high quality ensembles with lower cost and higher accuracy than the entire ensemble of  $M$  networks. Our HQ framework combines three novel ensemble pruning techniques. *First*, we leverage focal ensemble diversity metrics [39] to measure the failure independence among member networks of a deep ensemble. The higher focal diversity score indicates the higher level of failure independence among member networks of a deep ensemble, that is higher complementary capacity for improving ensemble predictions. Our focal diversity metrics can precisely capture such correlations among member networks of an ensemble team, which can be used to effectively guide ensemble pruning. *Second*, we present a novel hierarchical ensemble pruning method powered by focal diversity metrics. Our hierarchical pruning approach iteratively identifies subsets of member networks with low diversity, which tend to make similar prediction errors, and then prunes them out from the entire ensemble team. *Third*, we perform focal diversity consensus voting to combine multiple focal diversity metrics for ensemble pruning, which further refines the hierarchical ensemble pruning results by a single focal diversity metric. Comprehensive experiments are performed on four popular benchmark datasets, CIFAR-10 [16], ImageNet [29],

Cora [23] and MNIST [19]. The experimental results demonstrate that our focal diversity based hierarchical pruning approach is effective in identifying high quality deep ensembles with significantly smaller sizes and better ensemble accuracy than the entire deep ensemble team.

## 2 ENSEMBLE PRUNING WITH FOCAL DIVERSITY

Given an entire deep ensemble with a large size  $M$ , it consists of  $M$  individual member networks ( $F_i, i \in \{0, 1, \dots, M-1\}$ ) that are trained for a specific learning task and dataset. We denote the set of all possible sub-ensembles as  $EnsSet$ , which are composed of any subset of these  $M$  individual networks. For a specific team size  $S$ , let  $EnsSet(S)$  denote the set of all possible sub-ensembles of size  $S$  in  $EnsSet$ . The cardinality of  $EnsSet(S)$  is calculated based on the selection of  $S$  networks from all  $M$  base networks, that is  $|EnsSet(S)| = \binom{M}{S}$ . Therefore, the total number of candidate sub-ensembles for  $S = 2, \dots, M-1$  is  $|EnsSet| = \sum_{S=2}^{M-1} |EnsSet(S)| = \binom{M}{2} + \binom{M}{3} + \dots + \binom{M}{M-1} = 2^M - (2+M)$ , which grows exponentially with  $M$ . For example, when  $M = 3, 5, 10, 20$ , we have  $|EnsSet| = 3, 25, 1012, 1048554$  respectively. It may not be feasible to perform the exhaustive search of all possible ensembles in  $EnsSet$  with a large  $M$ . Hence, it is critical to develop efficient ensemble pruning methods for examining the sub-ensembles from the candidate set  $EnsSet$ , removing these sub-ensembles with low diversity and obtaining the set  $GEnsSet$  of high quality sub-ensembles with lower cost and better ensemble accuracy than the entire deep ensemble of  $M$  member networks.

Table 3. Example Deep Ensembles for CIFAR-10 and ImageNet




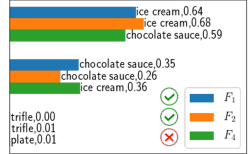
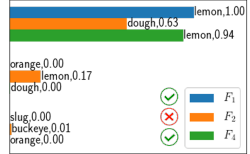
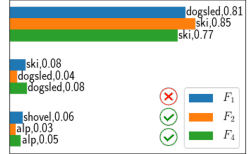
Dataset	CIFAR-10					ImageNet				
Ensemble Team	0123456789	0123	01238	123	1234	0123456789	12345	2345	1234	124
Ensemble Acc (%)	96.33	<b>97.15</b>	<b>96.87</b>	<b>96.81</b>	<b>96.63</b>	79.82	<b>80.77</b>	<b>80.70</b>	<b>80.29</b>	<b>79.84</b>
Team size	10	4	5	3	4	10	5	4	4	3
Acc Improv (%)	0	<b>0.82</b>	<b>0.54</b>	<b>0.48</b>	<b>0.30</b>	0	<b>0.95</b>	<b>0.88</b>	<b>0.47</b>	<b>0.02</b>
Cost Reduction (%)	0	<b>60</b>	<b>50</b>	<b>70</b>	<b>60</b>	0	<b>50</b>	<b>60</b>	<b>60</b>	<b>70</b>

Table 4. All Individual Member Models for Four Benchmark Datasets

Dataset	CIFAR-10		ImageNet		Cora		MNIST	
	10,000 testing samples		50,000 testing samples		1,000 testing samples		10,000 testing samples	
ID	Name	Acc (%)	Name	Acc (%)	Name	Acc (%)	Name	Acc (%)
0	<b>DenseNet190</b>	<b>96.68</b>	AlexNet	56.63	GCN	81.70	KNN	94.23
1	DenseNet100	95.46	DenseNet	77.15	GAT	82.80	Logistic Regression	91.89
2	ResNeXt	96.23	EfficientNet-B0	75.80	SGC	81.70	Linear SVM	92.48
3	WRN	96.21	ResNeXt50	77.40	ARMA	82.10	<b>RBF SVM</b>	<b>96.31</b>
4	VGG19	93.34	Inception3	77.25	APPNP	82.20	Random Forest	95.91
5	ResNet20	91.73	<b>ResNet152</b>	<b>78.25</b>	APPNP1	83.80	GBDT	92.89
6	ResNet32	92.63	ResNet18	69.64	APPNP2	88.70	Neural Network	96.18
7	ResNet44	93.10	SqueezeNet	58.00	<b>SplineCNN</b>	<b>88.90</b>		
8	ResNet56	93.39	VGG16	71.63	SplineCNN1	88.30		
9	ResNet110	93.68	VGG19-BN	74.22	SplineCNN2	88.50		
MIN	ResNet20	91.73	AlexNet	56.63	GCN/SGC	81.70	Logistic Regression	91.89
AVG		94.25		71.60		84.87		94.27
MAX	DenseNet190	96.68	ResNet152	78.25	SplineCNN	88.90	RBF SVM	96.31

We show 5 example deep ensemble teams in Table 3 for two datasets, CIFAR-10 and ImageNet. For each dataset, we list the entire deep ensemble (0123456789) of 10 member networks in addition to the 4 high quality deep ensembles that are identified by our HQ ensemble pruning approach. The 10 member networks in each entire ensemble for CIFAR-10 and ImageNet are given in Table 4. The

Table 5. Examples on ImageNet and Top-3 Classification Confidence

Image			
Ground Truth Label	ice cream	lemon	ski
Accuracy (%) $F_1 F_2 F_4$ : 79.84 (DenseNet: 77.15, EfficientNet-B0: 75.80, Inception3: 77.25)			
Ensemble Output	ice cream	lemon	ski

4 sub-ensembles recommended by our HQ have much smaller team sizes with only 3~5 individual member networks and achieve better ensemble accuracy than the given entire ensemble of 10 networks, significantly reducing the ensemble execution cost by 50%~70% for both CIFAR-10 and ImageNet. Table 5 further presents 3 image examples from ImageNet and the prediction results with Top-3 classification confidence from the member networks of the sub-ensemble team 124. This ensemble team achieves higher accuracy than each individual member network. For each image, one member model makes a prediction error. But the ensemble team can still give the correct predictions by repairing the wrong predictions by its member networks. This sub-ensemble team of 3 member networks also outperforms the entire ensemble of 10 models on ImageNet. For a given deep ensemble of size  $M = 10$ , we have a total of 1012 possible sub-ensembles to be considered in ensemble pruning. It is challenging to design and develop an efficient ensemble pruning approach to identify high quality sub-ensembles to improve both ensemble accuracy and time and space efficiency of ensemble execution.

**Problems with Baseline Ensemble Pruning.** In related work, we discussed several recent studies that utilize deep ensembles to strengthen the robustness of individual deep neural network against adversarial attacks [8, 9, 20, 34]. Most of these existing methods leverage Cohen’s Kappa (CK) [25] for measuring ensemble diversity since early studies [17, 24, 31, 32] in the literature have shown that both pairwise and non-pairwise diversity metrics share similar diversity evaluation results with regard to ensemble predictive performance, including the CK, BD, KW and GD metrics mentioned in related work. However, we show that these existing diversity metrics may not precisely capture the inherent failure independence among member networks of a deep ensemble team, which may not produce the optimal performance in guiding ensemble pruning.

We first study the baseline diversity metrics for ensemble pruning and analyze their inherent problems. For a given diversity metric, such as BD or GD, the baseline ensemble pruning method calculates the ensemble diversity scores for each candidate ensemble in *EnsSet* using a set of random samples drawn from the validation set as suggested by [17]. A practical method for pruning a given large ensemble of  $M$  member networks follows three steps: (1) we first compute the ensemble diversity score for each candidate sub-ensemble by using CK, BD, KW or GD; (2) we calculate the mean diversity score as the pruning threshold; and (3) we choose these sub-ensembles in *EnsSet* with their ensemble diversity scores above this pruning threshold and add them into the set

*GE<sub>ns</sub>Set* of high quality ensembles. The rest sub-ensembles will be pruned out given their lower ensemble diversity than the pruning threshold.

