

Generalized Unified Decomposition of Ensemble Loss

Remco R. Bouckaert

COMPUTER SCIENCE DEPARTMENT
UNIVERSITY OF WAIKATO
Xtal Mountain Information Technology
New Zealand

REMCO@CS.WAIKATO.AC.NZ, RRB@XM.CO.NZ

Michael Goebel

Pat Riddle

Department of Computer Science
University of Auckland
New Zealand

MGOEBEL@CS.AUCKLAND.AC.NZ

PAT@CS.AUCKLAND.AC.NZ

Editor: EDITOR NAME HERE

Abstract

Goebel et al. (7) presented a unified decomposition of ensemble loss for explaining ensemble performance. They considered democratic voting schemes with uniform weights, where the various base classifiers each can vote for a single class once only. In this article, we generalize their decomposition to cover weighted, probabilistic voting schemes and non-uniform (progressive) voting schemes. Empirical results suggest that democratic voting schemes can be outperformed by probabilistic and progressive voting schemes. This makes the generalization worth exploring and we show how to use the generalization to analyze ensemble loss.

Keywords: ensemble, loss decomposition, diversity, voting scheme

1. Introduction

Ensemble classifiers like bagging (2), boosting (6) bumping (13), stacking (15) and arcing (3) are popular items for research. However, they are not well understood from a theoretical, quantitative perspective. A number of theories have been proposed for decomposing the loss into bias and variance (2; 3; 5; 9), but none of these decompositions are very satisfactory in explaining the expected loss of an ensemble for a given dataset.

This article is inspired by Goebel et al. (7), hereafter referred to as UDEL to reflect the title **Unified Decomposition of Ensemble Loss**. They decomposed the loss of a classifier ensemble into the mean loss of the individual ensemble members and a loss term D which is a measure of the diversity of the ensemble members. However, they considered only ensembles where the various base classifiers each can vote for a single class once only. Here a G UDEL, a generalized UDEL is presented. This generalized decomposition additionally applies to ensembles with weighted votes, as well as to ensembles where the member classifiers output a probability distribution instead of a single vote. We also investigate several other voting schemes which are covered neither by UDEL or by G UDEL. Experiments show that simple democratic voting schemes can sometimes be outperformed by other voting schemes (1).

The rest of this paper is organized as follows: In Section 2 we introduce some notation and formally define the various voting schemes considered here. Sections 3 and 4 introduce the generalized definitions for mean member loss and diversity, and present the generalized unified decomposition of ensemble loss for binary classifiers. We proceed with some experiments in Section 5 and finish with conclusions, open questions, and directions for further research in Section 6.

2. Voting schemes

Given input space $\mathbf{X} = \mathbf{X}_1 \times \dots \times \mathbf{X}_v$, a set of class labels $Y = \{Y_1, \dots, Y_n\}$, and an unknown but stationary probability distribution P over $\mathbf{X} \times Y$, a learner is usually given a finite sample $D = \{\langle \mathbf{x}_1, y_1 \rangle, \dots, \langle \mathbf{x}_m, y_m \rangle\}$ drawn from $\mathbf{X} \times Y$ according to P and is required to produce a classifier c . Given a (previously unseen) test example $\mathbf{x} = \langle x_1, \dots, x_v \rangle$, this classifier produces a prediction $\hat{y}_c(\mathbf{x})$.

The performance of models (and therefore implicitly the learners that produced those models) is measured using loss functions: A loss function $l : Y \times Y \rightarrow \mathfrak{R}$ measures the cost of making the prediction \hat{y} when the true value is y . For the case where Y is a set of class labels $Y = \{Y_1, \dots, Y_n\}$, the most commonly used loss function is the zero-one-loss ($l_{01}(\hat{y}, y) := 0$ iff $\hat{y} = y$; $l_{01}(\hat{y}, y) := 1$ otherwise). The goal of a learner should be to produce a model with the smallest possible expected loss; i.e., a model which minimizes the average loss over examples drawn independently from $\mathbf{X} \times Y$ according to P .

Some classifiers, rather than producing a prediction $\hat{y}_c(\mathbf{x})$ directly, output instead a probability distribution $\hat{P}_c(Y|\mathbf{x})$ over the set of class labels Y , such that, for any $y \in Y$, $\hat{P}_c(y|\mathbf{x})$ reflects the degree of the classifier's internal belief that the test example to be \mathbf{x} belongs to class y . In those cases, the classifier prediction $\hat{y}_c(\mathbf{x})$ is taken to be

$$\hat{y}_c(\mathbf{x}) = \arg \min_{y' \in Y} \int_{y'' \in Y} \hat{P}_c(y''|\mathbf{x}) l(y'', y') dy''. \quad (1)$$

In recent years, there has been considerable interest in learners that produce models of small expected loss by generating and aggregating multiple individual models (4; 11; 12). The resulting model C (hereafter referred to as an *ensemble*) often has smaller expected loss than any of the k individual models c_i ($i = 1, \dots, k$), hereafter referred to as *member models* or simply *members*, making up the ensemble. It is well known (4; 8; 10) that a good ensemble is one whose members are both accurate and diverse. For the remainder of this article, the prediction of an individual ensemble member $c \in C$ will be denoted by $\hat{y}_c(\mathbf{x})$, whereas the prediction of the ensemble classifier will be denoted by $\hat{y}(\mathbf{x})$.

A typical ensemble algorithm under 0-1 loss takes the predictions of the base classifiers giving the prediction

$$\hat{y}(\mathbf{x}) := \arg \max_{y \in Y} \sum_{c \in C} w_c I(y = \hat{y}_c(\mathbf{x})) \quad (2)$$

where $I(\cdot)$ is the indicator function ($I(y = \hat{y}_c) = 1$ iff \hat{y}_c equals y , and 0 otherwise). When two or more classes get the same vote, one is chose at random. This voting scheme is called *democratic voting* also known as *majority vote*. The base classifiers predictions are possibly weighted with a weight w_c . Throughout this article, we assume the weights are normalized to 1, i.e., $\sum_{c \in C} w_c = 1$.

```

learn(data set  $\mathbf{D}$ , ensemble size  $k$ , sample size  $s$ ):
  for  $i = 1$  to  $k$  [
     $D_{train} =$  select bootstrap sample from  $D$  of size  $s$ 
     $c_i =$  learn base classifier from  $D_{train}$ 
  ]
  return  $c_1, \dots, c_k$ 

classify( $\mathbf{x}$ ):
  return (voting scheme formula (2), (3), (4), (5) or (6) goes here)
    
```

Figure 1: The Bagging algorithm

Notice that there are two orthogonal dimensions along which an ensemble voting scheme can lie. The first dimension is democratic/probabilistic and it specifies whether the classifiers return a vote for a single class or a probabilistic distribution over the set of classes. The second dimension is uniform/non-uniform and it determines whether each classifier gets a single vote or the classifiers each have an associated weight.

A lot of commonly used base classifiers calculate a probability distribution $\widehat{P}_c(y|\mathbf{x})$ over y and pick the class with the highest probability as their prediction, for instance, decision trees, naive Bayes, and Bayesian network variants. So, instead of taking the class with the highest probability as a vote, the vote can be weighed with the probabilities of the base classifiers, giving

$$\widehat{y}(\mathbf{x}) := \arg \max_{y \in Y} \sum_{c \in C} w_c \widehat{P}_c(y|\mathbf{x}) \quad (3)$$

where $\widehat{P}_c(y|\mathbf{x})$ is the probability that classifier c assigns to class y given observation \mathbf{x} . This voting scheme will be called *probabilistic voting*.

Alternatively, it can be argued that the model with the highest confidence in its ability to classify should be making the decision. This *aristocratic voting* scheme returns the class

$$\widehat{y}(\mathbf{x}) := \arg \max_{y \in Y} (\max_{c \in C} w_c \widehat{P}_c(y|\mathbf{x})) \quad (4)$$

Balancing aristocratic and democratic arguments, the vote can be weighed according to a convex function of \widehat{P}_c , for example quadratic or exponential. The *progressive quadratic voting* scheme results in the following classification:

$$\widehat{y}(\mathbf{x}) := \arg \max_{y \in Y} \sum_{c \in C} w_c (\widehat{P}_c(y|\mathbf{x}))^2 \quad (5)$$

and *progressive exponential voting* scheme:

$$\widehat{y}(\mathbf{x}) := \arg \max_{y \in Y} \sum_{c \in C} w_c e^{\widehat{P}_c(y|\mathbf{x})} \quad (6)$$

The standard bagging algorithm (2) uses the democratic voting scheme. It can be easily adapted to apply each of the above voting schemes by plugging in Equation (2), (3), (4), (5) or (6) as shown in Figure 1.

Ensemble				
	c_1	c_2	c_3	
$\widehat{P}_c(y = 0)$	0.55	0.60	0.10	
$\widehat{P}_c(y = 1)$	0.45	0.40	0.90	
Democratic votes				Prediction
$y = y_1$	1	1	0	2 *
$y = y_2$	0	0	1	1
Probabilistic votes				Prediction
$y = y_1$	0.55	0.60	0.10	1.25
$y = y_2$	0.45	0.40	0.90	1.75 *
Aristocratic votes				Prediction
$y = y_1$	0	0.60	0	0.6
$y = y_2$	0	0	0.90	0.9 *
Progr. square votes				Prediction
$y = y_1$	0.30	0.36	0.01	0.67
$y = y_2$	0.20	0.16	0.81	1.17 *
Progr. exp. votes				Prediction
$y = y_1$	1.73	1.82	1.10	4.66
$y = y_2$	1.57	1.49	2.46	5.52 *

Table 1: Illustration of voting schemes.

Table 1 shows an example that illustrates the various voting schemes for an ensemble with three probabilistic classifiers $C = \{c_1, c_2, c_3\}$ and binary class $Y = \{y_1, y_2\}$. The democratic scheme classifies y as y_1 because two of the three member classifiers have more confidence in class y_1 than in class y_2 . However, the probabilistic voting scheme classifies y as y_2 because the two classifiers c_1 and c_2 having more confidence in class y_1 do so with only a very small probability, while c_3 prefers class y_2 with overwhelming confidence. The other voting schemes also prefer class y_2 as shown in Table 1.

3. Decomposition

First, we need to define some terms, highlighting the differences with UDEL wherever they occur.

The *ensemble prediction* of an ensemble C for input \mathbf{x} is denoted as $\widehat{y}(\mathbf{x})$ or \widehat{y} if the input \mathbf{x} is clear from the context. The way the ensemble prediction is calculated depends on the voting scheme applied. For example, with the democratic voting scheme, the ensemble prediction for input \mathbf{x} is calculated using Equation 2.

Note that UDEL defines the ensemble prediction in terms of a loss function, while this definition does not take this into account. However, for the probabilistic voting scheme as given by Equation 3, the ensemble prediction can be defined in terms of the loss function as shown in Equation 1. This specializes to the definition used by (7), which was $\widehat{y}(x) := \arg \min_{y \in Y} E_{c \in C} [l(y, \widehat{y}_c(x))]$, with $w_c := 1/|C|$ and $\widehat{P}_c(y'|\mathbf{x}) := 1$ iff $y' = \widehat{y}_c(\mathbf{x})$; $\widehat{P}_c(y'|\mathbf{x}) := 0$ otherwise for all $c \in C$ and $y' \in Y$.

Definition 1 *The loss of ensemble C on instance $\langle \mathbf{x}, y \rangle$ under loss function l is given by $L(\langle \mathbf{x}, y \rangle) := l(\hat{y}(\mathbf{x}), y)$. The expected loss of ensemble C on the domain $\langle \mathbf{X}, Y, P \rangle$ is given by $L := E_P[L(\langle \mathbf{x}, y \rangle)] = \int_{X \times Y} L(\langle \mathbf{x}, y \rangle) p(\langle \mathbf{x}, y \rangle) d\langle \mathbf{x}, y \rangle$.*

This does not differ from UDEL. Typically, we are interested in learning ensembles that have a low expected loss, where the expectation is taken over the instance distribution P .

Definition 2 *The mean member loss of ensemble C on instance $\langle \mathbf{x}, y \rangle$ under loss function l is given by $\bar{L}(\langle \mathbf{x}, y \rangle) := \sum_{c \in C} w_c E_{\hat{P}_c}[l(\hat{y}_c(\mathbf{x}), y)] = \sum_{c \in C} \int_{y' \in Y} w_c l(y', y) \hat{P}_c(y' | \mathbf{x}) dy'$. The expected mean member loss of ensemble C on the domain $\langle \mathbf{X}, Y, P \rangle$ is given by $\bar{L} := E_P[\bar{L}(\langle \mathbf{x}, y \rangle)] = \int_{X \times Y} \bar{L}(\langle \mathbf{x}, y \rangle) p(\langle \mathbf{x}, y \rangle) d\langle \mathbf{x}, y \rangle$.*

The mean member loss indicates how much the predictions of the individual base classifiers in the ensemble C differ from the true value of the class y . This definition differs from UDEL in that the classifier weights w_c are taken into account when taking the expectation over the set of classifiers C . This specializes to the definition in UDEL when $w_c = 1/|C|$ for all classifiers $c \in C$. Also, the expectation is taken over both C and Y while in UDEL only the expectation over C is taken. This will take into account their class probabilities instead of just their final predictions in the case where the base classifier is a distribution classifier. If the base classifier is not a distribution classifier, the definition still applies with $\hat{P}_c(y' | \mathbf{x}) = I(y' = \hat{y}_c(\mathbf{x}))$, where $I(\cdot)$ is the indicator function, i.e, $I(y' = \hat{y}_c(\mathbf{x})) := 1$ iff y' equals $\hat{y}_c(\mathbf{x})$, and 0 otherwise. In this case, the definition of mean member loss coincides with the one from UDEL, which was $\bar{L}(\langle x, y \rangle) := E_{c \in C}[l(\hat{y}_c(x), y)]$.