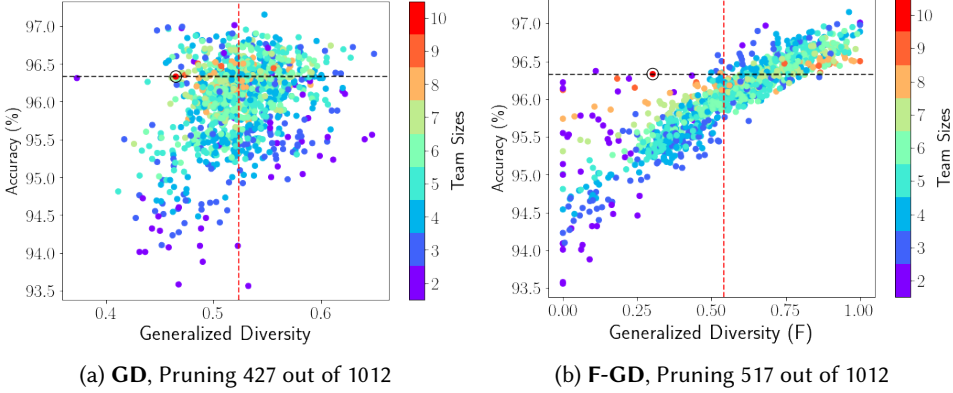


Fig. 1. Pruning All Possible Deep Ensembles by The Mean Threshold on CIFAR-10: (a) Baseline Diversity (GD), (b) Focal Diversity (F-GD)

We compute the baseline GD scores and ensemble accuracy for all 1013 (1012 sub-ensembles + 1 entire ensemble) deep ensembles on CIFAR-10 as shown in Figure 1a, where each dot denotes an ensemble team and each color marks the team size in the right color mapping bar. The baseline GD ensemble pruning approach utilizes the GD mean threshold of 0.524 (vertical red dashed line) to select sub-ensemble teams on the right of this diversity threshold line into *GE<sub>ns</sub>Set*. The ensemble accuracy 96.33% of the entire ensemble (marked by the black circle) is represented by the horizontal black dashed line, which serves as the reference accuracy for evaluating ensemble pruning methods. Ideally, the ensemble pruning algorithms should identify high quality sub-ensembles with their ensemble accuracy above this reference accuracy.

For an entire ensemble of a large size  $M$ , we use four performance metrics to assess the ensemble pruning algorithms: (1) the accuracy range of the selected sub-ensembles in *GE<sub>ns</sub>Set*, (2) the precision, which measures the proportion of the selected sub-ensembles with equal or higher ensemble accuracy than the reference accuracy of the given entire ensemble, e.g., 96.33% accuracy for CIFAR-10 and 79.82% accuracy for ImageNet, among all selected sub-ensembles, (3) the recall, which measures the proportion of the selected sub-ensembles in *GE<sub>ns</sub>Set* over all the sub-ensembles in *EnsSet* whose ensemble accuracy is equal to or higher than the reference accuracy of the entire deep ensemble, and (4) the cost reduction, which measures the reduction in the ensemble team size of the selected sub-ensembles relative to the entire ensemble size  $M$ . For example, we show a small sub-ensemble, 123, in Table 3 with the cost reduction of 70% ( $= (M - S)/M = (10 - 3)/10$ ). We evaluate the baseline mean threshold based ensemble pruning algorithm using the baseline BD (2nd column) and GD (4th column) diversity metrics on the CIFAR-10 dataset in Table 6 and ImageNet dataset in Table 7. The BD-based baseline pruning suffers from low precision (6.18% for CIFAR-10 and 8.53% for ImageNet) and recall (11.72% for CIFAR-10 and 23.15% for ImageNet). Even though the GD-based baseline pruning has higher precision (36.90% for CIFAR-10 and 17.74% for ImageNet) and recall (63.10% for CIFAR-10 and 46.31% for ImageNet), which is still below the acceptable values, such as precision of over 50%. Moreover, the accuracy lower bound for both BD and GD based baseline ensemble pruning, i.e., 93.56% on CIFAR-10 and 61.39% (BD) and 70.79% (GD) on ImageNet, is much lower than the reference accuracy of the entire ensemble team, i.e.,

96.33% accuracy for CIFAR-10 and 79.82% accuracy for ImageNet. We find similar observations on other diversity metrics, including CK and KW. This motivates us to investigate how to properly measure ensemble diversity and optimize the ensemble pruning process.

Table 6. Baseline Ensemble Pruning by Mean Threshold (CIFAR-10)

Methods	BD>0.339 (baseline)	F-BD>0.602 (optimized)	GD>0.524 (baseline)	F-GD>0.543 (optimized)
Acc Range (%) of $G_{EnsSet}$	93.56~96.72	<b>95.71</b> ~97.15	93.56~97.15	<b>95.71</b> ~97.15
Precision (%)	6.18	<b>63.53</b>	36.90	<b>53.90</b>
Recall (%)	11.72	<b>95.51</b>	63.10	<b>97.59</b>
Cost Reduction	10%~80%	10%~80%	10%~80%	10%~80%

Table 7. Baseline Ensemble Pruning by Mean Threshold (ImageNet)

Methods	BD>0.314 (baseline)	F-BD>0.632 (optimized)	GD>0.335 (baseline)	F-GD>0.594 (optimized)
Acc Range (%) of $G_{EnsSet}$	61.39~80.54	<b>76.77</b> ~80.77	70.79~80.60	<b>76.41</b> ~80.77
Precision (%)	8.53	<b>41.93</b>	17.74	<b>36.75</b>
Recall (%)	23.15	<b>98.52</b>	46.31	<b>99.01</b>
Cost Reduction	10%~80%	10%~80%	10%~80%	10%~80%

**Focal Diversity based Hierarchical Pruning.** We improve the ensemble diversity measurement and ensemble pruning process with three novel techniques in this paper. *First*, we use focal diversity metrics to optimize the ensemble diversity measurements, which utilizes the concept of focal models to sample negative samples and precisely capture the failure independence of member networks of a deep ensemble team. *Second*, we introduce a novel hierarchical pruning approach, which leverages focal diversity metrics and iteratively prunes out subsets of redundant member networks with low diversity from the entire ensemble. *Third*, we combine multiple focal diversity metrics through focal diversity consensus voting to further enhance the hierarchical ensemble pruning performance.

### 3 FOCAL DIVERSITY BASED HIERARCHICAL PRUNING

#### 3.1 Focal Diversity Concept and Algorithm

Our focal diversity metrics are designed to provide more accurate measurements of failure independence of the member networks in a deep ensemble team, including both pairwise (F-CK and F-BD) and non-pairwise (F-GD and F-KW) diversity metrics. Unlike traditional ensemble diversity evaluation approaches, which randomly select samples from the validation set to evaluate ensemble diversity, our focal diversity measurements draw random negative samples from  $NegSampSet(F_f)$  based on a specific focal model  $F_f$  on which the focal model  $F_f$  makes prediction errors and then calculate the focal negative correlation score. For a given deep ensemble of  $S$  member networks, each of the  $S$  member networks will serve as the focal model ( $F_f$ ) to draw negative samples. Therefore, we have a total of  $S$  focal negative correlation scores, where each corresponds to a member network serving as the focal model. Then we obtain the focal diversity score for this deep ensemble of  $S$  member networks through the (weighted) average of  $S$  focal negative correlation scores. The focal model concept is motivated by ensemble defense against adversarial attacks [9, 34], where the attack victim model is protected in a defense ensemble team. The focal model can be viewed as the victim model to evaluate the failure independence of other member models to this focal model in an ensemble team, i.e., the focal negative correlation score. Thus, the ensemble diversity score for

this ensemble of size  $S$  can be computed as the (weighted) average of  $S$  focal negative correlation scores by taking each member model as a focal model to mitigate the bias in diversity evaluation using only one focal model. Our preliminary results have demonstrated promising performance of the focal diversity metrics in capturing the failure independence of the member networks of a deep ensemble team [37–39].

---

**Algorithm 1** Focal Diversity Metric Calculation
 

---

```

1: procedure GETFQ(NegSampSet, Q, EnsSet)
2:   Input: NegSampSet: negative sample sets for each focal model  $F_f$ ; Q: the focal diversity metric, such
   as F-CK, F-BD, F-KW and F-GD; EnsSet: the set of candidate sub-ensemble teams to be considered;
3:   Output: FQ: focal diversity scores
4:   Initialize  $D(Q) = \{\}$ ,  $\bar{D}(Q) = \{\}$ 
5:   Initialize  $FQ = \{\}$  ▷ A map of diversity scores and teams
6:   for  $S = 2$  to  $M - 1$  do
7:     for  $f = 0$  to  $M - 1$  do
8:       Obtain  $EnsSet(F_f, S)$  with candidate sub-ensembles of size  $S$  and containing the focal model
        $F_f$ .
9:       Initialize  $D(Q, S, F_f) = \{\}$ 
10:      for  $i = 1$  to  $|EnsSet(F_f, S)|$  do
11:        ▷ calculate the focal negative correlation score for  $T_i \in EnsSet(F_f, S)$  using the definition
        of  $Q$ 
12:           $q_i = FocalNegativeCorrelation(Q, T_i, NegSampSet(F_f))$ 
13:           $D(Q, S, F_f).append(q_i)$ 
14:        end for
15:      for  $i = 1$  to  $|EnsSet(F_f, S)|$  do
16:        ▷ scale focal negative correlation scores for sub-ensembles of the same size  $S$ 
17:           $\bar{D}(Q, S, F_f, T_i) = \frac{q_i - \min(D(Q, S, F_f))}{\max(D(Q, S, F_f)) - \min(D(Q, S, F_f))}$ 
18:        end for
19:      end for
20:      Obtain  $EnsSet(S)$  with candidate sub-ensemble teams of size  $S$ 
21:      for  $i = 1$  to  $|EnsSet(S)|$  do
22:        Initialize  $tmpD = \{\}$ 
23:        for  $j = 1$  to  $|T_i|$  do
24:           $tmpD.append(\bar{D}(Q, S, F_f = T_i[j], T_i))$ 
25:        end for
26:        ▷ Obtain the member model accuracy ranks as the member model weights.
27:         $w = MemberModelAccuracyRank(T_i)$ 
28:         $FQ(T_i) = WeightedAverage(w, tmpD)$ 
29:      end for
30:    end for
31:  return FQ
32: end procedure

```

---

In general, our focal diversity metrics compare the ensemble diversity scores among the ensembles of the same size  $S$ , randomly draw negative samples from each focal model ( $F_f$ , i.e.,  $NegSampSet(F_f)$ ) and calculate the focal diversity score as the average of  $S$  focal negative correlation scores by taking each member model as the focal model. Algorithm 1 gives a skeleton of calculating the focal diversity scores for all the candidate sub-ensembles in  $EnsSet$ . For each team size  $S$  (Line 6~30), we follow two general steps to calculate the focal diversity scores for each ensemble. *First*, for each member model  $F_f$ , let  $EnsSet(F_f, S)$  denote all candidate ensembles



of size  $S$ , each containing the focal model  $F_f$ . We first compute the focal negative correlation score for each ensemble in  $EnsSet(F_f, S)$  with the negative samples randomly drawn from the focal model  $F_f$  ( $NegSampSet(F_f)$ ) and store them in  $D(Q, S, F_f)$  (Line 10~14). Then, in order to make them comparable across different member models ( $F_f$ ), we scale  $D(Q, S, F_f)$  into  $[0, 1]$  and store them in  $\bar{D}(Q, S, F_f, T_i)$  for each ensemble team  $T_i$  (Line 15~18). *Second*, for each candidate sub-ensemble ( $T_i$ ) of size  $S$ , we perform a weighted average of the scaled focal negative correlation scores  $\bar{D}(Q, S, F_f = T_i[j], T_i)$  associated with each of its focal (member) model  $F_f = T_i[j]$  to obtain the focal diversity score. The weight is calculated with the corresponding rank of the accuracy of the member model ( $T_i[j]$ ) in the ensemble team ( $T_i$ ), i.e., the member models with higher accuracy will have higher weights (Line 21~29). In addition, we subtract the CK value from 1 when using the CK formula [25] to present the consistent view that high diversity values correspond to high ensemble diversity.

We show a visual comparison between our focal diversity metric F-GD in Figure 1b and the baseline GD metric in Figure 1a for CIFAR-10. Here the baseline mean threshold based ensemble pruning method is used for both F-GD and GD diversity metrics. Compared to the GD based ensemble pruning, the F-GD powered pruning obtains better ensemble pruning results and identifies a larger portion of sub-ensembles (on the right side of the vertical red dashed line) with ensemble accuracy above the reference accuracy of 96.33% (horizontal black dashed line).

When the focal diversity metrics are used in our hierarchical pruning algorithm with a desired team size  $S_d$ , we only calculate the diversity scores for the ensembles of size  $2 \sim S_d$  by replacing  $M - 1$  with  $S_d$  in Line 6 of Algorithm 1 to reduce the computation cost. Compared to the baseline diversity metrics, our focal diversity score computation takes about 1.2~2.4 $\times$  the time of the baseline diversity score computation for all candidate sub-ensembles in  $EnsSet$ . The computation of all focal diversity metrics can be finished in several seconds on MNIST, CIFAR-10 and Cora and in several minutes on ImageNet (see Table 16). Given that our focal diversity metrics significantly outperform the baseline diversity metrics in capturing the failure independence of a group of member models, it is beneficial and worthwhile to use our focal diversity metrics to measure the ensemble diversity and identify high quality sub-ensembles. In addition, our hierarchical pruning can effectively prune out about 80% of the ensemble teams in  $EnsSet$ , which also compensates for the increased computation time of the focal diversity metrics, making our entire solution very efficient.

We show the baseline mean threshold based ensemble pruning results using our focal diversity F-BD and F-GD in the 3rd and 5th columns of Table 6 for CIFAR-10 and Table 7 for ImageNet. Both F-BD and F-GD (optimized) substantially outperform the baseline BD and GD based ensemble pruning, as measured by the accuracy range of the selected sub-ensembles in  $GEnsSet$ , precision and recall. For the other two diversity metrics, CK and KW, we observed similar performance improvements of using our focal diversity metrics, F-CK and F-KW, in the baseline ensemble pruning. Even though our focal diversity metrics can significantly improve the baseline mean threshold based ensemble pruning method, achieving very high recall of over 95%, we can still enhance the 63.53% precision of F-BD and 53.90% precision of F-GD for CIFAR-10 and the 41.93% precision of F-BD and 36.75% precision of F-GD for ImageNet to more accurately identify high-quality sub-ensembles. In addition, the baseline mean threshold based ensemble pruning examines every sub-ensemble in  $EnsSet$  with high computational cost. These observations have motivated us to explore new methods to enhance ensemble pruning efficiency and accuracy.

---

**Algorithm 2** Hierarchical Pruning
 

---

```

1: procedure HQ-PRUNING( $Q, S_d, \beta, EnsSet$ )
2:   Input:  $Q$ : the focal diversity metric;  $S_d$ : the desired ensemble size;  $\beta$ : the percentage of the number of
   ensemble teams to be pruned out in each iteration;  $EnsSet$ : the set of sub-ensemble teams to be considered;
3:   Output:  $GEnsSet(Q, S_d)$ : the set of good ensemble teams of size  $S_d$  identified by the focal diversity
   metric  $Q$ .
4:   Initialize  $pruneSet = \{\}$  ▷ Subsets of member models to prune out.
5:   for  $S = 2$  to  $S_d$  do
6:     Initialize  $GEnsSet(Q, S) = \{\}, D = \{\}$ 
7:     Construct  $EnsSet(S)$  of size  $S$  ensembles
8:     for  $i = 1$  to  $|EnsSet(S)|$  do
9:       if  $T_i$  contains any subset in  $pruneSet$  then
10:        continue ▷ Prune out this sub-ensemble  $T_i$  and avoid expanding along this branch.
11:       else
12:          $q_i = FQ(T_i)$  ▷  $FQ(T_i)$  is obtained through Algorithm 1 by using the focal diversity  $Q$ .
13:          $D.append(q_i)$ 
14:          $GEnsSet(Q, S).add(T_i)$ 
15:       end if
16:     end for
17:      $n = \beta \times |GEnsSet(Q, S)|$ 
18:     Sort  $T_i \in GEnsSet(Q, S)$  by  $q_i \in D$ 
19:     Remove  $n$  sub-ensembles of the lowest ensemble diversity from  $GEnsSet(Q, S)$  and add them into
      $pruneSet$ 
20:   end for
21:   return  $GEnsSet(Q, S_d)$ 
22: end procedure

```

---

### 3.2 Hierarchical Pruning Overview and Algorithm

Our focal diversity metrics reveal some anti-monotonicity property: a superset of a low diversity ensemble has also low diversity. Specifically, a low focal diversity score for an ensemble team, e.g.,  $F_0F_2$ , often implies insufficient ensemble diversity, that is a high correlation of its member models in making similar prediction errors, making the member models highly redundant for a large superset ensemble team, such as  $F_0F_1F_2$ . Hence, those large ensembles that contain a small sub-ensemble with low diversity (e.g.,  $F_0F_2$ ), such as  $F_0F_1F_2$ ,  $F_0F_2F_3$ ,  $F_0F_1F_2F_3$ , and  $F_0F_1F_2F_4$  tend to have lower ensemble diversity than other ensembles with the same size and more diverse member models. Therefore, when we identify a sub-ensemble with low ensemble diversity, these large ensembles that are the supersets of this sub-ensemble can be preemptively pruned out, which motivates our hierarchical ensemble pruning approach. The pseudo code is given in Algorithm 2. Overall, our hierarchical pruning is an iterative process of composing and selecting deep ensembles for a desired ensemble team size  $S_d$ . *First*, we start the hierarchical pruning process with the ensembles of size  $S = 2$ , that is  $|EnsSet(S = 2)| = \binom{M}{2} = 10(10 - 1)/2 = 45$  candidate ensemble teams. *Second*, for a given focal diversity metric, we rank candidate ensembles of size  $S$ , such as  $S = 2$ , by their focal diversity scores ( $q_i$ ), prune out the bottom  $\beta$  percentage of ensembles with low focal diversity scores, and add them into  $pruneSet$  as the pruning targets (Line 17~19). Here, a dynamic  $\beta$  can be configured to accommodate the concrete number of candidate ensembles and diversity score distribution. A conservative strategy is recommended to set a small  $\beta$ , e.g., by default  $\beta = 10\%$ . *Third*, we preemptively prune out all subsequent ensemble teams with a larger size  $S + 1$  that contain one or more pruned sub-ensembles in  $pruneSet$  (Line 7~16). By iterating through these

pruning steps, our hierarchical pruning can efficiently identify high quality sub-ensembles of the desired team size  $S_d$  and add them into  $GENsSet(Q, S_d)$ .

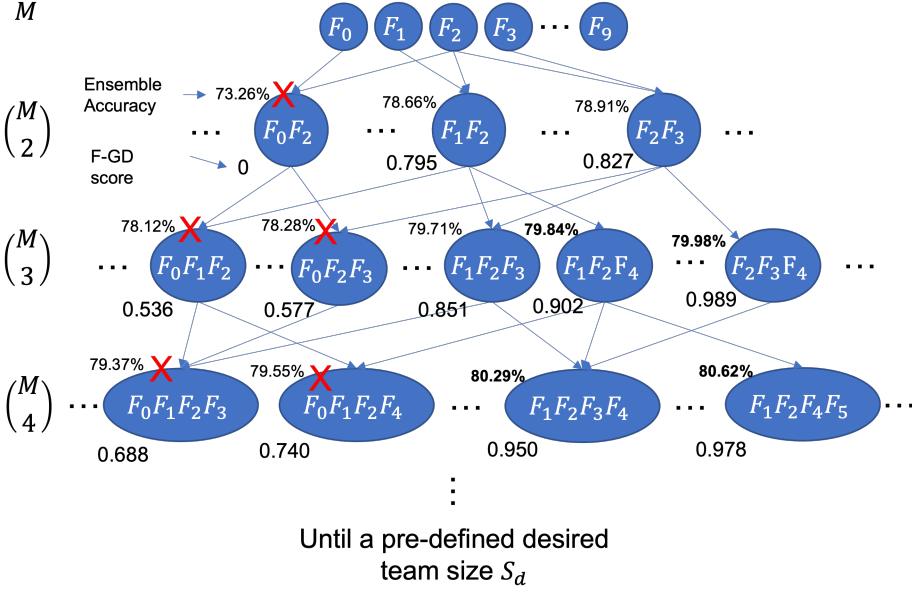


Fig. 2. Hierarchical Pruning On ImageNet

We present a hierarchical ensemble pruning example on ImageNet in Figure 2. With  $M = 10$ , we have all 10 member networks on the top, followed by all sub-ensembles of size  $S = 2$ . For each tier, we add one additional network to the candidate ensemble teams of size  $S$  ( $2 \leq S < S_d$ ) and place these extended ensembles of size  $S + 1$  in the next tier until  $S = S_d$ , where the ensembles of the desired team size  $S_d$  will be placed in the last tier. Meanwhile, for each pruned ensemble, such as  $F_0F_2$ , our hierarchical pruning algorithm will preemptively cut off these branches of candidate ensemble teams that contain this pruned ensemble, i.e., all ensembles that are supersets of  $F_0F_2$ , marked by the red cross in Figure 2. This way of building ensemble teams allows us to compose high quality sub-ensembles and strategically remove the ensembles with low diversity. As the example in Figure 2 shows that our hierarchical pruning indeed can effectively prune out these ensemble teams with low diversity and low accuracy, such as  $F_0F_1F_2$ ,  $F_0F_2F_3$ ,  $F_0F_1F_2F_3$  and  $F_0F_1F_2F_4$ , and avoid exploring those unpromising branches.