Definition 3 *The diversity of ensemble C on input \mathbf{x} under loss function l is given by $\bar{D}(\mathbf{x}) := \sum_{c \in C} w_c E_{\hat{P}_c}[l(\hat{y}_c(\mathbf{x}), \hat{y}(\mathbf{x}))] = \sum_{c \in C} \int_{y' \in Y} w_c l(y', \hat{y}(\mathbf{x})) \hat{P}_c(y' | \mathbf{x}) dy'$. The expected diversity of ensemble C on the domain $\langle \mathbf{X}, Y, P \rangle$ is given by $\bar{D} := E_P[\bar{D}(\mathbf{x})] = \int_{X \times Y} \bar{D}(\mathbf{x}) p(\langle \mathbf{x}, y \rangle) d\langle \mathbf{x}, y \rangle$.*

The diversity indicates how much the predictions of the individual base classifiers in the ensemble C differ from the ensemble prediction $\hat{y}(\mathbf{x})$. Again, this definition differs from UDEL in that here the classifier weights w_c are taken into account, and in that the expectation is taken over both C and Y while in UDEL only the expectation over C is taken, such as $\bar{D}(x) := E_{c \in C}[l(\hat{y}_c(x), \hat{y}(x))]$.

Figure 2 shows the relations between the various terms just defined. It shows that the loss L is a function of the real class y and the ensemble prediction \hat{y} , the expected mean member loss \bar{L} is a function of y and the various base classifiers predictions \hat{y}_c , and the expected diversity \bar{D} is a function of the various base classifiers predictions \hat{y}_c and the ensemble prediction \hat{y} .

Following UDEL, a decomposition of L is proposed as a function of \bar{L} and \bar{D} , that is,

$$L(\langle \mathbf{x}, y \rangle) = f(\bar{L}(\langle \mathbf{x}, y \rangle), \bar{D}(\mathbf{x})) \tag{7}$$

for ensembles of weighted, probabilistic classifiers. This covers ensembles voting democratically as in Equation 2 or probabilistically as in Equation 3, with or without individual classifier weights. Though it does not cover the aristocratic or progressive voting schemes

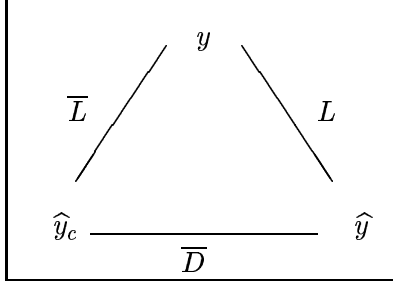


Figure 2: Intuitive relations between L , \bar{L} , \bar{D} and y , \hat{y} and \hat{y}_c .

in Equations 4, 5, and 6, the theory can be easily extended by normalizing the votes in the schemes so that the votes add to unity for each particular instance. The extensions to the theory which cover these cases are discussed shortly in Section 4.4.

In the following section, details for 0-1 loss will be given.

4. Instantiating the decomposition for 0-1 Loss

We first instantiate Equation 7 for the case where the class is binary, that is $Y = \{0, 1\}$, and the loss function l is the 0-1 loss, i.e., $l(y_1, y_2) = I(y_1 \neq y_2)$.

Lemma 1 *For 0-1 loss in two-class problems, the ensemble loss $L(\langle \mathbf{x}, y \rangle)$ can be written as*

$$L(\langle \mathbf{x}, y \rangle) = f(\bar{L}(\langle \mathbf{x}, y \rangle), \bar{D}(\mathbf{x})) = \bar{L}(\langle \mathbf{x}, y \rangle) + k(x, y)\bar{D}(\mathbf{x}) \quad (8)$$

where $k(x, y) = -1$ iff $\hat{y}(\mathbf{x}) = y$ and $k(x, y) = 1$ iff $\hat{y}(\mathbf{x}) \neq y$.

Proof. If $\hat{y}(\mathbf{x}) = y$, we have

$$\begin{aligned} L(\langle \mathbf{x}, y \rangle) &= l(\hat{y}(\mathbf{x}), y) = 0 \\ &= \sum_{c \in C} \int_{y' \in Y} w_c l(y', \hat{y}(\mathbf{x})) \hat{P}_c(y' | \mathbf{x}) dy' \\ &\quad - \sum_{c \in C} \int_{y' \in Y} w_c l(y', y) \hat{P}_c(y' | \mathbf{x}) dy' \\ &= \bar{L}(\langle \mathbf{x}, y \rangle) - \bar{D}(\mathbf{x}) \end{aligned}$$

which shows $k(x, y) = -1$ for Equation 8. If $\hat{y}(\mathbf{x}) \neq y$, we have

$$\begin{aligned} L(\langle \mathbf{x}, y \rangle) &= l(\hat{y}, y) = 1 = \sum_{c \in C} w_c \\ &= \sum_{c \in C} \int_{y' \in Y} w_c \hat{P}_c(y' | \mathbf{x}) dy' \\ &= \sum_{c \in C} \int_{y' \in Y} w_c I(y' = y) \hat{P}_c(y' | \mathbf{x}) dy' \end{aligned}$$

$$\begin{aligned}
 & + \sum_{c \in C} \int_{y' \in Y} w_c I(y' \neq y) \hat{P}_c(y' | \mathbf{x}) dy' \\
 = & \sum_{c \in C} \int_{y' \in Y} w_c I(y', \hat{y}(\mathbf{x})) \hat{P}_c(y' | \mathbf{x}) dy' \\
 & + \sum_{c \in C} \int_{y' \in Y} w_c I(y', y) \hat{P}_c(y' | \mathbf{x}) dy' \\
 = & \bar{L}(\langle \mathbf{x}, y \rangle) + \bar{D}(\mathbf{x})
 \end{aligned}$$

which shows $k(x, y) = 1$ for Equation 8. ■

This generalizes Theorem 3 from UDEL.

For a given ensemble C , let $T := \{\langle \mathbf{x}, y \rangle | \hat{y}(\mathbf{x}) = y\}$ be the set of instances which the ensemble classifies correctly, and let $F := \{\langle \mathbf{x}, y \rangle | \hat{y}(\mathbf{x}) \neq y\}$ be the set of instances which the ensemble classifies incorrectly. Then, we can define \bar{D}_T as the diversity over T and \bar{D}_F as the diversity over F . More formally, $\bar{D}_T := \int_{\langle \mathbf{x}, y \rangle \in T} \bar{D}(\mathbf{x}) p(\langle \mathbf{x}, y \rangle | \langle \mathbf{x}, y \rangle \in T) d\langle \mathbf{x}, y \rangle$ and $\bar{D}_F := \int_{\langle \mathbf{x}, y \rangle \in F} \bar{D}(\mathbf{x}) p(\langle \mathbf{x}, y \rangle | \langle \mathbf{x}, y \rangle \in F) d\langle \mathbf{x}, y \rangle$. \bar{D}_T denotes the expected ensemble diversity on correctly predicted examples, and \bar{D}_F denotes the expected ensemble diversity on incorrectly predicted examples. \bar{D}_T , \bar{D}_F , and \bar{D} will always be between 0 and 0.5. This can be easily seen for a two class dataset. If \bar{D} became more than .5 then the ensemble would predict a new class thereby making \bar{D} less than .5 again. The same argument holds for \bar{D}_F and \bar{D}_T and becomes even more obvious when there are more than two classes. The following lemmas hold:

Lemma 2 *Under 0-1 loss, the expected diversity \bar{D} can be written as*

$$\bar{D} = (1 - L)\bar{D}_T + L\bar{D}_F. \quad (9)$$

Proof.

$$\begin{aligned}
 \bar{D} & = E_P[\bar{D}(\mathbf{x}, y)] \\
 & = \int_{X \times Y} \bar{D}(\mathbf{x}) p(\langle \mathbf{x}, y \rangle) d\langle \mathbf{x}, y \rangle \\
 & = \int_{\langle \mathbf{x}, y \rangle \in T} \bar{D}(\mathbf{x}) p(\langle \mathbf{x}, y \rangle) d\langle \mathbf{x}, y \rangle + \int_{\langle \mathbf{x}, y \rangle \in F} \bar{D}(\mathbf{x}) p(\langle \mathbf{x}, y \rangle) d\langle \mathbf{x}, y \rangle \\
 & = \frac{p(\langle \mathbf{x}, y \rangle \in T)}{p(\langle \mathbf{x}, y \rangle \in T)} \int_{\langle \mathbf{x}, y \rangle \in T} \bar{D}(\mathbf{x}) p(\mathbf{x}, y \wedge \langle \mathbf{x}, y \rangle \in T) d\langle \mathbf{x}, y \rangle \\
 & \quad + \frac{p(\langle \mathbf{x}, y \rangle \in F)}{p(\langle \mathbf{x}, y \rangle \in F)} \int_{\langle \mathbf{x}, y \rangle \in F} \bar{D}(\mathbf{x}) p(\mathbf{x}, y \wedge \langle \mathbf{x}, y \rangle \in F) d\langle \mathbf{x}, y \rangle \\
 & = p(\langle \mathbf{x}, y \rangle \in T) \int_{\langle \mathbf{x}, y \rangle \in T} \bar{D}(\mathbf{x}) \frac{p(\langle \mathbf{x}, y \rangle \wedge \langle \mathbf{x}, y \rangle \in T)}{p(\langle \mathbf{x}, y \rangle \in T)} d\langle \mathbf{x}, y \rangle \\
 & \quad + p(\langle \mathbf{x}, y \rangle \in F) \int_{\langle \mathbf{x}, y \rangle \in F} \bar{D}(\mathbf{x}) \frac{p(\langle \mathbf{x}, y \rangle \wedge \langle \mathbf{x}, y \rangle \in F)}{p(\langle \mathbf{x}, y \rangle \in F)} d\langle \mathbf{x}, y \rangle
 \end{aligned}$$

$$\begin{aligned}
&= p(\langle \mathbf{x}, y \rangle \in T) \int_{\langle \mathbf{x}, y \rangle \in T} \overline{D}(\mathbf{x}) p(\mathbf{x}, y | \langle \mathbf{x}, y \rangle \in T) d\langle \mathbf{x}, y \rangle \\
&\quad + p(\langle \mathbf{x}, y \rangle \in F) \int_{\langle \mathbf{x}, y \rangle \in T} \overline{D}(\mathbf{x}) p(\mathbf{x}, y | \langle \mathbf{x}, y \rangle \in F) d\langle \mathbf{x}, y \rangle \\
&= (1 - L) \overline{D}_T + L \overline{D}_F
\end{aligned}$$

■

Note that Equation 9 is a generalization of a similar result for non-weighted, democratically voting ensembles given in UDEL.

Lemma 3 *For 0-1 loss in two-class problems, the expected ensemble loss L can be written as*

$$L = \frac{\overline{L} - \overline{D}_T}{1 - \overline{D}_T - \overline{D}_F}.$$

The proof exactly follows the proof of UDEL's Theorem 4, of which this is a generalization.

Lemma 4 *For 0-1 loss in two-class problems, the expected ensemble loss L can be written as*

$$L = \frac{\overline{L} + \overline{D} - 2\overline{D}_T}{1 - 2\overline{D}_T}. \quad (10)$$

Proof. Lemma 3 can be rewritten as

$$L = \overline{L} - (1 - L)\overline{D}_T + L\overline{D}_F. \quad (11)$$

From Lemma 2, $\overline{D} = (1 - L)\overline{D}_T + L\overline{D}_F$ can be rewritten to $L\overline{D}_F = \overline{D} - (1 - L)\overline{D}_T$. With substitution in (11), we get

$$\begin{aligned}
L &= \overline{L} - (1 - L)\overline{D}_T + \overline{D} - (1 - L)\overline{D}_T \\
&= \overline{L} + \overline{D} - 2\overline{D}_T + 2L\overline{D}_T
\end{aligned}$$

Bringing terms to one side gives

$$\begin{aligned}
L - 2L\overline{D}_T &= \overline{L} + \overline{D} - 2\overline{D}_T \\
L(1 - 2\overline{D}_T) &= \overline{L} + \overline{D} - 2\overline{D}_T
\end{aligned}$$

which can be rewritten to the statement in the lemma.

■

Lemma 5 *For 0-1 loss in two-class problems, the expected ensemble loss L can be written as*

$$L = \frac{\overline{L} - \overline{D}}{1 - 2\overline{D}_F}. \quad (12)$$

Proof. This is similar to the proof of Lemma 4, but now we rewrite $\overline{D} = (1 - L)\overline{D}_T + L\overline{D}_F$ from Lemma 2 to $(1 - L)\overline{D}_T = \overline{D} - L\overline{D}_F$. Then, substitution in (11) gives

$$\begin{aligned} L &= \overline{L} - \overline{D} + L\overline{D}_F + L\overline{D}_F \\ &= \overline{L} + \overline{D} + 2L\overline{D}_F \end{aligned}$$

Similar manipulation as above gives the required result. ■

As discussed in the following section, Lemmas 3, 4 and 5 provide better insight in the behavior of ensemble classifiers.

4.1 Relating L , \overline{L} , \overline{D} , \overline{D}_T and \overline{D}_F

There are two ways to manipulate \overline{D}_T and \overline{D}_F :

1. by keeping T and F constant. In this case, the ensemble decisions do not change, hence the 0-1 loss does not change, but the diversities \overline{D}_T and \overline{D}_F can be manipulated independently.
2. by changing the boundaries of T and F . In this case, the 0-1 loss is affected, and so are \overline{D}_T and \overline{D}_F .

Case 1: When keeping T and F constant and increasing \overline{D}_T by α , \overline{D} increases by $\alpha(1 - L)$ (by Lemma 2) and \overline{L} increases by $\alpha(1 - L)$ (because of Equation 11 and L remains constant). Lemma 4 shows that L remains constant under such change. If \overline{D}_F is increased by α , \overline{D} increases by αL (again by Lemma 2), and likewise \overline{L} decreases by αL (because of Equation 11). Lemma 5 confirms that L remains constant under such change.

Case 2: Now consider manipulating T and F by moving instances from F to T such that L decreases by α . Let M be the set of instances $\{\langle \mathbf{x}, y \rangle \in X \times Y\}$ that moved from F to T . For those instances holds $\overline{L} - \overline{D} = 0$ by Lemma 2. So, by Lemma 4, we have $L - \alpha = (\overline{L} - \overline{D} - \alpha)/(1 - 2\overline{D}_F)$, which after some manipulations gives $\overline{D}_F = L - \overline{L} + \overline{D}/2(L - \alpha)$, that is, \overline{D}_F increases. This also works the other way around; increasing \overline{D}_F helps decrease the expected ensemble loss L .

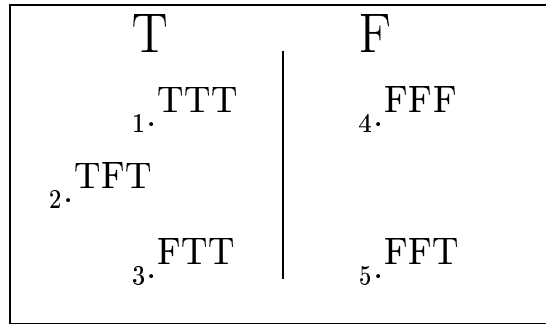


Figure 3: Case study of relations between L , \overline{L} , \overline{D} and y , \hat{y} and \hat{y}_c .

Let us look at a case study to make this a bit clearer. In Figure 3 we have three instances in the set T and two instances in the set F. We have 3 member classifiers for classifying each of these instances; their votes are given beside each instance. In case 1 we increase \overline{D}_T while keeping T and F constant; for example changing the votes at instance 1 from TTT to TTF. This will cause an increase in \overline{D} and \overline{L} but L remains constant. Likewise a change in instance 4, from FFF to FTF will cause an increase in \overline{D}_F , \overline{D} , but a decrease in \overline{L} while L remains constant. This highlights that just increasing or decreasing diversity can have no effect on the benefits derived from an ensemble.

Now let us examine case 2. If we move instance 3 from the T set to the F set by changing its votes from FTT to FTF, we will cause L to increase while \overline{D} hasn't changed. \overline{L} will also increase and \overline{D}_F could either have increased or decreased. If we move instance 5 from the F set to the T set by changing its votes from FFT to FTT, we will cause L to decrease while \overline{D} still hasn't changed. \overline{L} will also decrease and \overline{D}_T could either have increased or decreased. From this we can see, that increasing the diversity over the F set is the only way to lower the ensemble loss.

4.2 Why weighting may work

Instead of letting the classifiers in the ensemble vote uniformly ($\forall c \in C : w_c = 1/|C|$), the weights can be made proportional to the classification accuracy on the training data. This can be achieved by setting $\forall c \in C : w_c \propto 1 - L_c$, where $L_c = \int_{X \times Y} l(y, \hat{y}_c(\mathbf{x})) p(\langle \mathbf{x}, y \rangle) d(\mathbf{x}, y)$ empirically estimates the expected loss of member classifier c on the training data, with m being the number of instances in the training set. We assume that when the classification accuracy on the training data is smaller, then the classification accuracy in general will be smaller, that is, performance on the training data can be seen as a predictor on the performance in general. This seems a reasonable assumption since it underlies all classifier learning algorithms (as long as overfitting is taken into account).

So, let's look at what happens when moving weight to classifiers in the ensemble that perform well on the training data. From the definition for expected loss \overline{L} (definition 2) we see that it decreases. Also, the diversity \overline{D}_F will decrease, hence from Lemma 5 we have that the expected loss will decrease (since the numerator decreases and the denominator increases).

Obviously the effect of increased performance by weighting is not the only mechanism at work. If this were the case, just taking the classifier in the ensemble with the highest accuracy on the training data and discarding the others would give a better classifier than using the whole ensemble in uniform voting. Experimental results suggest that this is generally not true.

4.3 What diversity does for binary classes

Let's look at an example with a binary class, a set of four instances $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$, and a set of three classifiers $C = \{c_1, c_2, c_3\}$, each classifier being able to classify two instances correctly and two incorrectly. For simplicity, a democratic voting scheme is used, so that it is easy to illustrate when classifiers are correct (denoted by C in the following tables) or incorrect (denoted by I in the following tables). For the purpose of this example, it is also

\mathbf{x}	c_1	c_2	c_3	Ensemble	
\mathbf{x}_1	C	C	C	C	$L = 2/4$
\mathbf{x}_2	C	C	C	C	$\bar{L} = 6/12$
\mathbf{x}_3	I	I	I	I	$\bar{D} = 0/12$
\mathbf{x}_4	I	I	I	I	$\bar{D}_T = 0/6$
					$\bar{D}_F = 0/6$

Table 2: Example where there is no diversity ($\bar{D} = 0$).

\mathbf{x}	c_1	c_2	c_3	Ensemble	
\mathbf{x}_1	C	C	I	C	$L = 1/4$
\mathbf{x}_2	C	I	C	C	$\bar{L} = 6/12$
\mathbf{x}_3	I	C	C	C	$\bar{D} = 3/12$
\mathbf{x}_4	I	I	I	I	$\bar{D}_T = 3/9$
					$\bar{D}_F = 0/3$

Table 3: Example with high diversity and high \bar{D}_T .

\mathbf{x}	c_1	c_2	c_3	Ensemble	
\mathbf{x}_1	C	C	C	C	$L = 3/4$
\mathbf{x}_2	C	I	I	I	$\bar{L} = 6/12$
\mathbf{x}_3	I	C	I	I	$\bar{D} = 3/12$
\mathbf{x}_4	I	I	C	I	$\bar{D}_T = 0/3$
					$\bar{D}_F = 3/9$

Table 4: Example with high diversity and high \bar{D}_F .

assumed that the weights are uniform ($w_c = 1/3$), that there is no class noise ($\forall i \in \{1, \dots, 4\} : y_i = f(\mathbf{x}_i)$), and that the examples are uniformly distributed ($\forall \langle \mathbf{x}_i, y_i \rangle : P(\langle \mathbf{x}_i, y_i \rangle) = 1/4$).

If all of the classifiers' predictions are exactly the same, that is, when there is no diversity, we get a situation as shown in Table 2. Only 2 cases will be classified correctly, and no gain over the individual base classifiers is made.

Table 3 shows what happens when the diversity is increased; one more case is classified correctly now. However, Table 4 shows an example with the same diversity but with only one item classified correctly. The difference is that in Table 3 $\bar{D} = (1 - L)\bar{D}_T$ and $\bar{D}_F = 0$, while in Table 4 $\bar{D} = L\bar{D}_F$ and $\bar{D}_T = 0$.

The same effect appears when the set of classifiers is larger than 3. So, in general, we want to have a high as possible \bar{D}_T with an as low as possible \bar{D}_F .

Probabilistic voting can be interpreted as democratic voting where, instead of a single vote by a single classifier c_i , a set of classifiers $c_{i,1}, \dots, c_{i,|Y|}$ vote democratically, and a portion $\hat{P}_c(y = 0)$ of those votes for $y = 0$ while the rest votes for $y = 1$. Therefore, the same effect can be observed in probabilistically voting ensembles as well.

4.4 Auxiliaries

Other domains are possible. For instance, instead of a binary class, a multinomial class or a real-valued class can be taken. For the following cases the same results as in UDEL applies: multinomial class with 0-1 loss, real valued class with squared loss. The relevant results are Lemma 6 and Lemma 7.

Lemma 6 *Let the class be multinomial, that is, $Y = \{y_1, \dots, y_n\}$, and let the loss function be the 0-1 loss. Furthermore, let $\overline{D}_P := \int_{\langle \mathbf{x}, y \rangle \in F} (1 - \overline{L}(\mathbf{x})) p(\langle \mathbf{x}, y \rangle | \langle \mathbf{x}, y \rangle \in F) d\langle \mathbf{x}, y \rangle$ be the expected mean member accuracy on those instances that the ensemble predicts incorrectly. Then the expected ensemble loss L can be written as*

$$L = \frac{\overline{L} - \overline{D}_T}{\overline{L} - \overline{D}_T - \overline{D}_P}.$$

The proof follows exactly that of Theorem 6 from UDEL.

Lemma 7 *Let the class be real valued ($Y \subseteq \mathfrak{R}$) and the loss function l be the squared loss, that is, $l(y, y') = (y - y')^2$. Then the expected ensemble loss L can be written as*

$$L = \overline{L} - \overline{D}.$$

Lemma 7 is significant for two reasons. Firstly, it shows the decomposition commonly used for real-valued classes under squared loss is a special case of GUDEL. Secondly, it extends this decomposition to the case where the base classifiers, rather than producing a single real number as their prediction, may output a probability distribution over \mathfrak{R} instead.

The decomposition of loss (7) does not cover the aristocratic or progressive voting schemes in Equations 4, 5, and 6. The proof of Lemma 1 relies on the votes adding up to 1, which is not the case for these voting schemes. However, the theory can be easily extended by normalizing the votes in the schemes so that the votes add to unity for each particular instance.

5. Experiments

We performed some experimental evaluations to illustrate the behavior of the various voting schemes.

5.1 Experimental set up

Using Weka (14) version 3.4 and a modified version of its bagging algorithm where the voting scheme is configurable, an experiment was carried out with some of the data sets provided with Weka. The weights of the classifiers were kept uniform ($\forall c \in C : w_c = \frac{1}{|C|}$). As base algorithm, the Weka implementation of C4.5, called J4.8, was used with its default settings. The ensemble algorithm was bagging with its default settings (10 bags, 100% bagsize, seed for random number generation = 1), except the voting schemes were varied. The results are for 10 runs with 10 fold cross validation, so each algorithm is run 100 times.

Table 6 shows the voting schemes and their average accuracies. The numbers in parentheses are the standard deviations over the 10 runs. Algorithms are compared pairwise by

GENERALIZED UNIFIED DECOMPOSITION OF ENSEMBLE LOSS

Dataset	# Number of attributes			Classes		Number of Instances
	Total	Cont.	Nominal	Number	Distribution	
anneal	38	6	32	5	8/99/684/67/40	898
audiology	69	-	69	24	24 * [1..57]	226
autos	25	15	10	7	0/3/22/67/54/32/27	205
balance	4	4	0	3	288/49/288	625
breast-c	9	0	9	2	201/85	286
breast-w	9	9	0	2	458/241	699
colic	22	7	15	2	232/136	368
credit-a	15	6	9	2	307/383	690
credit-g	20	7	13	2	700/300	1,000
diabetes	8	8	0	2	500/268	768
glass	8	9	0	7	70/76/17/0/13/9/29	214
heart-c	13	6	7	5	165/138/0/0/0	303
heart-h	13	6	7	5	188/106/0/0/0	294
heart-s	13	6	7	2	150/120	270
hepatitis	19	6	13	2	32/123	155
hypothyroid	29	7	22	4	3481/194/95/2	3,772
ionosphere	34	34	-	2	126/225	351
kr-vs-kp	36	-	36	2	1669/1527	3,196
iris	4	4	0	3	50/50/50	150
labor	16	8	8	2	20/37	57
letter	16	16	0	26	26 * ca. 750	20,000
lymph	18	3	15	4	2/81/61/4	148
mushroom	22	0	22	2	4208/3916	8,124
primary	17	0	17	22	22 * [0..84]	339
segment	19	19	0	7	7 * 330	2,310
sick	29	7	22	2	3541/231	3,772
sonar	60	60	-	2	97/111	208
soybean	35	-	35	19	19 * [8..92]	683
splice	61	-	61	3	767/768/1655	3,190
vehicle	18	18	0	4	212/217/218/199	846
vote	16	0	16	2	267/168	435
vowel	13	10	3	11	11 * 90	990
waveform	40	40	-	3	1692/1653/1655	5,000
zoo	17	1	16	7	41/20/5/13/4/8/10	101

Table 5: List of datasets.

Dataset	Average accuracies				
	democratic	aristocratic	probabilistic	squared	exponential
autos	80.2 (8.95)	66.14 (9.22)	81.48 (8.66)	81.37 (8.78)	81.27 (8.88)
balance	82.51 (3.61)	77.33 (4.56)	82.68 (3.62)	82.89 (3.48)	82.78 (3.51)
breast-c	73.08 (5.44)	68.02 (6.75)	73.1 (6.06)	73.1 (6.06)	73.0 (6.05)
breast-w	96.1 (2.53)	90.5 (3.49)	96.18 (2.39)	96.18 (2.39)	96.18 (2.39)
colic	85.18 (6.02)	83.61 (6.04)	84.99 (5.99)	84.99 (5.99)	84.99 (5.99)
credit-a	86.29 (4.0)	82.83 (4.39)	86.38 (3.98)	86.38 (3.98)	86.33 (4.03)
credit-g	73.63 (3.54)	69.69 (2.53)	73.38 (3.87)	73.38 (3.87)	73.4 (3.86)
diabetes	75.39 (4.73)	69.33 (3.15)	75.23 (5.02)	75.23 (5.02)	75.23 (5.07)
glass	73.61 (8.57)	61.84 (8.15)	73.9 (9.22)	74.05 (9.43)	73.86 (9.36)
heart-c	79.83 (6.85)	72.1 (6.45)	79.86 (6.76)	79.86 (6.76)	79.83 (6.76)
heart-h	79.94 (7.31)	76.88 (7.63)	80.24 (6.92)	80.24 (6.92)	80.24 (6.92)
heart-s	81.3 (6.52)	71.48 (6.06)	80.93 (6.85)	80.93 (6.85)	80.93 (6.85)
hepatitis	81.33 (9.81)	75.99 (9.85)	82.05 (8.64)	82.05 (8.64)	82.05 (8.64)
iris	94.67 (4.83)	92.67 (5.23)	94.73 (4.86)	94.73 (4.86)	94.73 (4.86)
labor	86.0 (14.85)	81.33 (17.02)	83.67 (14.89)	83.67 (14.89)	83.67 (14.89)
letter	92.43 (0.63)	80.61 (0.98)	92.68 (0.62)	92.59 (0.63)	92.64 (0.62)
lymph	78.25 (10.3)	72.62 (8.38)	78.44 (10.29)	78.3 (10.27)	78.5 (10.23)
mushroom	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
primary	43.1 (6.97)	40.01 (7.89)	42.81 (7.04)	43.1 (7.04)	43.16 (6.97)
segment	97.22 (1.08)	93.72 (1.51)	97.25 (1.12)	97.26 (1.11)	97.26 (1.1)
vehicle	74.25 (4.0)	63.33 (4.0)	74.4 (4.06)	74.34 (3.99)	74.34 (4.01)
vote	96.23 (2.7)	96.06 (2.62)	96.25 (2.6)	96.25 (2.6)	96.25 (2.6)
vowel	88.58 (3.81)	62.32 (4.48)	89.54 (3.71)	89.35 (3.67)	89.46 (3.72)
zoo	93.51 (7.44)	91.32 (7.67)	93.7 (7.34)	93.7 (7.34)	93.7 (7.34)

Table 6: Experimental results of Bagging C4.5 using five different voting schemes.

democratic	-	0	3	4	3
aristocratic	22	-	22	22	22
probabilistic	1	0	-	1	1
squared	1	0	2	-	1
exponential	1	0	1	1	-

Table 7: Experimental results summary: Number of datasets where [column] is better than [row].

applying a double tailed t-test on a 0.05 confidence level. Table 7 gives a summary of comparisons, which shows the number of datasets on which an algorithm in a column performs better than one in a row. Finally, Table 8 shows the ranking of the voting schemes with the number of wins and losses from the second table where the score is the difference between them.