We first apply our F-BD and F-GD hierarchical ensemble pruning algorithms on CIFAR-10 to prune the entire deep ensemble of 10 member networks by setting  $\beta = 10\%$  and  $S_d = 5$ . Figure 3 visualizes the ensemble pruning results, where the black and red dots denote the pruned and selected ensemble teams respectively. Three interesting observations should be highlighted. *First*, for our focal diversity metrics, including F-BD and F-GD, the ensemble teams with high focal diversity scores tend to have high ensemble accuracy, especially when we compare the ensemble teams of the same size  $S$ , e.g.,  $S=3, 4, 5$ , which is an important property to guide ensemble pruning. *Second*, our hierarchical ensemble pruning approach can effectively identify these ensemble teams with insufficient ensemble diversity by their low focal diversity scores. The effectiveness of our hierarchical pruning can be attributed to two factors. On the one hand, our focal diversity powered hierarchical pruning encourages more fair comparison of focal diversity scores among these

ensembles of the same size  $S$ . Comparing to Figure 1b showing all possible ensembles of mixed team sizes, Figure 3 shows much clearer correlation between the focal diversity score and ensemble accuracy for the ensembles of the same size  $S = 3, 4, 5$ . On the other hand, our focal diversity metrics can more precisely capture the failure independence of member networks of a deep ensemble with some anti-monotonicity property, ensuring that the focal diversity based ensemble pruning is highly accurate and efficient. *Third*, the time and space complexity of the hierarchical pruning algorithm and pruned ensemble execution cost can be ultimately bounded by the desired team size  $S_d$ . For example, we recommend setting the desired ensemble team size  $S_d$  up to  $50\% \times M$ , which allows our hierarchical pruning to effectively find these sub-ensembles that offer on par or improved ensemble accuracy over the entire ensemble of  $M$  member networks and have a substantially smaller team size, such as one third or one half of  $M$  with a significant cost reduction in ensemble execution. *Finally*, for Figure 3c and 3f, all the selected sub-ensembles (red dots) with the desired team size  $S_d = 5$  provide higher ensemble accuracy than the reference accuracy of 96.33% for CIFAR-10, which leads to 100% precision of our hierarchical pruning with F-BD and F-GD. We observe similar ensemble pruning results using the other two focal diversity metrics, F-CK and F-KW.

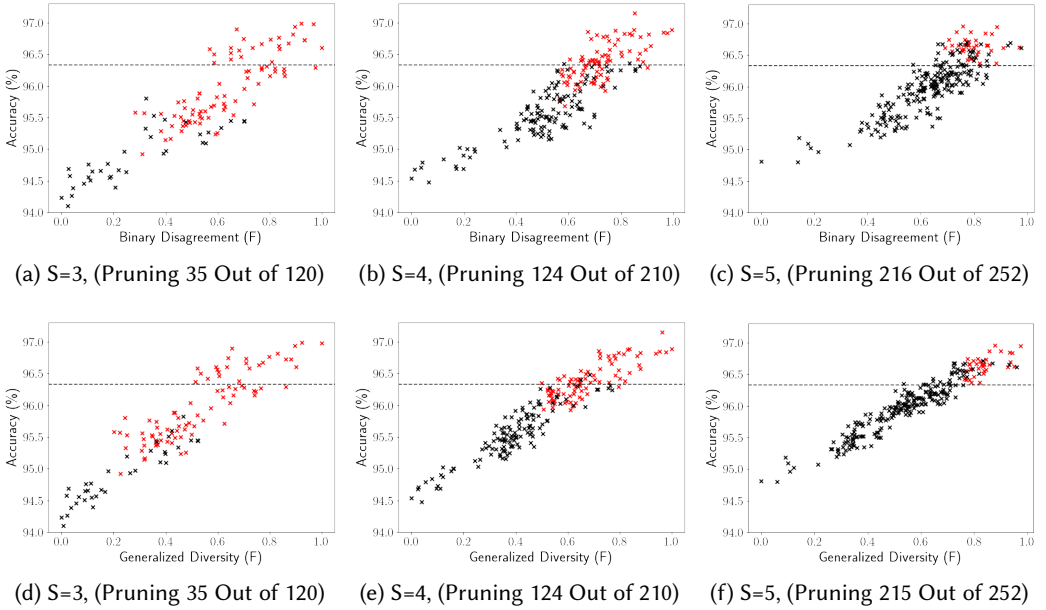


Fig. 3. Deep Ensembles of Size  $S = 3, 4, 5$  on CIFAR-10: top three figures for **F-BD** and bottom three figures for **F-GD** ( $\beta = 10\%$ ,  $S_d = 5$ )

**Focal Diversity Pruning by Focal Diversity Consensus Voting.** Using different focal diversity metrics in the hierarchical ensemble pruning may produce different sets of selected ensembles ( $GENS_{Set}$ ). It is observed that these ensemble teams that are chosen by the majority of our focal diversity metrics in ensemble pruning (e.g., F-BD pruning, F-KW pruning, F-GD pruning, etc.) tend to consistently outperform the entire deep ensemble in terms of the ensemble accuracy and execution cost. Therefore, we introduce the third step in our hierarchical ensemble pruning, which consolidates these ensembles that are selected by different focal diversity metrics into a majority voting based final selection. Concretely, an ensemble team will only be added to this final selection

if it is chosen by the majority of focal diversity metrics, i.e., at least three focal diversity metrics, via hierarchical pruning. This third step further refines the focal diversity powered hierarchical pruning results and consistently delivers enhanced precision and robustness in ensemble pruning.

## 4 FORMAL ANALYSIS

Deep neural network ensembles use multiple deep neural networks to form a team to collaborate and combine the predictions of individual member networks to make the final prediction. It is a commonly held view that an ensemble team consisting of diverse member networks has high prediction performance [18, 24, 32]. Our focal diversity based ensemble pruning methods can efficiently identify small pruned deep ensembles with highly diverse member networks. These pruned deep ensembles can achieve the same or even improved prediction performance with significantly reduced space and time costs compared to the entire deep ensemble. In this section, we present a formal analysis to show the desired properties and features of these pruned deep ensembles to further demonstrate the effectiveness of our proposed methods.

### 4.1 Diversity by Uncorrelated Errors

For an ensemble team of size  $S$ , following [33, 39], we can derive the added error for its ensemble prediction  $E_{add}^{avg}$  using model averaging (*avg*) as the ensemble consensus method as Formula (1) shows.

$$E_{add}^{avg} = E_{add} \left( \frac{1 + (S - 1)\delta}{S} \right) \quad (1)$$

where  $E_{add}$  is the added error of a single network and  $\delta$  is the expected average correlation of all member networks in the ensemble. Therefore, the ideal scenario is when all member networks in an ensemble team of size  $S$  are diverse. They can learn and predict with uncorrelated errors (failure independence), i.e.,  $\delta \leq 0$ . Then a simple model averaging method can significantly reduce the overall prediction error by at least  $S$  times. Meanwhile, the worst scenario happens when errors of individual networks are highly correlated with  $\delta = 1$ . For example, when all  $S$  member networks are perfect duplicates, the error of the ensemble is identical to the initial error without any improvement. In practice, given that it is challenging to directly measure the correlation  $\delta$ , many ensemble diversity metrics are proposed to quantify the correlation among member networks of an ensemble team. Our focal diversity metrics can significantly improve ensemble diversity measurement, and they are closely correlated to the ensemble prediction performance, which can be directly leveraged for identifying high-quality ensemble teams.

### 4.2 Ensemble Robustness

The deep neural network model is typically trained to minimize a cross-entropy loss and output a probability vector to approximate posteriori probabilities for the corresponding classes. Let  $f_i(\mathbf{x})$  denote the  $i$ th element in the probability vector of a classifier  $F$  for input  $\mathbf{x}$ , which predicts the probability of class  $i$ . The classifier  $F$  will output the predicted label  $c$  with the highest probability for input  $\mathbf{x}$ , that is  $c = \operatorname{argmax}_{1 \leq i \leq C} f_i(\mathbf{x})$ .

The robustness of a classifier  $F$  can be assessed based on its ability to maintain consistent prediction for a given input  $\mathbf{x}$  under input perturbation ( $\mu$ ), such as the noise from adversarial samples. When the magnitude of the input perturbation  $\mu$  remains within a certain bound, the prediction made by the classifier  $F$  remains unaffected. We can leverage such a bound to compare the level of robustness of different classifiers, where a higher bound indicates a higher level of robustness. In Theorem 1, we introduce the concept of robustness bound ( $R$ ) and formally show that this bound fulfills the aforementioned requirement.

**Theorem 1** (Robustness Bound ( $R$ )). The robustness bound ( $R$ ) for a classifier  $F$  can be defined as the following Formula (2),

$$R = \min_{j \neq c} \frac{f_c(\mathbf{x}_0) - f_j(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(f_c(\mathbf{x}) - f_j(\mathbf{x}))\|_q} \quad (2)$$

where  $\mathbf{x} = \mathbf{x}_0 + \mu$ . When the magnitude of the perturbation  $\mu$  on the input  $\mathbf{x}_0$  is limited by  $\|\mu\|_p \leq R$  and  $p, q$  satisfy  $\frac{1}{p} + \frac{1}{q} = 1$  and  $1 \leq p, q \leq \infty$ , the predicted labels for  $\mathbf{x}$  and  $\mathbf{x}_0$  by the classifier  $F$  will be identical, indicating the input perturbation ( $\mu$ ) will not change the prediction.

We first introduce Lemma 1 on Lipschitz continuity below and then show the formal proof of Theorem 1.