5.2 Discussion

In figure 6 and 7 the voting schemes are compared in a pairwise fashion. Using democratic voting instead of probabilistic voting performs slightly worse in 17 out of 24 datasets, whereas democratic is better than probabilistic in only 6 out of 24. In three datasets probabilistic performs significantly better than democratic (autos, letter, vowel) and performs significantly worse on one of the datasets. This indicates that there are quite a few situations where a majority of the base classifiers in the ensemble have a light preference for a certain class, but they are outvoted by a few models that have a high preference for another, namely the correct class. It suggests that the model that has the highest confidence in its classification should be listened to. However, the aristocratic voting scheme turns out to perform considerably and significantly worse than both democratic and probabilistic voting. The progressive voting schemes (both quadratic and exponential variants) perform slightly better than democratic voting. Exponential beats democratic in 16 out of 24 datasets and loses in only 6. Squared voting beats democratic in 17 out of 24 datasets and loses in only 5. The progressive voting schemes perform slightly worse over all than probabilistic voting. Comparison of the two progressive voting algorithms shows no large difference, making it hard to justify selecting one of the functions.

Table 9 shows a comparison of a single base classifier with probabilistic and democratic bagging with only 3 bags. The democratic scheme is outperformed by a single C4.5 tree, while probabilistic voting never is. So the difference in voting schemes appears even with as few as 3 bags.

These results are encouraging the search for voting schemes other than straightforward democratic voting.

Algorithm	Score	Wins	Losses	Draws
probabilistic	25	28	3	65
exponential	24	27	3	66
squared	24	28	4	65
democratic	15	25	10	61
aristocratic	-88	0	88	12

Table 8: Experimental results summary: Ranking of voting schemes.

Algorithm	Score	Wins	Losses	Draws
probabilistic (3 bags)	5	5	0	43
C4.5	0	3	3	42
democratic (3 bags)	-5	1	6	41

Table 9: Comparison of democratic and probabilistic voting schemes with single classifier.

Dataset	L
anneal	1.43±0.08
audiology	22.74±0.33
autos	18.23±1.06
balance	22.18±0.34
breastc	25.72±0.77
breastw	4.99±0.22
colic	14.84±0.18
credita	14.43±0.34
creditg	28.75±0.30
diabetes	25.51±0.45
glass	32.37±1.30
heartc	23.06±0.68
hearth	19.78±0.46
hearts	21.85±1.12
hepatitis	20.78±1.30
hypo	0.46±0.03
ionosphere	10.26±0.72
iris	5.27±0.40
krkp	0.56±0.02
labor	21.40±1.62
letter	11.97±0.05
lymph	24.16±1.32
mushroom	0.00±0.00
primary	58.61±0.85
segment	3.21±0.15
sick	1.28±0.07
sonar	26.39±1.31
soybean	8.22±0.40
splice	5.97±0.14
vehicle	27.72±0.65
vote	3.43±0.09
vowel	19.80±0.50
waveform	24.75±0.23
zoo	7.39±0.35
Mean	16.40

Table 10: Performance of base classifier.

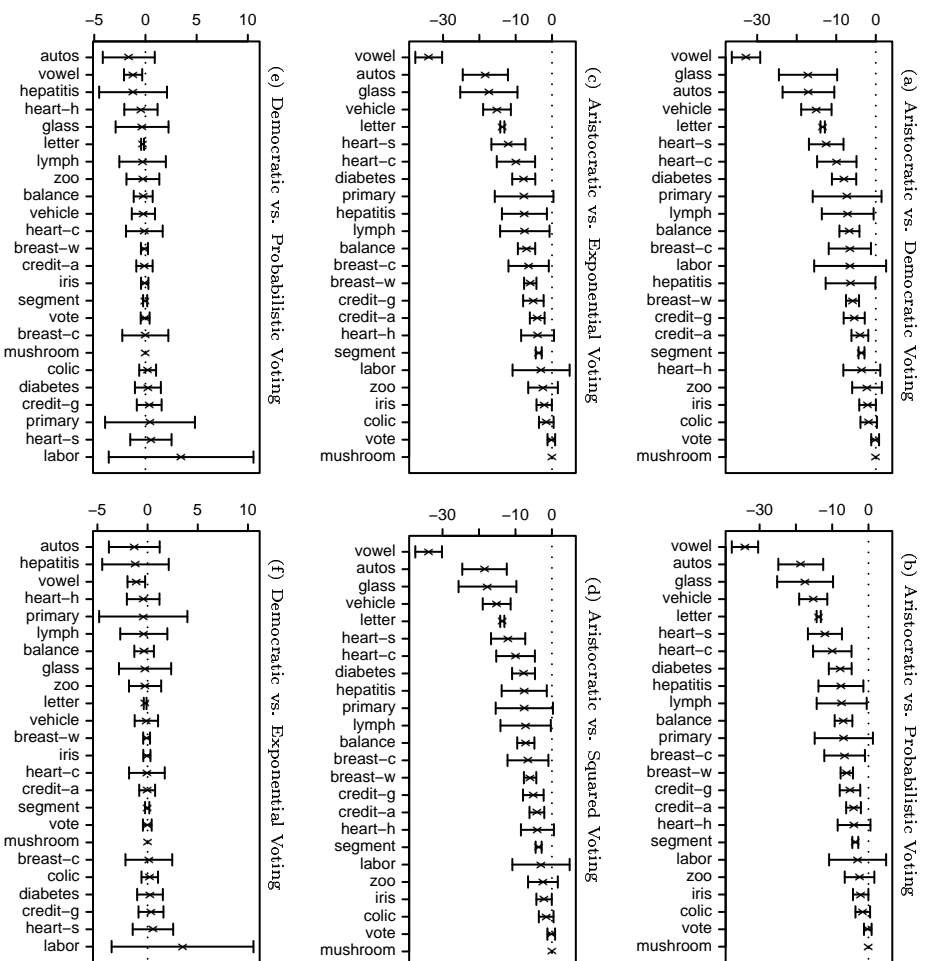


Figure 4: OLD Comparisons of voting schemes (Part 1).

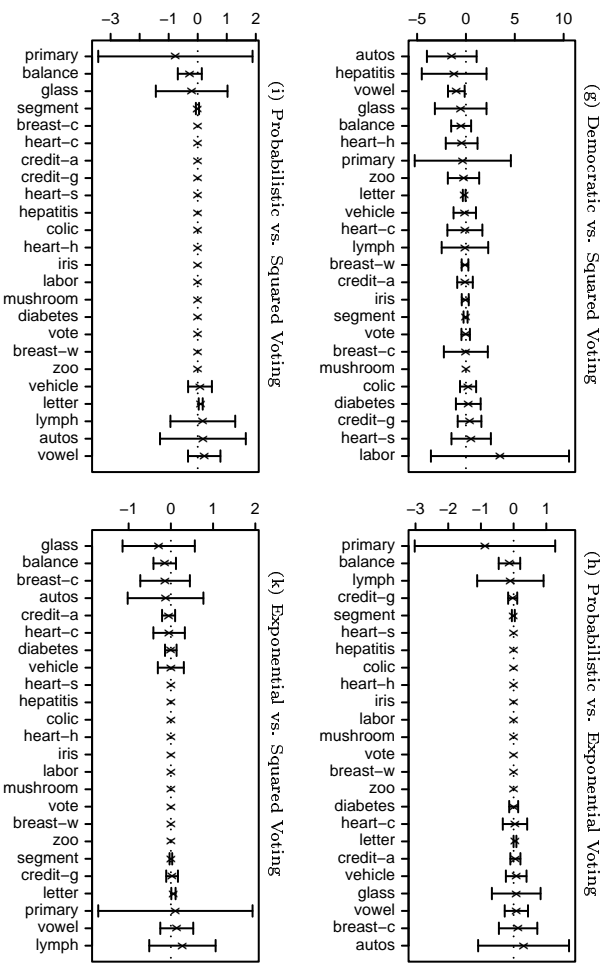


Figure 5: OLD Comparisons of voting schemes (Part 2).

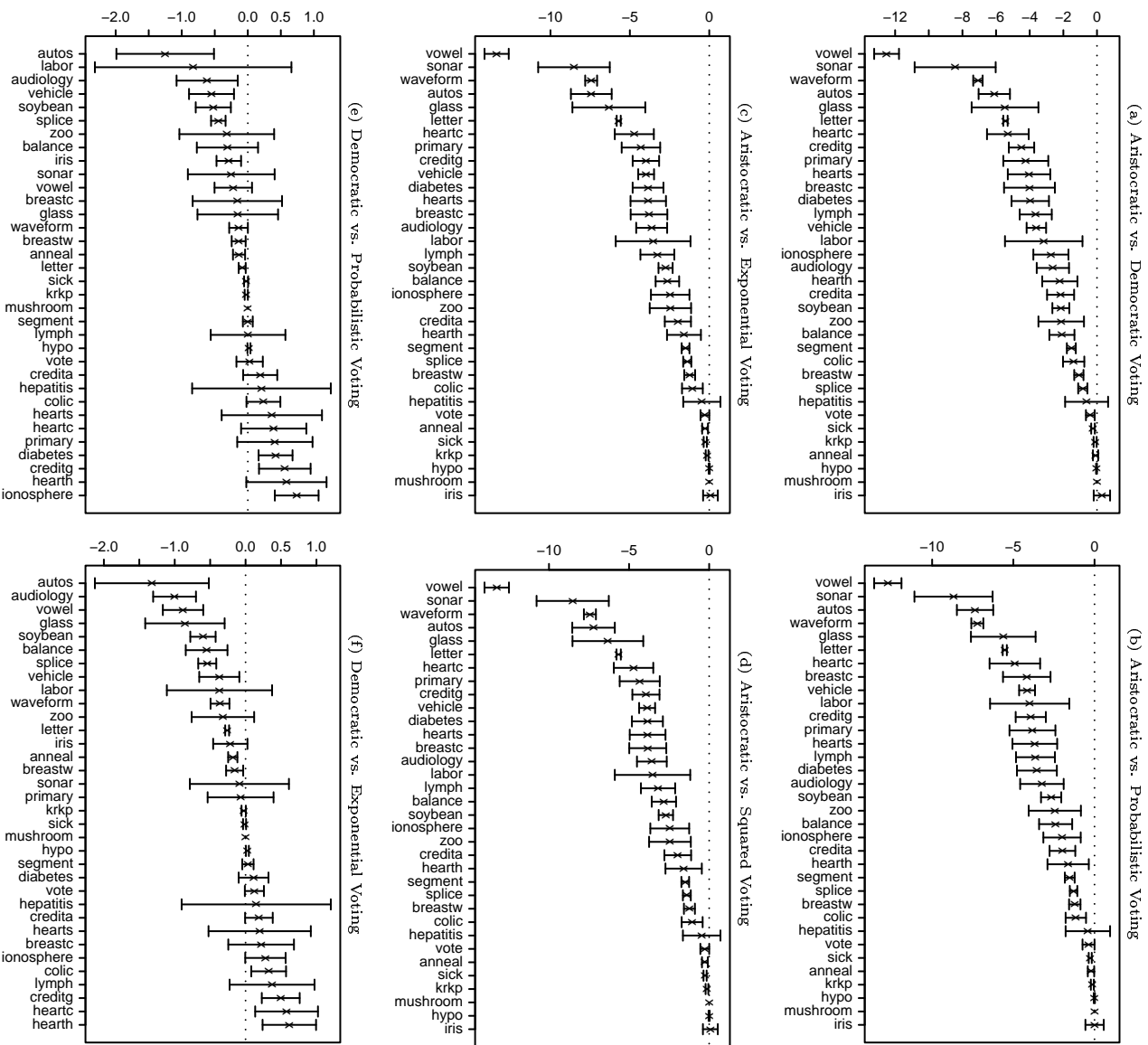


Figure 6: Comparisons of voting schemes (Part 1).

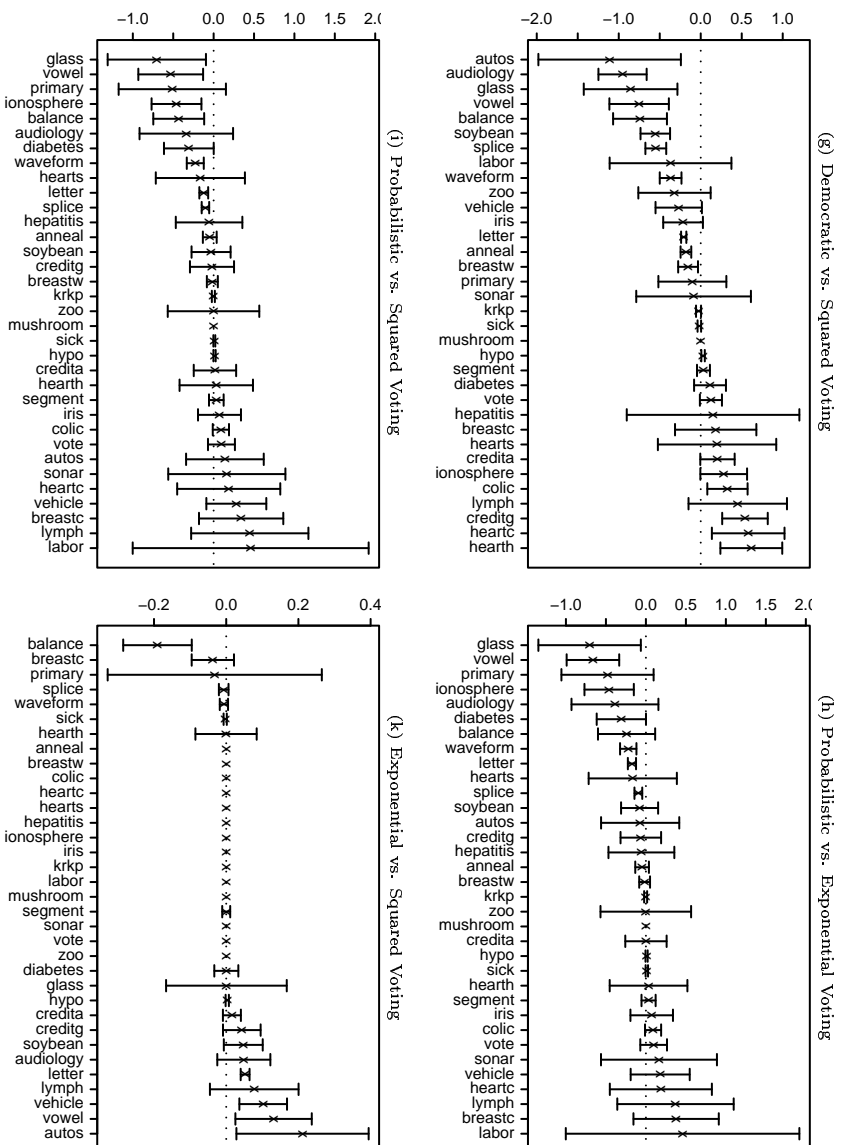


Figure 7: Comparisons of voting schemes (Part 2).

Dataset	\bar{L}	\bar{D}	\bar{D}_T	\bar{D}_F	\bar{D}_P	L	L^*
anneal	1.81±0.04	1.27±0.06	0.93±0.04	19.17±1.86	18.22±1.91	1.23±0.06	1.09
audiology	27.84±0.47	17.77±0.39	13.45±0.49	34.78±1.44	15.05±1.04	20.01±0.84	20.14
autos	32.31±0.78	25.66±0.51	20.57±0.65	44.97±1.29	23.39±0.97	20.65±0.78	20.96
balance	23.95±0.14	17.45±0.20	11.71±0.26	44.45±0.81	18.19±0.50	17.44±0.48	17.47
breastc	36.85±0.27	24.76±0.26	23.82±0.38	27.12±0.37	27.12±0.37	26.55±0.54	26.56
breastw	6.59±0.05	4.44±0.10	3.59±0.11	23.55±1.07	23.55±1.07	4.15±0.14	4.12
colic	23.50±0.19	14.12±0.28	13.43±0.28	18.26±0.68	18.26±0.68	14.67±0.17	14.74
credita	18.78±0.19	11.95±0.13	9.92±0.17	25.11±0.75	25.11±0.75	13.55±0.37	13.64
creditg	32.87±0.26	22.21±0.20	19.67±0.21	29.45±0.31	29.45±0.31	25.90±0.40	25.95
diabetes	30.55±0.17	20.15±0.20	17.17±0.33	29.23±0.47	29.23±0.47	24.86±0.57	24.96
glass	36.63±0.43	25.52±0.36	20.83±0.48	37.89±0.49	22.50±0.71	27.41±1.06	27.88
heartc	27.12±0.30	18.70±0.20	15.80±0.33	29.31±0.84	29.31±0.84	20.79±0.61	20.63
hearth	25.36±0.38	16.11±0.29	13.48±0.39	26.66±0.82	26.66±0.82	19.80±0.70	19.85
hearts	25.45±0.40	16.84±0.32	14.34±0.52	27.59±0.73	27.59±0.73	18.96±0.73	19.13
hepatitis	23.84±0.53	15.24±0.65	12.29±0.71	26.57±1.02	26.57±1.02	19.04±0.62	18.89
hypo	0.68±0.01	0.43±0.02	0.33±0.02	18.90±2.47	16.21±2.58	0.43±0.02	0.41
ionosphere	12.05±0.14	8.17±0.18	6.73±0.22	24.30±1.64	24.30±1.64	7.78±0.36	7.71
iris	6.92±0.34	3.12±0.17	2.17±0.16	12.05±1.54	11.97±1.53	5.93±0.46	5.53
krknp	1.07±0.02	0.70±0.02	0.58±0.02	17.60±2.36	17.60±2.36	0.62±0.03	0.60
labor	25.26±1.05	19.31±0.83	17.54±0.78	16.30±2.58	16.30±2.58	14.07±1.83	11.66
letter	15.56±0.05	13.50±0.06	10.19±0.06	53.62±0.19	19.40±0.16	7.63±0.08	7.63
lymph	26.46±0.55	17.88±0.34	14.06±0.42	31.01±1.47	24.75±1.21	20.66±0.85	20.27
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00
primary	68.67±0.25	48.31±0.45	38.31±0.59	55.49±0.51	9.50±0.24	58.17±0.70	58.17
segment	4.72±0.05	3.52±0.08	2.73±0.06	32.22±0.78	23.29±0.91	2.69±0.08	2.69
sick	1.70±0.04	1.16±0.04	0.86±0.03	27.08±1.08	27.08±1.08	1.16±0.04	1.17
sonar	30.33±0.34	23.59±0.40	20.22±0.67	35.04±0.41	35.04±0.41	22.42±1.31	22.59
soybean	14.15±0.21	11.19±0.18	8.83±0.16	38.41±0.73	25.34±1.01	8.05±0.32	8.08
splice	9.44±0.09	6.41±0.09	5.05±0.05	27.16±0.69	23.63±0.86	6.15±0.17	6.15
vehicle	30.57±0.22	21.69±0.27	15.95±0.25	37.94±0.48	28.46±0.41	26.25±0.38	26.31
vote	6.67±0.09	4.53±0.06	3.89±0.08	18.13±1.96	18.13±1.96	3.70±0.14	3.57
vowel	27.43±0.27	25.20±0.24	22.16±0.26	49.91±0.49	29.66±0.54	10.97±0.43	10.95
waveform	26.06±0.09	19.90±0.08	16.65±0.10	34.01±0.11	33.12±0.12	18.72±0.10	18.72
zoo	9.02±0.35	6.38±0.44	4.35±0.46	16.74±2.82	11.81±2.26	6.56±0.81	5.57
Mean	20.30	14.33	11.81	29.12	22.23	14.62	14.52

Table 11: Loss decomposition results for Democratic Voting.

GENERALIZED UNIFIED DECOMPOSITION OF ENSEMBLE LOSS

Dataset	\bar{L}	\bar{D}	\bar{D}_T	\bar{D}_F	\bar{D}_P	L	L^*
anneal	1.81±0.04	1.62±0.08	1.06±0.08	28.58±4.58	26.79±4.26	1.30±0.15	1.05
audiology	27.84±0.47	20.06±0.66	13.34±0.33	42.90±2.52	21.98±2.30	22.44±1.08	22.42
autos	32.31±0.78	30.22±0.58	20.49±0.92	57.35±1.28	34.78±1.27	26.30±0.89	26.43
balance	23.95±0.14	19.86±0.40	11.79±0.31	53.14±1.30	25.27±1.39	19.37±0.65	19.32
breastc	36.85±0.27	27.32±0.52	24.36±0.29	33.84±1.01	33.84±1.01	30.27±1.12	29.89
breastw	6.59±0.05	5.36±0.19	3.58±0.13	36.36±2.02	36.36±2.02	5.15±0.21	5.01
colic	23.50±0.19	14.72±0.40	13.24±0.32	22.43±1.62	22.43±1.62	15.95±0.48	15.95
credita	18.78±0.19	13.99±0.33	10.18±0.22	34.50±1.41	34.50±1.41	15.55±0.49	15.55
creditg	32.87±0.26	26.29±0.43	20.78±0.30	39.05±0.82	39.05±0.82	30.05±0.84	30.10
diabetes	30.55±0.17	24.72±0.35	18.68±0.38	39.84±0.87	39.84±0.87	28.54±0.88	28.60
glass	36.63±0.43	31.46±0.65	22.19±0.58	51.12±1.11	33.78±1.53	32.46±1.60	32.79
heartc	27.12±0.30	22.47±0.45	16.06±0.47	40.96±1.07	40.96±1.07	25.70±1.10	25.74
hearth	25.36±0.38	18.77±0.60	14.22±0.37	34.71±1.72	34.71±1.72	21.84±0.83	21.81
hearts	25.45±0.40	20.83±0.67	15.21±0.56	39.59±2.02	39.59±2.02	22.70±0.89	22.65
hepatitis	23.84±0.53	17.22±0.90	13.25±0.46	32.55±2.16	32.55±2.16	19.61±1.28	19.53
hypo	0.68±0.01	0.49±0.02	0.35±0.02	26.16±2.71	22.85±2.70	0.46±0.03	0.43
ionosphere	12.05±0.14	10.58±0.39	6.83±0.41	42.49±2.45	42.49±2.45	10.31±0.85	10.30
iris	6.92±0.34	3.89±0.36	2.70±0.20	14.21±4.30	14.12±4.29	5.67±0.59	5.07
krkbp	1.07±0.02	0.83±0.03	0.59±0.04	27.41±4.42	27.41±4.42	0.72±0.07	0.66
labor	25.26±1.05	21.10±0.95	17.60±1.00	24.39±3.43	24.39±3.43	17.00±1.76	13.20
letter	15.56±0.05	16.48±0.12	8.68±0.10	70.21±0.40	37.04±0.34	12.67±0.11	12.67
lymph	26.46±0.55	21.72±0.58	14.83±0.75	42.69±2.05	36.48±1.56	24.03±1.16	23.90
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00
primary	68.67±0.25	52.38±0.36	37.67±0.73	61.42±0.47	12.35±0.46	62.09±0.80	62.03
segment	4.72±0.05	4.54±0.15	2.51±0.11	51.44±1.61	42.89±1.70	4.10±0.23	4.05
sick	1.70±0.04	1.40±0.06	0.87±0.04	38.84±1.95	38.84±1.95	1.38±0.09	1.37
sonar	30.33±0.34	30.40±0.78	21.73±0.57	49.90±1.40	49.90±1.40	30.23±1.39	30.31
soybean	14.15±0.21	12.87±0.25	8.69±0.33	49.45±1.68	36.45±1.44	10.04±0.63	9.96
splice	9.44±0.09	7.36±0.10	5.15±0.08	36.85±0.98	32.87±0.92	6.93±0.16	6.91
vehicle	30.57±0.22	26.60±0.37	17.32±0.28	48.58±0.71	38.03±0.37	29.60±0.52	29.68
vote	6.67±0.09	4.79±0.10	3.84±0.14	22.49±3.35	22.49±3.35	4.07±0.27	3.84
vowel	27.43±0.27	31.63±0.49	20.60±0.23	69.60±0.37	49.07±0.62	22.56±0.74	22.54
waveform	26.06±0.09	25.76±0.19	17.54±0.17	50.02±0.44	48.74±0.40	25.28±0.24	25.25
zoo	9.02±0.35	8.31±0.70	4.27±0.57	31.35±5.90	25.24±4.90	8.54±1.15	6.73
Mean	20.30	16.94	12.06	39.54	32.30	17.44	17.23

Table 12: Loss decomposition results for Aristocratic Voting.

Dataset	\bar{L}	\bar{D}	\bar{D}_T	\bar{D}_F	\bar{D}_P	L	L^*
anneal	1.81±0.04	1.26±0.06	0.99±0.06	17.18±2.58	16.44±2.38	1.10±0.06	0.99
audiology	27.84±0.47	17.70±0.38	13.77±0.60	33.38±1.41	14.53±0.75	19.44±1.02	19.62
autos	32.31±0.78	25.52±0.53	20.96±0.57	43.38±1.35	20.81±1.15	19.49±1.02	19.50
balance	23.95±0.14	17.46±0.19	11.88±0.25	44.31±0.65	17.84±0.55	17.16±0.51	17.17
breastc	36.85±0.27	24.66±0.24	23.81±0.39	26.82±0.31	26.82±0.31	26.40±0.55	26.41
breastw	6.59±0.05	4.44±0.10	3.65±0.15	22.91±1.24	22.91±1.24	4.02±0.21	4.00
colic	23.50±0.19	14.05±0.28	13.30±0.29	18.45±0.60	18.45±0.60	14.89±0.27	14.95
credita	18.78±0.19	11.92±0.13	9.82±0.23	25.24±0.72	25.24±0.72	13.72±0.44	13.80
creditg	32.87±0.26	22.12±0.21	19.39±0.25	29.68±0.25	29.68±0.25	26.42±0.34	26.47
diabetes	30.55±0.17	20.13±0.19	17.00±0.33	29.52±0.50	29.52±0.50	25.25±0.61	25.33
glass	36.63±0.43	25.61±0.37	20.98±0.46	37.80±0.96	22.66±1.05	27.27±0.84	27.76
heartc	27.12±0.30	18.66±0.21	15.62±0.36	29.81±0.88	29.81±0.88	21.15±0.64	21.07
hearth	25.36±0.38	16.12±0.29	13.24±0.39	27.54±0.67	27.54±0.67	20.33±0.74	20.46
hearts	25.45±0.40	16.80±0.31	14.18±0.51	27.72±0.81	27.72±0.81	19.30±0.97	19.39
hepatitis	23.84±0.53	15.08±0.64	12.10±0.71	27.05±1.47	27.05±1.47	19.23±0.76	19.28
hypo	0.68±0.01	0.43±0.02	0.32±0.01	19.42±2.25	16.61±2.26	0.45±0.02	0.43
ionosphere	12.05±0.14	8.18±0.18	6.41±0.24	25.32±1.43	25.32±1.43	8.46±0.37	8.26
iris	6.92±0.34	3.12±0.17	2.31±0.19	11.03±1.89	10.95±1.89	5.67±0.50	5.31
krknp	1.07±0.02	0.70±0.02	0.59±0.02	17.38±2.48	17.38±2.48	0.59±0.03	0.58
labor	25.26±1.05	19.18±0.84	17.76±0.85	15.27±1.75	15.27±1.75	13.30±1.28	11.19
letter	15.56±0.05	13.50±0.06	10.25±0.04	53.29±0.23	19.43±0.20	7.55±0.07	7.55
lymph	26.46±0.55	17.79±0.37	14.03±0.62	31.83±1.70	25.27±0.83	20.67±0.75	20.48
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00
primary	68.67±0.25	47.83±0.44	37.72±0.47	54.87±0.54	9.41±0.35	58.55±0.84	58.54
segment	4.72±0.05	3.53±0.08	2.73±0.08	32.53±0.73	23.74±0.77	2.69±0.11	2.70
sick	1.70±0.04	1.16±0.04	0.87±0.04	26.46±0.91	26.46±0.91	1.14±0.04	1.14
sonar	30.33±0.34	23.58±0.39	20.28±0.59	34.48±0.51	34.48±0.51	22.19±1.19	22.22
soybean	14.15±0.21	11.12±0.18	9.07±0.15	36.49±1.09	24.00±1.40	7.57±0.34	7.59
splice	9.44±0.09	6.29±0.07	5.19±0.07	24.38±0.38	20.80±0.53	5.73±0.11	5.74
vehicle	30.57±0.22	21.71±0.27	16.23±0.14	37.57±0.57	28.13±0.49	25.74±0.43	25.79
vote	6.67±0.09	4.51±0.07	3.87±0.12	17.49±2.47	17.49±2.47	3.72±0.20	3.56
vowel	27.43±0.27	25.26±0.24	22.24±0.25	50.29±0.60	29.53±0.42	10.77±0.46	10.77
waveform	26.06±0.09	19.91±0.09	16.71±0.10	33.92±0.19	33.02±0.19	18.59±0.16	18.59
zoo	9.02±0.35	6.31±0.44	4.47±0.32	15.63±2.96	10.83±2.30	6.26±0.85	5.37
Mean	20.30	14.28	11.82	28.78	21.92	14.55	14.47

Table 13: Loss decomposition results for Probabilistic Voting.