**Lemma 1** (Lipschitz Continuity). If  $g(\mathbf{x})$  is Lipschitz continuous, according to [28], the following inequality (3) holds:

$$|g(\mathbf{x}) - g(\mathbf{y})| \leq L_q \|\mathbf{x} - \mathbf{y}\|_p \quad (3)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are inputs, Lipschitz constant  $L_q = \max_{\mathbf{x}} \|\nabla g(\mathbf{x})\|_q$ ,  $\frac{1}{p} + \frac{1}{q} = 1$ , and  $1 \leq p, q \leq \infty$ .

**PROOF OF THEOREM 1.** Without loss of generality, we assume  $g(\mathbf{x}) = f_c(\mathbf{x}) - f_j(\mathbf{x})$  is Lipschitz continuous with Lipschitz constant  $L_q^j$  and  $j \neq c$ . Following Lemma 1, let  $\mathbf{x} = \mathbf{x}_0 + \mu$  and  $\mathbf{y} = \mathbf{x}_0$ , we derive Formula (4) below,

$$|g(\mathbf{x}_0 + \mu) - g(\mathbf{x}_0)| \leq L_q^j \|\mu\|_p \quad (4)$$

where  $L_q^j = \max_{\mathbf{x}} \|\nabla g(\mathbf{x})\|_q = \max_{\mathbf{x}} \|\nabla(f_c(\mathbf{x}) - f_j(\mathbf{x}))\|_q$ ,  $\frac{1}{p} + \frac{1}{q} = 1$ , and  $1 \leq p, q \leq \infty$ .

Formula (4) can be rearranged as Formula (5),

$$g(\mathbf{x}_0) - L_q^j \|\mu\|_p \leq g(\mathbf{x}_0 + \mu) \leq g(\mathbf{x}_0) + L_q^j \|\mu\|_p \quad (5)$$

when  $g(\mathbf{x}_0 + \mu) < 0$ , the predicted class label will change. However,  $g(\mathbf{x}_0 + \mu)$  is lower bounded by  $g(\mathbf{x}_0) - L_q^j \|\mu\|_p \leq g(\mathbf{x}_0 + \mu)$ . If  $0 \leq g(\mathbf{x}_0) - L_q^j \|\mu\|_p$ , we have  $g(\mathbf{x}_0 + \mu) \geq 0$  to ensure that the prediction will not change with the small perturbation  $\mu$  on the input  $\mathbf{x}_0$ . This leads to Formula (6),

$$g(\mathbf{x}_0) - L_q^j \|\mu\|_p \geq 0 \Rightarrow \|\mu\|_p \leq \frac{g(\mathbf{x}_0)}{L_q^j} \quad (6)$$

that is Formula (7):

$$\|\mu\|_p \leq \frac{f_c(\mathbf{x}_0) - f_j(\mathbf{x}_0)}{L_q^j} \quad (7)$$

where  $L_q^j = \max_{\mathbf{x}} \|\nabla(f_c(\mathbf{x}) - f_j(\mathbf{x}))\|_q$ ,  $\frac{1}{p} + \frac{1}{q} = 1$ , and  $1 \leq p, q \leq \infty$ .

In order to ensure the classification result will not change, that is  $\operatorname{argmax}_{1 \leq i \leq C} f_i(\mathbf{x}_0 + \mu) = c$ , we use the minimum of the bound on  $\mu$  over  $j \neq c$  to obtain the inequality (8) with  $\frac{1}{p} + \frac{1}{q} = 1$  and  $1 \leq p, q \leq \infty$ ,

$$\|\mu\|_p \leq \min_{j \neq c} \frac{f_c(\mathbf{x}_0) - f_j(\mathbf{x}_0)}{L_q^j} = \min_{j \neq c} \frac{f_c(\mathbf{x}_0) - f_j(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(f_c(\mathbf{x}) - f_j(\mathbf{x}))\|_q} = R \quad (8)$$

which indicates that as long as  $\|\mu\|_p$  is small enough to fulfill the above bound, the classifier decision will never be changed, which marks the robustness of this classifier  $F$ . Hence, we formally prove Theorem 1 on the robustness bound ( $R$ ) for a classifier  $F$ .  $\square$

For a deep neural network  $F_k$ , we have its robustness bound as Formula (9) shows.

$$R^k = \min_{j \neq c} \frac{f_c^k(\mathbf{x}_0) - f_j^k(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(f_c^k(\mathbf{x}) - f_j^k(\mathbf{x}))\|_q} \quad (9)$$

Let  $g_j^k(\mathbf{x}) = f_c^k(\mathbf{x}) - f_j^k(\mathbf{x})$ , we have the robustness bound as Formula (10) shows.

$$R^k = \min_{j \neq c} \frac{g_j^k(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q} \quad (10)$$

Given  $S$  networks, combining their predictions with model averaging (*avg*), we have the  $i$ th element in the combined probability vector as  $f_i^{avg}(\mathbf{x}) = \frac{1}{S} \sum_{k=1}^S f_i^k(\mathbf{x})$  corresponding to the robustness bound as the following Formula (11) shows.

$$\begin{aligned} R^{avg} &= \min_{j \neq c} \frac{f_c^{avg}(\mathbf{x}_0) - f_j^{avg}(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(f_c^{avg}(\mathbf{x}) - f_j^{avg}(\mathbf{x}))\|_q} \\ &= \min_{j \neq c} \frac{g_c^{avg}(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(g_c^{avg}(\mathbf{x}))\|_q} \end{aligned} \quad (11)$$

Assume the minimum of the robustness bound can be achieved with the prediction result  $c$  and  $j$  for each model  $F^k$  including the ensemble  $F^{avg}$  as Formula (12) shows.

$$\begin{aligned} R^k &= \frac{g_j^k(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q} \\ R^{avg} &= \frac{g_j^{avg}(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(g_j^{avg}(\mathbf{x}))\|_q} \end{aligned} \quad (12)$$

where  $g_j^k(\mathbf{x}) = f_c^k(\mathbf{x}) - f_j^k(\mathbf{x})$  and  $g_j^{avg}(\mathbf{x}) = \frac{1}{S} \sum_{k=1}^S g_j^k(\mathbf{x})$ .

We can then derive Theorem 2 that states  $\exists k, 1 \leq k \leq S, R^k \leq R^{avg}$ , where the equal sign corresponds to the case that all member networks are the perfect duplicates. This theorem further indicates that the ensembles of high diversity can improve the robustness of individual member networks.

**Theorem 2** (Ensemble Robustness Bound Enhancement). Let  $R^{avg}$  denote the robustness bound for an ensemble, which combines its member network predictions through model averaging (*avg*). We can always find one member network  $F^k$  with its robustness bound  $R^k$  satisfying  $R^k \leq R^{avg}$ .

**PROOF OF THEOREM 2.** We prove the ensemble robustness bound enhancement by contradiction. First, we assume  $\forall k, 1 \leq k \leq S, R^k > R^{avg}$ , that is as Formula (13) shows.

$$g_j^k(\mathbf{x}_0) (\max_{\mathbf{x}} \|\nabla(g_j^{avg}(\mathbf{x}))\|_q) > g_j^{avg}(\mathbf{x}_0) (\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q) \quad (13)$$

For each  $k \in \{1, \dots, S\}$ , this inequality holds. To add them all, we have Formula (14).

$$\sum_{k=1}^S g_j^k(\mathbf{x}_0) (\max_{\mathbf{x}} \|\nabla(g_j^{avg}(\mathbf{x}))\|_q) > \sum_{k=1}^S g_j^{avg}(\mathbf{x}_0) (\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q) \quad (14)$$

That is Formula (15).

$$(\max_{\mathbf{x}} \|\nabla(g_j^{avg}(\mathbf{x}))\|_q) \sum_{k=1}^S g_j^k(\mathbf{x}_0) > g_j^{avg}(\mathbf{x}_0) \sum_{k=1}^S (\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q) \quad (15)$$

Given  $g_j^{avg}(\mathbf{x}) = \frac{1}{S} \sum_{k=1}^S g_j^k(\mathbf{x})$ , we have Formula (16):

$$(\max_{\mathbf{x}} \|\nabla(\sum_{k=1}^S g_j^k(\mathbf{x}))\|_q) \frac{1}{S} \sum_{k=1}^S g_j^k(\mathbf{x}_0) > \frac{1}{S} \sum_{k=1}^S g_j^k(\mathbf{x}_0) \sum_{k=1}^S (\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q) \quad (16)$$

Therefore, we have the following Formula (17).

$$(\max_{\mathbf{x}} \|\nabla(\sum_{k=1}^S g_j^k(\mathbf{x}))\|_q) > \sum_{k=1}^S (\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q) \quad (17)$$

According to the triangle inequality, we have Formula (18).

$$\begin{aligned} \max_{\mathbf{x}} \|\nabla(\sum_{k=1}^S g_j^k(\mathbf{x}))\|_q &\leq \max_{\mathbf{x}} (\sum_{k=1}^S \|\nabla(g_j^k(\mathbf{x}))\|_q) \\ &\leq \sum_{k=1}^S (\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q) \end{aligned} \quad (18)$$

which contradicts with the derived inequality (Formula (17)). Therefore, the previous assumption does not hold. We show that  $\exists k, 1 \leq k \leq S, R^k \leq R^{avg}$ , demonstrating that the robustness of a member network can be further improved with an ensemble team. Furthermore, for a network  $F^k$ , if its robustness bound  $R^k$  was not obtained with class  $j$ . We have  $\exists i \neq j$  ( $i, j \neq c$ ) and  $R^k = \frac{g_i^k(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(g_i^k(\mathbf{x}))\|_q} \leq \frac{g_j^k(\mathbf{x}_0)}{\max_{\mathbf{x}} \|\nabla(g_j^k(\mathbf{x}))\|_q}$ , where Theorem 2 still holds.  $\square$

The above analysis formally certifies that an ensemble team of diverse member networks can further improve the robustness of individual networks.

## 5 EXPERIMENTAL EVALUATION

We conducted a comprehensive experimental evaluation on four benchmark datasets, CIFAR-10, ImageNet, Cora and MNIST, for pruning the given entire ensemble of 10 individual member models for CIFAR-10, ImageNet and Cora and 7 member models for MNIST (see Table 4 for all member models). All the experiments were performed on an Intel i7-10700K server with the NVIDIA GeForce RTX 3090 (24GB) GPU on Ubuntu 20.04.