GENERALIZED UNIFIED DECOMPOSITION OF ENSEMBLE LOSS

Dataset	\bar{L}	\bar{D}	\bar{D}_T	\bar{D}_F	\bar{D}_P	L	L^*
anneal	1.81±0.04	1.26±0.06	1.01±0.06	16.47±2.24	15.91±2.18	1.06±0.07	0.96
audiology	27.84±0.47	17.63±0.38	13.89±0.49	33.08±1.44	13.94±1.09	19.12±0.84	19.33
autos	32.31±0.78	25.40±0.52	20.82±0.54	43.16±1.46	20.62±1.28	19.62±1.12	19.63
balance	23.95±0.14	17.35±0.19	12.06±0.13	43.58±0.74	17.10±0.36	16.76±0.33	16.78
breastc	36.85±0.27	24.57±0.25	23.66±0.30	27.07±0.31	27.07±0.31	26.72±0.29	26.78
breastw	6.59±0.05	4.43±0.10	3.65±0.16	22.29±1.17	22.29±1.17	4.01±0.24	3.96
colic	23.50±0.19	14.04±0.28	13.25±0.28	18.64±0.66	18.64±0.66	14.97±0.24	15.05
credita	18.78±0.19	11.86±0.12	9.79±0.20	25.18±0.59	25.18±0.59	13.74±0.43	13.84
creditg	32.87±0.26	21.97±0.20	19.30±0.22	29.41±0.34	29.41±0.34	26.40±0.42	26.46
diabetes	30.55±0.17	19.98±0.19	17.01±0.27	28.96±0.45	28.96±0.45	24.97±0.48	25.06
glass	36.63±0.43	25.43±0.35	21.15±0.52	37.08±0.79	21.66±0.95	26.61±1.16	27.06
heartc	27.12±0.30	18.54±0.20	15.44±0.19	29.63±0.68	29.63±0.68	21.33±0.62	21.27
hearth	25.36±0.38	16.03±0.28	13.16±0.37	27.18±0.64	27.18±0.64	20.37±0.68	20.44
hearts	25.45±0.40	16.72±0.31	14.18±0.46	27.31±0.82	27.31±0.82	19.15±0.92	19.25
hepatitis	23.84±0.53	15.02±0.62	12.11±0.61	26.62±1.25	26.62±1.25	19.18±0.71	19.14
hypo	0.68±0.01	0.43±0.02	0.32±0.02	20.43±1.74	17.65±1.78	0.46±0.01	0.44
ionosphere	12.05±0.14	8.15±0.18	6.59±0.21	24.63±1.72	24.63±1.72	8.04±0.38	7.93
iris	6.92±0.34	3.12±0.17	2.27±0.16	10.90±1.59	10.82±1.60	5.73±0.42	5.35
krknp	1.07±0.02	0.70±0.02	0.59±0.02	17.37±1.95	17.37±1.95	0.59±0.03	0.58
labor	25.26±1.05	19.09±0.81	17.46±0.67	16.02±2.75	16.02±2.75	13.73±1.84	11.72
letter	15.56±0.05	13.44±0.06	10.29±0.06	52.67±0.20	18.80±0.21	7.44±0.09	7.44
lymph	26.46±0.55	17.69±0.36	13.75±0.54	31.72±1.75	25.44±1.04	21.08±0.97	20.91
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00
primary	68.67±0.25	47.79±0.41	38.07±0.74	54.76±0.47	9.25±0.34	58.07±0.77	58.09
segment	4.72±0.05	3.51±0.08	2.71±0.08	32.36±0.79	23.69±0.90	2.72±0.09	2.73
sick	1.70±0.04	1.16±0.04	0.86±0.03	26.63±0.99	26.63±0.99	1.15±0.04	1.15
sonar	30.33±0.34	23.48±0.39	20.21±0.64	34.84±0.54	34.84±0.54	22.34±1.08	22.52
soybean	14.15±0.21	11.09±0.18	9.07±0.20	36.10±0.72	23.98±1.14	7.54±0.31	7.59
splice	9.44±0.09	6.27±0.07	5.23±0.08	23.76±0.49	20.23±0.64	5.64±0.13	5.65
vehicle	30.57±0.22	21.55±0.27	15.99±0.20	37.39±0.52	28.04±0.36	26.00±0.37	26.06
vote	6.67±0.09	4.50±0.07	3.82±0.10	17.79±2.03	17.79±2.03	3.81±0.15	3.64
vowel	27.43±0.27	25.12±0.23	22.36±0.25	49.29±0.54	28.33±0.62	10.27±0.45	10.29
waveform	26.06±0.09	19.82±0.08	16.74±0.11	33.47±0.15	32.59±0.15	18.38±0.19	18.38
zoo	9.02±0.35	6.26±0.41	4.47±0.40	15.52±2.38	11.02±2.09	6.26±0.73	5.38
Mean	20.30	14.22	11.80	28.57	21.73	14.51	14.44

Table 14: Loss decomposition results for ProgressiveSquare Voting.

Dataset	\bar{L}	\bar{D}	\bar{D}_T	\bar{D}_F	\bar{D}_P	L	L^*
anneal	1.81±0.04	1.26±0.06	1.01±0.06	16.47±2.24	15.91±2.18	1.06±0.07	0.96
audiology	27.84±0.47	17.62±0.38	13.92±0.51	32.26±1.02	13.65±0.92	19.08±0.87	19.23
autos	32.31±0.78	25.38±0.52	20.92±0.55	42.95±1.38	20.66±1.31	19.43±1.08	19.50
balance	23.95±0.14	17.34±0.19	11.95±0.16	43.70±0.77	17.31±0.39	16.93±0.34	16.96
breastc	36.85±0.27	24.57±0.25	23.64±0.30	27.09±0.30	27.09±0.30	26.75±0.31	26.81
breastw	6.59±0.05	4.43±0.10	3.65±0.16	22.29±1.17	22.29±1.17	4.01±0.24	3.96
colic	23.50±0.19	14.04±0.28	13.25±0.28	18.64±0.66	18.64±0.66	14.97±0.24	15.05
credita	18.78±0.19	11.86±0.12	9.79±0.19	25.16±0.59	25.16±0.59	13.72±0.41	13.82
creditg	32.87±0.26	21.97±0.20	19.32±0.22	29.38±0.32	29.38±0.32	26.36±0.41	26.42
diabetes	30.55±0.17	19.98±0.19	17.01±0.27	28.96±0.45	28.96±0.45	24.97±0.49	25.05
glass	36.63±0.43	25.42±0.35	21.15±0.58	37.00±0.77	21.65±0.94	26.61±1.23	27.05
heartc	27.12±0.30	18.54±0.20	15.44±0.19	29.63±0.68	29.63±0.68	21.33±0.62	21.27
hearth	25.36±0.38	16.03±0.28	13.17±0.35	27.20±0.62	27.20±0.62	20.37±0.64	20.44
hearts	25.45±0.40	16.72±0.31	14.18±0.46	27.31±0.82	27.31±0.82	19.15±0.92	19.25
hepatitis	23.84±0.53	15.02±0.62	12.11±0.61	26.62±1.25	26.62±1.25	19.18±0.71	19.14
hypo	0.68±0.01	0.43±0.02	0.32±0.02	20.13±1.75	17.45±1.80	0.45±0.01	0.44
ionosphere	12.05±0.14	8.15±0.18	6.59±0.21	24.63±1.72	24.63±1.72	8.04±0.38	7.93
iris	6.92±0.34	3.12±0.17	2.27±0.16	10.90±1.59	10.82±1.60	5.73±0.42	5.35
krkbp	1.07±0.02	0.70±0.02	0.59±0.02	17.37±1.95	17.37±1.95	0.59±0.03	0.58
labor	25.26±1.05	19.09±0.81	17.46±0.67	16.02±2.75	16.02±2.75	13.73±1.84	11.72
letter	15.56±0.05	13.43±0.06	10.31±0.06	52.50±0.20	18.70±0.19	7.39±0.09	7.39
lymph	26.46±0.55	17.69±0.37	13.79±0.52	31.69±1.72	25.39±1.00	21.00±0.92	20.84
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00
primary	68.67±0.25	47.64±0.43	38.01±0.59	54.54±0.48	9.21±0.32	58.10±0.82	58.10
segment	4.72±0.05	3.51±0.08	2.71±0.08	32.38±0.81	23.69±0.90	2.72±0.09	2.73
sick	1.70±0.04	1.16±0.04	0.86±0.03	26.65±1.00	26.65±1.00	1.15±0.05	1.15
sonar	30.33±0.34	23.48±0.39	20.21±0.64	34.84±0.54	34.84±0.54	22.34±1.08	22.52
soybean	14.15±0.21	11.08±0.18	9.09±0.19	35.81±0.81	23.86±1.21	7.49±0.31	7.54
splice	9.44±0.09	6.27±0.07	5.22±0.08	23.81±0.50	20.26±0.66	5.65±0.14	5.65
vehicle	30.57±0.22	21.54±0.27	16.05±0.20	37.28±0.49	28.02±0.34	25.91±0.37	25.97
vote	6.67±0.09	4.50±0.07	3.82±0.10	17.79±2.03	17.79±2.03	3.81±0.15	3.64
vowel	27.43±0.27	25.11±0.23	22.41±0.24	49.11±0.53	28.16±0.61	10.15±0.41	10.16
waveform	26.06±0.09	19.82±0.08	16.74±0.11	33.47±0.14	32.60±0.15	18.39±0.19	18.39
zoo	9.02±0.35	6.26±0.41	4.47±0.40	15.52±2.38	11.02±2.09	6.26±0.73	5.38
Mean	20.30	14.21	11.81	28.50	21.70	14.49	14.42

Table 15: Loss decomposition results for ProgressiveExp Voting.

GENERALIZED UNIFIED DECOMPOSITION OF ENSEMBLE LOSS

Dataset	\bar{L}	\bar{D}	\bar{D}_T	\bar{D}_F	\bar{D}_P	L	L^*
anneal	1.81±0.04	1.26±0.06	1.00±0.06	17.31±2.40	16.48±2.26	1.08±0.07	0.98
audiology	27.84±0.47	17.64±0.38	13.61±0.47	32.93±1.05	14.28±0.96	19.65±0.83	19.74
autos	32.31±0.78	25.41±0.52	20.76±0.54	43.69±1.54	21.41±1.21	19.77±0.95	19.97
balance	23.95±0.14	17.35±0.19	11.86±0.15	43.80±0.81	17.49±0.41	17.11±0.33	17.12
breastc	36.85±0.27	24.59±0.25	23.79±0.28	26.78±0.26	26.78±0.26	26.37±0.41	26.42
breastw	6.59±0.05	4.43±0.10	3.66±0.17	22.26±1.19	22.26±1.19	3.99±0.24	3.95
colic	23.50±0.19	14.05±0.28	13.31±0.28	18.54±0.60	18.54±0.60	14.86±0.21	14.95
credita	18.78±0.19	11.87±0.12	9.85±0.21	24.91±0.55	24.91±0.55	13.61±0.41	13.70
creditg	32.87±0.26	21.99±0.20	19.38±0.20	29.30±0.49	29.30±0.49	26.24±0.56	26.29
diabetes	30.55±0.17	19.99±0.19	17.07±0.30	28.91±0.42	28.91±0.42	24.85±0.54	24.95
glass	36.63±0.43	25.43±0.35	21.18±0.47	37.00±0.62	21.63±0.88	26.61±1.06	27.01
heartc	27.12±0.30	18.55±0.20	15.48±0.18	29.50±0.69	29.50±0.69	21.23±0.58	21.16
hearth	25.36±0.38	16.05±0.28	13.36±0.33	26.73±0.86	26.73±0.86	19.96±0.62	20.03
hearts	25.45±0.40	16.72±0.31	14.24±0.43	27.11±0.82	27.11±0.82	19.04±0.87	19.11
hepatitis	23.84±0.53	15.02±0.62	12.09±0.58	26.64±1.29	26.64±1.29	19.24±0.80	19.17
hypo	0.68±0.01	0.43±0.02	0.32±0.02	18.77±1.92	16.20±1.95	0.44±0.02	0.42
ionosphere	12.05±0.14	8.15±0.18	6.61±0.20	24.22±1.24	24.22±1.24	8.01±0.34	7.86
iris	6.92±0.34	3.12±0.17	2.27±0.16	10.90±1.59	10.82±1.60	5.73±0.42	5.35
krkcp	1.07±0.02	0.70±0.02	0.59±0.02	17.37±1.95	17.37±1.95	0.59±0.03	0.58
labor	25.26±1.05	19.10±0.81	17.47±0.67	15.60±2.32	15.60±2.32	13.73±1.84	11.64
letter	15.56±0.05	13.43±0.06	10.30±0.05	52.54±0.22	18.81±0.18	7.42±0.08	7.42
lymph	26.46±0.55	17.70±0.36	13.72±0.57	31.89±1.82	25.61±1.01	21.13±0.76	21.00
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00
primary	68.67±0.25	47.72±0.44	37.95±0.55	54.70±0.56	9.26±0.31	58.20±0.62	58.21
segment	4.72±0.05	3.51±0.08	2.71±0.08	32.26±0.81	23.64±0.90	2.72±0.09	2.73
sick	1.70±0.04	1.16±0.04	0.87±0.04	26.41±1.00	26.41±1.00	1.14±0.05	1.14
sonar	30.33±0.34	23.48±0.39	20.23±0.62	34.83±0.49	34.83±0.49	22.29±1.02	22.47
soybean	14.15±0.21	11.09±0.18	9.03±0.17	36.08±0.95	24.00±1.15	7.61±0.33	7.64
splice	9.44±0.09	6.28±0.07	5.21±0.07	24.08±0.47	20.57±0.66	5.69±0.13	5.70
vehicle	30.57±0.22	21.55±0.27	16.03±0.20	37.32±0.49	28.01±0.35	25.95±0.37	26.00
vote	6.67±0.09	4.51±0.06	3.79±0.11	17.95±2.58	17.95±2.58	3.88±0.19	3.68
vowel	27.43±0.27	25.11±0.23	22.39±0.24	49.16±0.46	28.31±0.64	10.22±0.41	10.23
waveform	26.06±0.09	19.82±0.08	16.75±0.12	33.46±0.16	32.59±0.15	18.37±0.20	18.37
zoo	9.02±0.35	6.26±0.41	4.47±0.40	15.53±2.39	11.02±2.09	6.26±0.73	5.38
Mean	20.30	14.22	11.80	28.48	21.68	14.50	14.42

Table 16: Loss decomposition results for ProgressiveSquareNorm Voting.