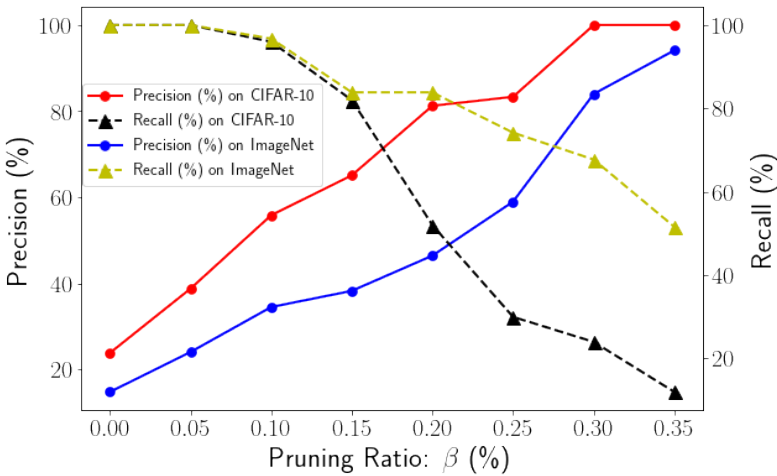


Fig. 4. Impact of  $\beta$  on Precision and Recall ( $S_d = 4$ , F-GD)



### 5.1 Efficiency of Pruning with Varying $\beta$

The hyperparameter  $\beta$  in our hierarchical ensemble pruning determines the percentage of candidate ensembles to be pruned out during each iteration for a specific ensemble team size  $S$ . In general, a higher  $\beta$  tends to remove more ensemble teams in our hierarchical pruning, potentially lowering the recall score. On the other hand, a high precision score is expected when the ensemble pruning algorithm can precisely find the high quality sub-ensembles with much smaller team sizes and equal or better ensemble accuracy of the reference accuracy of the entire ensemble team. We analyze the impacts of the hyperparameter  $\beta$  in Figure 4, where we vary  $\beta$  from 5% to 35% for both CIFAR-10 and ImageNet using F-GD based hierarchical pruning and the desired team size  $S_d = 4$ . As shown in Figure 4, when the  $\beta$  increases, the precision of ensemble pruning increases, and the recall decreases for both CIFAR-10 and ImageNet. Our focal diversity based hierarchical ensemble pruning is designed to achieve high precision and good recall, e.g., over 75% precision and 50% recall. Therefore, with  $S_d = 4$  on CIFAR-10, we set  $\beta = 20\%$  to obtain 81.25% precision and 52% recall of the F-GD based hierarchical pruning. We follow the same principles in determining the  $\beta$  value for other experiments.

Table 8. Impact of Desired Team Size  $S_d$  on Hierarchical Pruning (CIFAR-10)

$S_d$	4	5	6
$\beta$ (%)	20	10	4
Precision (%)	81.25	<b>100</b>	97.56
Recall (%)	52	47.14	57.14

### 5.2 Impact of Desired Team Size $S_d$

The hyperparameter  $S_d$  specifies the target ensemble team size after ensemble pruning. Table 8 presents the impacts of varying  $S_d$  on our focal diversity powered hierarchical pruning approach on CIFAR-10. For each  $S_d$ , we find the optimal  $\beta$  following the principles introduced in Section 5.1. We highlight two interesting observations. *First*, for different  $S_d$ , our hierarchical pruning can consistently deliver high precision of over 81.25% with good recall, which demonstrates the effectiveness of our hierarchical pruning approach. In particular,  $S_d = 5$  produces the highest precision of 100%. *Second*, as the desired team size  $S_d$  increases, there is a decrease in the optimal  $\beta$  value, indicating that fewer ensembles will be pruned out for each iteration to achieve a good recall rate. In practice, we recommend setting the desired ensemble team size  $S_d$  up to half the size of the entire ensemble, which allows our hierarchical pruning to efficiently find these sub-ensembles that offer comparable or improved ensemble accuracy over the entire ensemble and substantially reduce the ensemble execution cost.

### 5.3 Focal Diversity Pruning Methods

We then evaluate our hierarchical ensemble pruning algorithm with four focal diversity metrics and the focal diversity consensus voting based refinement. The precision, recall and cost reduction of ensemble pruning are primarily used to measure the pruning efficiency on four benchmark datasets.

**CIFAR-10.** We compare the four focal diversity metrics using our hierarchical ensemble pruning and the focal diversity consensus voting based pruning (MAJORITY-F) in Table 9, where we set  $\beta = 10\%$ ,  $S_d = 5$ , and the F-BD and F-GD pruning results correspond to Figure 3c and 3f. Two interesting observations should be highlighted. *First*, our hierarchical ensemble pruning approach achieves a very high precision of over 85% in ensemble pruning with all four focal diversity

Table 9. Hierarchical Pruning with  $S_d=5$  on CIFAR-10

Methods	F-CK	F-BD	F-KW	F-GD	MAJORITY-F
Precision (%)	85.71	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Recall (%)	17.14	51.43	51.43	52.86	47.14
Cost Reduction	50%	50%	50%	50%	50%

Table 10. Hierarchical Pruning with  $S_d=4$  on CIFAR-10

Methods	F-CK	F-BD	F-KW	F-GD	MAJORITY-F
Precision (%)	26.09	<b>81.25</b>	<b>81.25</b>	<b>81.25</b>	<b>81.25</b>
Recall (%)	12.00	52.00	52.00	52.00	52.00
Cost Reduction	60%	60%	60%	60%	60%

metrics. Especially, three focal diversity metrics, F-BD, F-KW and F-GD, produces 100% precision in identifying high quality sub-ensembles with much smaller team sizes and equal or better ensemble accuracy than the entire deep ensemble. *Second*, with the focal diversity consensus voting based ensemble pruning, coined as MAJORITY-F, the 100% precision is maintained. MAJORITY-F further refines the ensemble pruning results by pruning out these ensemble teams that are not chosen by the majority of four focal diversity metrics, which slightly reduces the total number of selected ensembles (i.e., true positives given 100% precision). Hence, the slightly lower recall score of MAJORITY-F is expected in comparison with F-BD, F-KW and F-GD. We find similar observations in Table 10 by setting  $\beta = 20\%$ ,  $S_d = 4$ . Even though the F-CK based hierarchical pruning achieves lower performance than the other three focal diversity metrics, the focal diversity consensus voting method, MAJORITY-F, can still deliver consistent high precision of 81.25% and recall of 52%.

Table 11. Hierarchical Pruning with  $S_d=5$  on ImageNet

Methods	F-CK	F-BD	F-KW	F-GD	MAJORITY-F
Precision (%)	0	75.61	75.61	74.42	<b>78.95</b>
Recall (%)	0	64.58	64.58	66.67	62.50
Cost Reduction	50%	50%	50%	50%	50%

**ImageNet.** We then evaluate our hierarchical ensemble pruning approach on ImageNet by setting  $\beta = 10\%$ ,  $S_d = 5$ . Table 11 shows the experimental results. We highlight two interesting observations. *First*, our hierarchical ensemble pruning approach performs very well with high precision of over 74.42% by using F-BD, F-KW and F-GD metrics, which is effective and accurate in selecting high quality sub-ensembles with a much smaller team size ( $S = 5$  vs.  $M = 10$ ) and higher ensemble accuracy than the 79.82% reference accuracy of the entire ensemble team on ImageNet. On the other hand, the F-CK based pruning failed to identify any satisfactory ensemble teams, which indicates the inherent limitation of the CK diversity, although it has been optimized by the focal diversity measures. *Second*, using our focal diversity consensus voting based ensemble pruning, MAJORITY-F, we can effectively improve the precision by over 3.34% from 74.42%~75.61% to 78.95%, which further demonstrates the enhanced ensemble pruning performance by consolidating the majority of focal diversity metrics.

**Cora.** We also evaluate and compare our focal diversity based hierarchical pruning approach on Cora, a graph dataset. Table 12 shows the results with  $\beta = 10\%$  and  $S_d = 4$ . All focal diversity based hierarchical pruning methods achieve over 84.93% precision in identifying high quality sub-ensembles with the same or improved ensemble accuracy over 87.90% (the reference accuracy of

Table 12. Hierarchical Pruning with  $S_d=4$  on Cora

Methods	F-CK	F-BD	F-KW	F-GD	MAJORITY-F
Precision (%)	<b>91.67</b>	86.30	86.30	84.93	87.14
Recall (%)	64.71	61.76	61.76	60.78	59.80
Cost Reduction	60%	60%	60%	60%	60%

the entire ensemble for Cora). Interestingly, the F-CK pruning for Cora achieves the best ensemble pruning performance in terms of the highest precision of 91.67% and highest recall of 64.71%, indicating the diverse utilities of different focal diversity pruning methods for different datasets. Therefore, the focal diversity consensus voting based ensemble pruning is beneficial for delivering more stable and consistent ensemble pruning efficiency in terms of precision and robustness.

Table 13. Hierarchical Pruning with  $S_d=3$  on MNIST

Methods	F-CK	F-BD	F-KW	F-GD	MAJORITY-F
Precision (%)	<b>100</b>	75	75	75	75
Recall (%)	60	60	60	60	60
Cost Reduction	57%	57%	57%	57%	57%

**MNIST.** We then use popular machine learning models, such as KNN, SVM and Logistic Regression in Table 4 to evaluate the effectiveness of our focal diversity based hierarchical pruning methods on MNIST. Table 13 shows the results with  $\beta = 50\%$  and  $S_d = 3$ . We highlight two interesting observations. *First*, all focal diversity based pruning methods achieved over 75% precision in identifying sub-ensembles with the same or improved ensemble accuracy over 96.36%, which is the reference accuracy of the entire ensemble for MNIST. It demonstrates that our focal diversity based hierarchical pruning methods are also effective on representative machine learning models. *Second*, among the Top-3 selected ensemble teams by F-GD or F-CK, the ensemble team 346 (RBF SVM, Random Forest and Neural Network), achieved the highest accuracy of 97.04%, significantly outperforming the entire ensemble of 7 machine learning models with 96.36% accuracy and reducing the ensemble execution cost by over 57%. Moreover, given the Random Forest is already an ensemble of decision trees, it shows that integrating ensemble models, such as the Random Forest, in an ensemble team can further improve the overall predictive performance.