Dataset	\bar{L}	\bar{D}	\bar{D}_T	\bar{D}_F	\bar{D}_P	L	L^*
anneal	1.81±0.04	1.26±0.06	1.02±0.06	15.97±2.06	15.41±2.02	1.05±0.07	0.95
audiology	27.84±0.47	17.62±0.38	13.92±0.51	32.26±1.02	13.65±0.92	19.08±0.87	19.23
autos	32.31±0.78	25.38±0.52	20.88±0.56	43.06±1.44	20.62±1.27	19.48±1.08	19.53
balance	23.95±0.14	17.34±0.19	11.92±0.14	43.70±0.79	17.33±0.38	16.98±0.31	17.00
breastc	36.85±0.27	24.57±0.25	23.67±0.30	27.04±0.31	27.04±0.31	26.68±0.31	26.74
breastw	6.59±0.05	4.43±0.10	3.65±0.16	22.29±1.17	22.29±1.17	4.01±0.24	3.96
colic	23.50±0.19	14.04±0.28	13.25±0.27	18.64±0.63	18.64±0.63	14.97±0.23	15.05
credita	18.78±0.19	11.86±0.12	9.81±0.21	25.09±0.66	25.09±0.66	13.70±0.45	13.79
creditg	32.87±0.26	21.97±0.20	19.31±0.22	29.39±0.37	29.39±0.37	26.38±0.44	26.44
diabetes	30.55±0.17	19.98±0.19	17.04±0.28	28.92±0.46	28.92±0.46	24.92±0.52	25.00
glass	36.63±0.43	25.42±0.35	21.22±0.54	36.93±0.73	21.59±0.94	26.52±1.17	26.94
heartc	27.12±0.30	18.54±0.20	15.44±0.19	29.63±0.68	29.63±0.68	21.33±0.62	21.27
hearth	25.36±0.38	16.03±0.28	13.16±0.37	27.18±0.64	27.18±0.64	20.37±0.68	20.44
hearts	25.45±0.40	16.72±0.31	14.18±0.46	27.31±0.82	27.31±0.82	19.15±0.92	19.25
hepatitis	23.84±0.53	15.02±0.62	12.11±0.61	26.62±1.25	26.62±1.25	19.18±0.71	19.14
hypo	0.68±0.01	0.43±0.02	0.32±0.02	20.38±1.66	17.69±1.76	0.46±0.02	0.44
ionosphere	12.05±0.14	8.15±0.18	6.57±0.21	24.72±1.70	24.72±1.70	8.10±0.36	7.98
iris	6.92±0.34	3.12±0.17	2.27±0.16	10.90±1.59	10.82±1.60	5.73±0.42	5.35
krkcp	1.07±0.02	0.70±0.02	0.59±0.02	17.12±2.00	17.12±2.00	0.59±0.03	0.58
labor	25.26±1.05	19.09±0.81	17.46±0.67	16.02±2.75	16.02±2.75	13.73±1.84	11.72
letter	15.56±0.05	13.43±0.06	10.31±0.06	52.50±0.19	18.70±0.19	7.39±0.09	7.39
lymph	26.46±0.55	17.69±0.37	13.79±0.52	31.69±1.72	25.39±1.00	21.00±0.92	20.84
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00
primary	68.67±0.25	47.63±0.43	37.96±0.60	54.56±0.49	9.23±0.32	58.16±0.79	58.16
segment	4.72±0.05	3.51±0.08	2.71±0.08	32.41±0.82	23.71±0.91	2.72±0.09	2.73
sick	1.70±0.04	1.16±0.04	0.86±0.03	26.63±0.99	26.63±0.99	1.15±0.04	1.15
sonar	30.33±0.34	23.48±0.39	20.23±0.62	34.82±0.54	34.82±0.54	22.29±1.04	22.47
soybean	14.15±0.21	11.08±0.18	9.10±0.19	35.72±0.73	23.81±1.22	7.48±0.31	7.52
splice	9.44±0.09	6.27±0.07	5.22±0.08	23.81±0.45	20.27±0.61	5.65±0.13	5.66
vehicle	30.57±0.22	21.54±0.27	16.04±0.19	37.28±0.51	27.99±0.37	25.91±0.37	25.97
vote	6.67±0.09	4.50±0.07	3.82±0.10	17.79±2.03	17.79±2.03	3.81±0.15	3.64
vowel	27.43±0.27	25.11±0.23	22.42±0.25	49.07±0.52	28.18±0.62	10.14±0.42	10.15
waveform	26.06±0.09	19.82±0.08	16.75±0.11	33.46±0.14	32.59±0.14	18.37±0.19	18.38
zoo	9.02±0.35	6.26±0.41	4.42±0.39	16.02±2.44	11.52±2.05	6.36±0.74	5.47
Mean	20.30	14.21	11.81	28.50	21.70	14.49	14.42

Table 17: Loss decomposition results for ProgressiveExpNorm Voting.

GENERALIZED UNIFIED DECOMPOSITION OF ENSEMBLE LOSS

Dataset	DEM	ARIST	PROB	PR-SQ	PR-EXP	PR-SQN	PR-EXPN
anneal	1.81±0.04	1.81±0.04	1.81±0.04	1.81±0.04	1.81±0.04	1.81±0.04	1.81± 0.04
audiology	27.84±0.47	27.84±0.47	27.84±0.47	27.84±0.47	27.84±0.47	27.84±0.47	27.84± 0.47
autos	32.31±0.78	32.31±0.78	32.31±0.78	32.31±0.78	32.31±0.78	32.31±0.78	32.31± 0.78
balance	23.95±0.14	23.95±0.14	23.95±0.14	23.95±0.14	23.95±0.14	23.95±0.14	23.95± 0.14
breastc	36.85±0.27	36.85±0.27	36.85±0.27	36.85±0.27	36.85±0.27	36.85±0.27	36.85± 0.27
breastw	6.59±0.05	6.59±0.05	6.59±0.05	6.59±0.05	6.59±0.05	6.59±0.05	6.59± 0.05
colic	23.50±0.19	23.50±0.19	23.50±0.19	23.50±0.19	23.50±0.19	23.50±0.19	23.50± 0.19
credita	18.78±0.19	18.78±0.19	18.78±0.19	18.78±0.19	18.78±0.19	18.78±0.19	18.78± 0.19
creditg	32.87±0.26	32.87±0.26	32.87±0.26	32.87±0.26	32.87±0.26	32.87±0.26	32.87± 0.26
diabetes	30.55±0.17	30.55±0.17	30.55±0.17	30.55±0.17	30.55±0.17	30.55±0.17	30.55± 0.17
glass	36.63±0.43	36.63±0.43	36.63±0.43	36.63±0.43	36.63±0.43	36.63±0.43	36.63± 0.43
heartc	27.12±0.30	27.12±0.30	27.12±0.30	27.12±0.30	27.12±0.30	27.12±0.30	27.12± 0.30
hearth	25.36±0.38	25.36±0.38	25.36±0.38	25.36±0.38	25.36±0.38	25.36±0.38	25.36± 0.38
hearts	25.45±0.40	25.45±0.40	25.45±0.40	25.45±0.40	25.45±0.40	25.45±0.40	25.45± 0.40
hepatitis	23.84±0.53	23.84±0.53	23.84±0.53	23.84±0.53	23.84±0.53	23.84±0.53	23.84± 0.53
hypo	0.68±0.01	0.68±0.01	0.68±0.01	0.68±0.01	0.68±0.01	0.68±0.01	0.68± 0.01
ionosphere	12.05±0.14	12.05±0.14	12.05±0.14	12.05±0.14	12.05±0.14	12.05±0.14	12.05± 0.14
iris	6.92±0.34	6.92±0.34	6.92±0.34	6.92±0.34	6.92±0.34	6.92±0.34	6.92± 0.34
krkcp	1.07±0.02	1.07±0.02	1.07±0.02	1.07±0.02	1.07±0.02	1.07±0.02	1.07± 0.02
labor	25.26±1.05	25.26±1.05	25.26±1.05	25.26±1.05	25.26±1.05	25.26±1.05	25.26± 1.05
letter	15.56±0.05	15.56±0.05	15.56±0.05	15.56±0.05	15.56±0.05	15.56±0.05	15.56± 0.05
lymph	26.46±0.55	26.46±0.55	26.46±0.55	26.46±0.55	26.46±0.55	26.46±0.55	26.46± 0.55
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00± 0.00
primary	68.67±0.25	68.67±0.25	68.67±0.25	68.67±0.25	68.67±0.25	68.67±0.25	68.67± 0.25
segment	4.72±0.05	4.72±0.05	4.72±0.05	4.72±0.05	4.72±0.05	4.72±0.05	4.72± 0.05
sick	1.70±0.04	1.70±0.04	1.70±0.04	1.70±0.04	1.70±0.04	1.70±0.04	1.70± 0.04
sonar	30.33±0.34	30.33±0.34	30.33±0.34	30.33±0.34	30.33±0.34	30.33±0.34	30.33± 0.34
soybean	14.15±0.21	14.15±0.21	14.15±0.21	14.15±0.21	14.15±0.21	14.15±0.21	14.15± 0.21
splice	9.44±0.09	9.44±0.09	9.44±0.09	9.44±0.09	9.44±0.09	9.44±0.09	9.44± 0.09
vehicle	30.57±0.22	30.57±0.22	30.57±0.22	30.57±0.22	30.57±0.22	30.57±0.22	30.57± 0.22
vote	6.67±0.09	6.67±0.09	6.67±0.09	6.67±0.09	6.67±0.09	6.67±0.09	6.67± 0.09
vowel	27.43±0.27	27.43±0.27	27.43±0.27	27.43±0.27	27.43±0.27	27.43±0.27	27.43± 0.27
waveform	26.06±0.09	26.06±0.09	26.06±0.09	26.06±0.09	26.06±0.09	26.06±0.09	26.06± 0.09
zoo	9.02±0.35	9.02±0.35	9.02±0.35	9.02±0.35	9.02±0.35	9.02±0.35	9.02± 0.35
Mean	20.30	20.30	20.30	20.30	20.30	20.30	20.30

Table 18: Comparison of \bar{L} .

Dataset	DEM	ARIST	PROB	PR-SQ	PR-EXP	PR-SQN	PR-EXPN
anneal	0.93±0.04	1.06±0.08	0.99±0.06	1.01±0.06	1.01±0.06	1.00±0.06	1.02±0.06
audiology	13.45±0.49	13.34±0.33	13.77±0.60	13.89±0.49	13.92±0.51	13.61±0.47	13.92±0.51
autos	20.57±0.65	20.49±0.92	20.96±0.57	20.82±0.54	20.92±0.55	20.76±0.54	20.88±0.56
balance	11.71±0.26	11.79±0.31	11.88±0.25	12.06±0.13	11.95±0.16	11.86±0.15	11.92±0.14
breastc	23.82±0.38	24.36±0.29	23.81±0.39	23.66±0.30	23.64±0.30	23.79±0.28	23.67±0.30
breastw	3.59±0.11	3.58±0.13	3.65±0.15	3.65±0.16	3.65±0.16	3.66±0.17	3.65±0.16
colic	13.43±0.28	13.24±0.32	13.30±0.29	13.25±0.28	13.25±0.28	13.31±0.28	13.25±0.27
credita	9.92±0.17	10.18±0.22	9.82±0.23	9.79±0.20	9.79±0.19	9.85±0.21	9.81±0.21
creditg	19.67±0.21	20.78±0.30	19.39±0.25	19.30±0.22	19.32±0.22	19.38±0.20	19.31±0.22
diabetes	17.17±0.33	18.68±0.38	17.00±0.33	17.01±0.27	17.01±0.27	17.07±0.30	17.04±0.28
glass	20.83±0.48	22.19±0.58	20.98±0.46	21.15±0.52	21.15±0.58	21.18±0.47	21.22±0.54
heartc	15.80±0.33	16.06±0.47	15.62±0.36	15.44±0.19	15.44±0.19	15.48±0.18	15.44±0.19
hearth	13.48±0.39	14.22±0.37	13.24±0.39	13.16±0.37	13.17±0.35	13.36±0.33	13.16±0.37
hearts	14.34±0.52	15.21±0.56	14.18±0.51	14.18±0.46	14.18±0.46	14.24±0.43	14.18±0.46
hepatitis	12.29±0.71	13.25±0.46	12.10±0.71	12.11±0.61	12.11±0.61	12.09±0.58	12.11±0.61
hypo	0.33±0.02	0.35±0.02	0.32±0.01	0.32±0.02	0.32±0.02	0.32±0.02	0.32±0.02
ionosphere	6.73±0.22	6.83±0.41	6.41±0.24	6.59±0.21	6.59±0.21	6.61±0.20	6.57±0.21
iris	2.17±0.16	2.70±0.20	2.31±0.19	2.27±0.16	2.27±0.16	2.27±0.16	2.27±0.16
krkp	0.58±0.02	0.59±0.04	0.59±0.02	0.59±0.02	0.59±0.02	0.59±0.02	0.59±0.02
labor	17.54±0.78	17.60±1.00	17.76±0.85	17.46±0.67	17.46±0.67	17.47±0.67	17.46±0.67
letter	10.19±0.06	8.68±0.10	10.25±0.04	10.29±0.06	10.31±0.06	10.30±0.05	10.31±0.06
lymph	14.06±0.42	14.83±0.75	14.03±0.62	13.75±0.54	13.79±0.52	13.72±0.57	13.79±0.52
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
primary	38.31±0.59	37.67±0.73	37.72±0.47	38.07±0.74	38.01±0.59	37.95±0.55	37.96±0.60
segment	2.73±0.06	2.51±0.11	2.73±0.08	2.71±0.08	2.71±0.08	2.71±0.08	2.71±0.08
sick	0.86±0.03	0.87±0.04	0.87±0.04	0.86±0.03	0.86±0.03	0.87±0.04	0.86±0.03
sonar	20.22±0.67	21.73±0.57	20.28±0.59	20.21±0.64	20.21±0.64	20.23±0.62	20.23±0.62
soybean	8.83±0.16	8.69±0.33	9.07±0.15	9.07±0.20	9.09±0.19	9.03±0.17	9.10±0.19
splice	5.05±0.05	5.15±0.08	5.19±0.07	5.23±0.08	5.22±0.08	5.21±0.07	5.22±0.08
vehicle	15.95±0.25	17.32±0.28	16.23±0.14	15.99±0.20	16.05±0.20	16.03±0.20	16.04±0.19
vote	3.89±0.08	3.84±0.14	3.87±0.12	3.82±0.10	3.82±0.10	3.79±0.11	3.82±0.10
vowel	22.16±0.26	20.60±0.23	22.24±0.25	22.36±0.25	22.41±0.24	22.39±0.24	22.42±0.25
waveform	16.65±0.10	17.54±0.17	16.71±0.10	16.74±0.11	16.74±0.11	16.75±0.12	16.75±0.11
zoo	4.35±0.46	4.27±0.57	4.47±0.32	4.47±0.40	4.47±0.40	4.47±0.40	4.42±0.39
Mean	11.81	12.06	11.82	11.80	11.81	11.80	11.81

Table 19: Comparison of \bar{D} .