#### 5.4 Focal vs. Baseline Diversity in Hierarchical Pruning

We next compare our focal diversity metrics with baseline diversity metrics by using the same hierarchical pruning algorithm. Figure 5 shows the comparison of our focal F-GD diversity (Figure 5a) and the baseline GD diversity (Figure 5b) in hierarchical pruning for the ensemble teams of size  $S = 4$ . The black dots mark the ensembles that are pruned out by the hierarchical pruning while the red ones represent the remaining selected ensembles. The reference ensemble accuracy 96.33% of the entire deep ensemble on CIFAR-10 is marked by the horizontal black dashed line. Overall, the hierarchical pruning with our focal F-GD diversity metric achieved much better performance than the baseline GD diversity metric. There are two primary reasons behind this observation. *First*, our focal diversity metrics can better capture the failure independence of member networks of a deep ensemble than baseline diversity metrics. Therefore, when pruning out a low focal diversity branch, such as in Figure 3d with  $S = 3$ , most ensembles of a larger size with low diversity (with low F-GD scores) will also be pruned out in Figure 3e (same as Figure 5a) with  $S = 4$ . *Second*, focal diversity

metrics are more effectively correlated to ensemble accuracy than baseline diversity metrics. Therefore, by pruning out low diversity ensembles, our focal diversity metrics can successfully identify high performance ensemble teams with high ensemble accuracy and low ensemble execution cost.

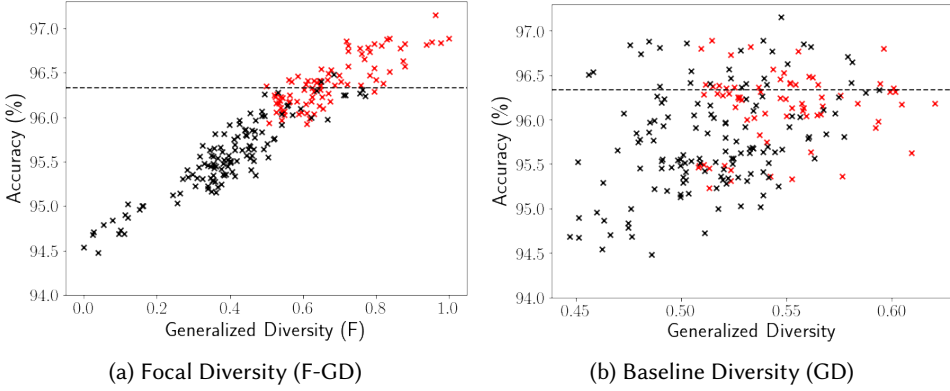


Fig. 5. Hierarchical Pruning with F-GD and GD for Ensemble Teams of  $S = 4$  (CIFAR-10,  $\beta = 10\%$ ,  $S_d = 5$ )

Table 14. Comparing Top-3 Ensemble Teams of  $S=4$  by F-GD and GD (CIFAR-10,  $\beta = 10\%$ ,  $S_d = 5$ ) with the 96.33% accuracy of the entire ensemble of 10 models shown in Table 3

Method	F-GD			GD		
Ensemble Team	0235	0234	0123	0268	0456	1345
Ensemble Acc (%)	96.89	96.84	97.15	96.18	95.63	96.17
Acc Improv (%) (Over 96.33%)	<b>0.56</b>	<b>0.51</b>	<b>0.82</b>	-0.15	-0.70	-0.16

Table 15. Comparing Top-3 Ensemble Teams of  $S=5$  by F-GD and GD (ImageNet,  $\beta = 10\%$ ,  $S_d = 5$ ) with the 96.33% accuracy of the entire ensemble of 10 models shown in Table 3

Method	F-GD			GD		
Ensemble Team	12345	23459	23458	03478	02356	01257
Ensemble Acc (%)	80.77	80.50	80.44	78.53	79.25	79.19
Acc Improv (%) (Over 79.82%)	<b>0.95</b>	<b>0.68</b>	<b>0.62</b>	-1.29	-0.57	-0.63

Table 14 lists the Top-3 sub-ensembles selected by F-GD and GD in Figure 5. All the Top-3 sub-ensembles selected by our hierarchical pruning using focal diversity F-GD improve the reference accuracy of 96.33% achieved by the entire ensemble on CIFAR-10. In comparison, the Top-3 sub-ensembles chosen by the baseline GD diversity all result in ensemble accuracy lower than 96.33%, the reference accuracy of the entire ensemble on CIFAR-10, failing to meet the ensemble pruning objective. Similar observations can be found for the ImageNet dataset in Table 15.

The empirical results on all four benchmark datasets demonstrate that our focal diversity based hierarchical pruning framework and algorithms are effective in identifying and selecting space

and time efficient sub-ensembles from a given large ensemble team, while offering competitive ensemble accuracy.

Table 16. Execution Time (s) Comparison of Baseline and Hierarchical Ensemble Pruning

Execution Time (s)		CIFAR-10	ImageNet
Baseline (No Pruning)	BD	0.90	79.08
	GD	0.84	75.83
	F-BD	2.01	97.26
	F-GD	2.01	93.51
Hierarchical Pruning	F-BD	<b>0.63</b>	<b>21.89</b>
	F-GD	<b>0.62</b>	<b>21.79</b>

### 5.5 Execution Time Comparison

We show the execution time of our hierarchical ensemble pruning based on F-BD and F-GD metrics in Table 16 for CIFAR-10 and ImageNet. The baseline is no pruning, where a diversity score, such as GD and F-GD, is computed for each possible ensemble team. We highlight three interesting observations. *First*, compared to the baseline focal diversity metrics, F-BD and F-GD, our hierarchical pruning significantly accelerates the process of identifying high-quality sub-ensembles by 3.2~4.4 $\times$ , substantially improving the ensemble pruning efficiency. Table 17 presents the execution time breakdown for each team size  $S$  from 2 to 5 using the baseline no pruning and hierarchical ensemble pruning. We list the number of all sub-ensemble teams for each team size in the 2nd column (#All Teams). For the baseline no pruning, we show the F-GD diversity computation time for all sub-ensemble teams in the 3rd column for CIFAR-10 and 8th column for ImageNet. For the hierarchical pruning approach, we list the execution time for calculating F-GD scores (4th column for CIFAR-10 and 9th column for ImageNet), the number of teams that participate in the diversity computation (5th column for CIFAR-10 and 10th column for ImageNet), and the reduction in both execution time and #Teams (6th & 7th columns for CIFAR-10 and 11th & 12th columns for ImageNet). Our hierarchical pruning approach will examine all the 45 smallest ensembles without #Teams reduction for  $S=2$ . The slight execution time reduction (0.05s, 3.65%) on ImageNet for  $S=2$  can be attributed to the random measurement errors. For both CIFAR-10 and ImageNet, the execution time reduction and #Teams reduction by our hierarchical pruning closely match with each other, such as 89.95% time reduction and 89.29% of the teams pruned out for  $S=5$  on ImageNet. This indicates that the overall performance improvement (3.2~4.4 $\times$ ) of our hierarchical pruning originates from the pruned teams that will not participate in diversity computation and thus reduce the execution time. *Second*, for the baseline without pruning, our focal diversity metrics, F-BD and F-GD, take longer execution time than BD and GD, which is due to the computation of focal diversity scores involving two steps for computing the focal negative correlation scores and aggregating the  $S$  focal negative correlation scores for each sub-ensemble of size  $S$ . *Third*, compared to the baseline diversity metrics, BD and GD, our F-BD and F-GD powered hierarchical pruning can reduce the execution time by 26%~72%, further demonstrating the efficiency of our hierarchical pruning approach.

Compared with the entire ensemble of all 10 base models, our focal diversity based hierarchical pruning approach can effectively reduce ensemble execution costs in terms of #Parameters (storage cost), GFLOPs (computing cost), and inference latency. Table 18 presents the ensemble execution costs for the entire ensemble and 3 Top-1 ensembles for  $S = 3, 4, 5$  identified by our F-GD powered hierarchical pruning approach. These three sub-ensembles all achieve higher ensemble accuracy than the large 10-model ensemble with much smaller team sizes, significantly improving the

Table 17. Execution Time (s) Breakdown Comparison of Baseline and Hierarchical Ensemble Pruning (F-GD)

Team Size	Dataset	CIFAR-10					ImageNet				
		No Pruning	Hierarchical Pruning				No Pruning	Hierarchical Pruning			
			Time (s)	Time (s)	#Teams	Time Reduction (%)		#Teams	Time Reduction (%)	Time (s)	Time (s)
2	45	0.06	0.06	45	<b>0</b>	<b>0</b>	1.37	1.32	45	<b>3.65</b>	<b>0</b>
3	120	0.26	0.18	85	<b>30.77</b>	<b>29.17</b>	9.17	5.82	82	<b>36.53</b>	<b>31.67</b>
4	210	0.63	0.23	86	<b>63.49</b>	<b>59.05</b>	28.34	9.78	78	<b>65.49</b>	<b>62.86</b>
5	252	1.03	0.15	37	<b>85.44</b>	<b>85.32</b>	52.94	5.32	27	<b>89.95</b>	<b>89.29</b>

Table 18. Ensemble Execution Cost Reduction by Hierarchical Ensemble Pruning (ImageNet,  $\beta = 5\%$ ,  $S_d = 5$ )

Ensemble Team		Ensemble Acc (%)	Execution Costs			
			#Params (M)	GFLOPs	Inference Time (ms)	
Baseline		0123456789	79.82	481.73	149.94	456.00
Top-1 teams for each size S	S=3	245	80.42 (+0.60)	92.64 (-81%)	58.50 (-61%)	153.69 (-66%)
	S=4	2345	80.70 (+0.88)	117.67 (-76%)	67.04 (-55%)	160.93 (65%)
	S=5	12345	<b>80.77 (+0.95)</b>	125.65 (-74%)	72.80 (-51%)	208.49 (-54%)

ensemble execution efficiency, i.e., cutting down the storage cost by 74%~81%, the computing cost by 51%~61%, and inference latency by 54%~66%.

## 6 CONCLUSION

This paper presents our hierarchical ensemble pruning approach powered by the focal diversity metrics. The hierarchical pruning can effectively identify high quality sub-ensembles with a significantly smaller team size and the same or better ensemble accuracy than the entire ensemble team. We made three original contributions. *First*, we optimize ensemble diversity measurements by using focal diversity metrics to accurately capture the failure independence among member networks of a deep ensemble, which closely correlates to ensemble predictive performance and provides effective guidance in ensemble pruning. *Second*, we introduce a hierarchical ensemble pruning algorithm, powered by our focal diversity metrics, which iteratively identifies high quality sub-ensembles and preemptively prunes out the member networks of low diversity. *Third*, we provide a systematic ensemble pruning approach (HQ), which consists of the focal ensemble diversity metric, hierarchical ensemble pruning algorithm and focal diversity consensus voting method. We conducted comprehensive experimental evaluations on four benchmark datasets, CIFAR-10, ImageNet, Cora and MNIST, which demonstrates that our HQ approach can efficiently prune large ensemble teams and obtain high quality sub-ensembles with enhanced ensemble accuracy and significantly reduced execution cost over the entire large ensemble team. Deep ensembles have been widely used in many domain-specific applications, including intelligent medical care [1, 11], intelligent transportation [22, 26], and intelligent manufacturing [12, 14]. One of our future works will focus on applying and optimizing our hierarchical ensemble pruning to enhance the deep ensemble efficiency for these domain-specific applications.

## ACKNOWLEDGMENTS

We acknowledge the partial support from the NSF CISE grants 1564097, 2038029, 2302720, and 2312758, an IBM Faculty Award, and a CISCO Edge AI grant.

## REFERENCES

- [1] Lerina Aversano, Mario Luca Bernardi, Marta Cimitile, and Riccardo Pecori. 2021. Deep neural networks ensemble to detect COVID-19 from CT scans. *Pattern Recognition* 120 (2021), 108135. <https://doi.org/10.1016/j.patcog.2021.108135>
- [2] Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer, and W.Philip Kegelmeyer. 2005. Ensemble diversity measures and their application to thinning. *Information Fusion* 6, 1 (2005), 49–62. <https://doi.org/10.1016/j.inffus.2004.04.005> Diversity in Multiple Classifier Systems.
- [3] Yijun Bian, Yijun Wang, Yaqiang Yao, and Huanhuan Chen. 2020. Ensemble Pruning Based on Objection Maximization With a General Distributed Framework. *IEEE Transactions on Neural Networks and Learning Systems* 31, 9 (2020), 3766–3774. <https://doi.org/10.1109/TNNLS.2019.2945116>
- [4] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.
- [5] Leo Breiman. 2001. Random Forests. In *Machine Learning*. 5–32.
- [6] Leo Breiman et al. 1998. Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics* 26, 3 (1998), 801–849.
- [7] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. Ensemble Selection from Libraries of Models. In *Proceedings of the Twenty-First International Conference on Machine Learning (Banff, Alberta, Canada) (ICML '04)*. Association for Computing Machinery, New York, NY, USA, 18. <https://doi.org/10.1145/1015330.1015432>
- [8] Ka-Ho Chow and Ling Liu. 2021. Robust Object Detection Fusion Against Deception. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (Virtual Event, Singapore) (KDD '21)*. Association for Computing Machinery, New York, NY, USA, 2703–2713. <https://doi.org/10.1145/3447548.3467121>
- [9] Ka-Ho Chow, Wenqi Wei, Yanzhao Wu, and Ling Liu. 2019. Denoising and Verification Cross-Layer Ensemble Against Black-box Adversarial Attacks. In *2019 IEEE International Conference on Big Data (Big Data)*. 1282–1291. <https://doi.org/10.1109/BigData47090.2019.9006090>
- [10] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. 2020. Deep Ensembles: A Loss Landscape Perspective. arXiv:1912.02757 [stat.ML]
- [11] Shenda Hong, Yanbo Xu, Alind Khare, Satria Priambada, Kevin Maher, Alaa Aljiffry, Jimeng Sun, and Alexey Tumanov. 2020. HOLMES: Health OnLine Model Ensemble Serving for Deep Learning Models in Intensive Care Units. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Virtual Event, CA, USA) (KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1614–1624. <https://doi.org/10.1145/3394486.3403212>
- [12] Chia-Yu Hsu and Ju-Chien Chien. 2022. Ensemble convolutional neural networks with weighted majority for wafer bin map pattern classification. *Journal of Intelligent Manufacturing* 33, 3 (2022), 831–844. <https://doi.org/10.1007/s10845-020-01687-7>
- [13] Cheng Ju, Aurélien Bibaut, and Mark Laan. 2017. The Relative Performance of Ensemble Methods with Deep Convolutional Neural Networks for Image Classification. *Journal of Applied Statistics* 45 (04 2017).
- [14] Myeongso Kim, Minyoung Lee, Minjeong An, and Hongchul Lee. 2020. Effective automatic defect classification process based on CNN with stacking ensemble model for TFT-LCD panel. *Journal of Intelligent Manufacturing* 31, 5 (2020), 1165–1174. <https://doi.org/10.1007/s10845-019-01502-y>
- [15] Ron Kohavi and David Wolpert. 1996. Bias plus Variance Decomposition for Zero-One Loss Functions. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning (Bari, Italy) (ICML '96)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 275–283.
- [16] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report. University of Toronto.
- [17] Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* 51, 2 (May 2003), 181–207. <https://doi.org/10.1023/A:1022859003006>
- [18] A. Lazarevic and Z. Obradovic. 2001. Effective pruning of neural network classifier ensembles. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, Vol. 2. 796–801 vol.2.
- [19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (Nov 1998), 2278–2324. <https://doi.org/10.1109/5.726791>
- [20] L. Liu, W. Wei, K. Chow, M. Loper, E. Gursoy, S. Truex, and Y. Wu. 2019. Deep Neural Network Ensembles Against Deception: Ensemble Diversity, Accuracy and Robustness. In *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. 274–282.
- [21] Tie-Yan Liu. 2011. *Learning to rank for information retrieval*. Springer.
- [22] Yang Liu, Zhiyuan Liu, Cheng Lyu, and Jieping Ye. 2020. Attention-Based Deep Ensemble Net for Large-Scale Online Taxi-Hailing Demand Prediction. *IEEE Transactions on Intelligent Transportation Systems* 21, 11 (2020), 4798–4807. <https://doi.org/10.1109/TITS.2019.2947145>

- [23] Qing Lu and Lise Getoor. 2003. Link-Based Classification. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning* (Washington, DC, USA) (ICML '03). AAAI Press, 496–503.
- [24] G. Martínez-Muñoz, D. Hernández-Lobato, and A. Suárez. 2009. An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 2 (2009), 245–259.
- [25] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282. <https://europepmc.org/articles/PMC3900052>
- [26] Kyushik Min, Dongchan Kim, Jongwon Park, and Kunsoo Huh. 2019. RNN-Based Path Prediction of Obstacle Vehicles With Deep Ensemble. *IEEE Transactions on Vehicular Technology* 68, 10 (2019), 10252–10256. <https://doi.org/10.1109/TVT.2019.2933232>
- [27] D. Partridge and W. Krzanowski. 1997. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology* 39, 10 (1997), 707 – 717. [https://doi.org/10.1016/S0950-5849\(97\)00023-2](https://doi.org/10.1016/S0950-5849(97)00023-2)
- [28] Remigijus Paulavičius and Julius Žilinskas. 2006. Analysis of different norms and corresponding Lipschitz constants for global optimization. *Ukio Technologinis ir Ekonominis Vystymas* 12, 4 (2006), 301–306.
- [29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- [30] David B. Skalak. 1996. 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*. 120–125.
- [31] E Ke Tang, Ponnuthurai N Suganthan, and Xin Yao. 2006. An analysis of diversity measures. *Machine learning* 65, 1 (2006), 247–271.
- [32] Grigorios Tsoumakas, Ioannis Partalas, and Ioannis Vlahavas. 2009. *An Ensemble Pruning Primer*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–13.
- [33] KAGAN TUMER and JOYDEEP GHOSH. 1996. Error Correlation and Error Reduction in Ensemble Classifiers. *Connection Science* 8, 3-4 (1996), 385–404. <https://doi.org/10.1080/095400996116839>
- [34] Wenqi Wei and Ling Liu. 2021. Robust Deep Learning Ensemble Against Deception. *IEEE Transactions on Dependable and Secure Computing* 18, 4 (2021), 1513–1527. <https://doi.org/10.1109/TDSC.2020.3024660>
- [35] W. Wei, L. Liu, M. Loper, K. Chow, E. Gursoy, S. Truex, and Y. Wu. 2020. Cross-Layer Strategic Ensemble Defense Against Adversarial Examples. In *2020 International Conference on Computing, Networking and Communications (ICNC)*. 456–460.
- [36] Yanzhao Wu, Ka-Ho Chow, Wenqi Wei, and Ling Liu. 2023. Exploring Model Learning Heterogeneity for Boosting Ensemble Robustness. In *2023 IEEE International Conference on Data Mining (ICDM)*.
- [37] Yanzhao Wu and Ling Liu. 2021. Boosting Deep Ensemble Performance with Hierarchical Pruning. In *2021 IEEE International Conference on Data Mining (ICDM)*. 1433–1438. <https://doi.org/10.1109/ICDM51629.2021.00184>
- [38] Yanzhao Wu, Ling Liu, Zhongwei Xie, Juhyun Bae, Ka-Ho Chow, and Wenqi Wei. 2020. Promoting High Diversity Ensemble Learning with EnsembleBench. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*. 208–217. <https://doi.org/10.1109/CogMI50398.2020.00034>
- [39] Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka-Ho Chow, and Wenqi Wei. 2021. Boosting Ensemble Accuracy by Revisiting Ensemble Diversity Metrics. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16464–16472. <https://doi.org/10.1109/CVPR46437.2021.01620>
- [40] Xu-Cheng Yin, Chun Yang, and Hong-Wei Hao. 2014. Learning to Diversify via Weighted Kernels for Classifier Ensemble. arXiv:1406.1167 [cs.LG]