GENERALIZED UNIFIED DECOMPOSITION OF ENSEMBLE LOSS

Dataset	DEM	ARIST	PROB	PR-SQ	PR-EXP	PR-SQN	PR-EXPN
anneal	19.17±1.86	28.58±4.58	17.18±2.58	16.47±2.24	16.47±2.24	17.31±2.40	15.97± 2.06
audiology	34.78±1.44	42.90±2.52	33.38±1.41	33.08±1.44	32.26± 1.02	32.93±1.05	32.26± 1.02
autos	44.97±1.29	57.35±1.28	43.38±1.35	43.16±1.46	42.95± 1.38	43.69±1.54	43.06± 1.44
balance	44.45±0.81	53.14±1.30	44.31±0.65	43.58± 0.74	43.70±0.77	43.80±0.81	43.70± 0.79
breastc	27.12±0.37	33.84±1.01	26.82±0.31	27.07±0.31	27.09±0.30	26.78± 0.26	27.04± 0.31
breastw	23.55±1.07	36.36±2.02	22.91±1.24	22.29±1.17	22.29±1.17	22.26± 1.19	22.29± 1.17
colic	18.26± 0.68	22.43±1.62	18.45±0.60	18.64±0.66	18.64±0.66	18.54±0.60	18.64± 0.63
credita	25.11± 0.75	34.50±1.41	25.24±0.72	25.18±0.59	25.16±0.59	24.91±0.55	25.09± 0.66
creditg	29.45± 0.31	39.05±0.82	29.68±0.25	29.41±0.34	29.38±0.32	29.30±0.49	29.39± 0.37
diabetes	29.23±0.47	39.84±0.87	29.52±0.50	28.96±0.45	28.96±0.45	28.91± 0.42	28.92± 0.46
glass	37.89±0.49	51.12±1.11	37.80±0.96	37.08±0.79	37.00±0.77	37.00±0.62	36.93± 0.73
heartc	29.31± 0.84	40.96±1.07	29.81±0.88	29.63±0.68	29.63±0.68	29.50±0.69	29.63± 0.68
hearth	26.66± 0.82	34.71±1.72	27.54±0.67	27.18±0.64	27.20±0.62	26.73±0.86	27.18± 0.64
hearts	27.59± 0.73	39.59±2.02	27.72±0.81	27.31±0.82	27.31±0.82	27.11±0.82	27.31± 0.82
hepatitis	26.57± 1.02	32.55±2.16	27.05±1.47	26.62±1.25	26.62±1.25	26.64±1.29	26.62± 1.25
hypo	18.90± 2.47	26.16±2.71	19.42±2.25	20.43±1.74	20.13±1.75	18.77±1.92	20.38± 1.66
ionosphere	24.30± 1.64	42.49±2.45	25.32±1.43	24.63±1.72	24.63±1.72	24.22±1.24	24.72± 1.70
iris	12.05±1.54	14.21±4.30	11.03± 1.89	10.90±1.59	10.90±1.59	10.90±1.59	10.90± 1.59
krkcp	17.60±2.36	27.41±4.42	17.38±2.48	17.37±1.95	17.37±1.95	17.37±1.95	17.12± 2.00
labor	16.30±2.58	24.39±3.43	15.27± 1.75	16.02±2.75	16.02±2.75	15.60±2.32	16.02± 2.75
letter	53.62±0.19	70.21±0.40	53.29±0.23	52.67±0.20	52.50± 0.20	52.54±0.22	52.50± 0.19
lymph	31.01± 1.47	42.69±2.05	31.83±1.70	31.72±1.75	31.69±1.72	31.89±1.82	31.69± 1.72
mushroom	0.00± 0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00± 0.00
primary	55.49±0.51	61.42±0.47	54.87±0.54	54.76± 0.47	54.54±0.48	54.70±0.56	54.56± 0.49
segment	32.22± 0.78	51.44±1.61	32.53±0.73	32.36±0.79	32.38±0.81	32.26±0.81	32.41± 0.82
sick	27.08±1.08	38.84±1.95	26.46±0.91	26.63±0.99	26.65±1.00	26.41± 1.00	26.63± 0.99
sonar	35.04±0.41	49.90±1.40	34.48± 0.51	34.84±0.54	34.84±0.54	34.83±0.49	34.82± 0.54
soybean	38.41±0.73	49.45±1.68	36.49±1.09	36.10±0.72	35.81±0.81	36.08±0.95	35.72± 0.73
splice	27.16±0.69	36.85±0.98	24.38±0.38	23.76± 0.49	23.81±0.50	24.08±0.47	23.81± 0.45
vehicle	37.94±0.48	48.58±0.71	37.57± 0.57	37.39±0.52	37.28±0.49	37.32±0.49	37.28± 0.51
vote	18.13± 1.96	22.49±3.35	17.49±2.47	17.79±2.03	17.79±2.03	17.95±2.58	17.79± 2.03
vowel	49.91±0.49	69.60±0.37	50.29±0.60	49.29±0.54	49.11±0.53	49.16±0.46	49.07± 0.52
waveform	34.01±0.11	50.02±0.44	33.92±0.19	33.47±0.15	33.47±0.14	33.46± 0.16	33.46± 0.14
zoo	16.74±2.82	31.35±5.90	15.63± 2.96	15.52±2.38	15.52±2.38	15.53±2.39	16.02± 2.44
Mean	29.12	39.54	28.78	28.57	28.50	28.48	28.50

Table 20: Comparison of \overline{D}_T .

Dataset	DEM	ARIST	PROB	PR-SQ	PR-EXP	PR-SQN	PR-EXPN
anneal	18.22±1.91	26.79±4.26	16.44±2.38	15.91±2.18	15.91±2.18	16.48±2.26	15.41± 2.02
audiology	15.05±1.04	21.98±2.30	14.53±0.75	13.94±1.09	13.65±0.92	14.28±0.96	13.65± 0.92
autos	23.39±0.97	34.78±1.27	20.81±1.15	20.62±1.28	20.66±1.31	21.41±1.21	20.62± 1.27
balance	18.19±0.50	25.27±1.39	17.84±0.55	17.10±0.36	17.31±0.39	17.49±0.41	17.33± 0.38
breastc	27.12±0.37	33.84±1.01	26.82±0.31	27.07±0.31	27.09±0.30	26.78±0.26	27.04± 0.31
breastw	23.55±1.07	36.36±2.02	22.91±1.24	22.29±1.17	22.29±1.17	22.26±1.19	22.29± 1.17
colic	18.26±0.68	22.43±1.62	18.45±0.60	18.64±0.66	18.64±0.66	18.54±0.60	18.64± 0.63
credita	25.11±0.75	34.50±1.41	25.24±0.72	25.18±0.59	25.16±0.59	24.91±0.55	25.09± 0.66
creditg	29.45±0.31	39.05±0.82	29.68±0.25	29.41±0.34	29.38±0.32	29.30±0.49	29.39± 0.37
diabetes	29.23±0.47	39.84±0.87	29.52±0.50	28.96±0.45	28.96±0.45	28.91±0.42	28.92± 0.46
glass	22.50±0.71	33.78±1.53	22.66±1.05	21.66±0.95	21.65±0.94	21.63±0.88	21.59± 0.94
heartc	29.31±0.84	40.96±1.07	29.81±0.88	29.63±0.68	29.63±0.68	29.50±0.69	29.63± 0.68
hearth	26.66±0.82	34.71±1.72	27.54±0.67	27.18±0.64	27.20±0.62	26.73±0.86	27.18± 0.64
hearts	27.59±0.73	39.59±2.02	27.72±0.81	27.31±0.82	27.31±0.82	27.11±0.82	27.31± 0.82
hepatitis	26.57±1.02	32.55±2.16	27.05±1.47	26.62±1.25	26.62±1.25	26.64±1.29	26.62± 1.25
hypo	16.21±2.58	22.85±2.70	16.61±2.26	17.65±1.78	17.45±1.80	16.20±1.95	17.69± 1.76
ionosphere	24.30±1.64	42.49±2.45	25.32±1.43	24.63±1.72	24.63±1.72	24.22±1.24	24.72± 1.70
iris	11.97±1.53	14.12±4.29	10.95±1.89	10.82±1.60	10.82±1.60	10.82±1.60	10.82± 1.60
krkcp	17.60±2.36	27.41±4.42	17.38±2.48	17.37±1.95	17.37±1.95	17.37±1.95	17.12± 2.00
labor	16.30±2.58	24.39±3.43	15.27±1.75	16.02±2.75	16.02±2.75	15.60±2.32	16.02± 2.75
letter	19.40±0.16	37.04±0.34	19.43±0.20	18.80±0.21	18.70±0.19	18.81±0.18	18.70± 0.19
lymph	24.75±1.21	36.48±1.56	25.27±0.83	25.44±1.04	25.39±1.00	25.61±1.01	25.39± 1.00
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00± 0.00
primary	9.50±0.24	12.35±0.46	9.41±0.35	9.25±0.34	9.21±0.32	9.26±0.31	9.23± 0.32
segment	23.29±0.91	42.89±1.70	23.74±0.77	23.69±0.90	23.69±0.90	23.64±0.90	23.71± 0.91
sick	27.08±1.08	38.84±1.95	26.46±0.91	26.63±0.99	26.65±1.00	26.41±1.00	26.63± 0.99
sonar	35.04±0.41	49.90±1.40	34.48±0.51	34.84±0.54	34.84±0.54	34.83±0.49	34.82± 0.54
soybean	25.34±1.01	36.45±1.44	24.00±1.40	23.98±1.14	23.86±1.21	24.00±1.15	23.81± 1.22
splice	23.63±0.86	32.87±0.92	20.80±0.53	20.23±0.64	20.26±0.66	20.57±0.66	20.27± 0.61
vehicle	28.46±0.41	38.03±0.37	28.13±0.49	28.04±0.36	28.02±0.34	28.01±0.35	27.99± 0.37
vote	18.13±1.96	22.49±3.35	17.49±2.47	17.79±2.03	17.79±2.03	17.95±2.58	17.79± 2.03
vowel	29.66±0.54	49.07±0.62	29.53±0.42	28.33±0.62	28.16±0.61	28.31±0.64	28.18± 0.62
waveform	33.12±0.12	48.74±0.40	33.02±0.19	32.59±0.15	32.60±0.15	32.59±0.15	32.59± 0.14
zoo	11.81±2.26	25.24±4.90	10.83±2.30	11.02±2.09	11.02±2.09	11.02±2.09	11.52± 2.05
Mean	22.23	32.30	21.92	21.73	21.70	21.68	21.70

Table 21: Comparison of \overline{D}_F .

GENERALIZED UNIFIED DECOMPOSITION OF ENSEMBLE LOSS

Dataset	DEM	ARIST	PROB	PR-SQ	PR-EXP	PR-SQN	PR-EXPN
anneal	1.27±0.06	1.62±0.08	1.26±0.06	1.26±0.06	1.26±0.06	1.26±0.06	1.26±0.06
audiology	17.77±0.39	20.06±0.66	17.70±0.38	17.63±0.38	17.62±0.38	17.64±0.38	17.62±0.38
autos	25.66±0.51	30.22±0.58	25.52±0.53	25.40±0.52	25.38±0.52	25.41±0.52	25.38±0.52
balance	17.45±0.20	19.86±0.40	17.46±0.19	17.35±0.19	17.34±0.19	17.35±0.19	17.34±0.19
breastc	24.76±0.26	27.32±0.52	24.66±0.24	24.57±0.25	24.57±0.25	24.59±0.25	24.57±0.25
breastw	4.44±0.10	5.36±0.19	4.44±0.10	4.43±0.10	4.43±0.10	4.43±0.10	4.43±0.10
colic	14.12±0.28	14.72±0.40	14.05±0.28	14.04±0.28	14.04±0.28	14.05±0.28	14.04±0.28
credita	11.95±0.13	13.99±0.33	11.92±0.13	11.86±0.12	11.86±0.12	11.87±0.12	11.86±0.12
creditg	22.21±0.20	26.29±0.43	22.12±0.21	21.97±0.20	21.97±0.20	21.99±0.20	21.97±0.20
diabetes	20.15±0.20	24.72±0.35	20.13±0.19	19.98±0.19	19.98±0.19	19.99±0.19	19.98±0.19
glass	25.52±0.36	31.46±0.65	25.61±0.37	25.43±0.35	25.42±0.35	25.43±0.35	25.42±0.35
heartc	18.70±0.20	22.47±0.45	18.66±0.21	18.54±0.20	18.54±0.20	18.55±0.20	18.54±0.20
hearth	16.11±0.29	18.77±0.60	16.12±0.29	16.03±0.28	16.03±0.28	16.05±0.28	16.03±0.28
hearts	16.84±0.32	20.83±0.67	16.80±0.31	16.72±0.31	16.72±0.31	16.72±0.31	16.72±0.31
hepatitis	15.24±0.65	17.22±0.90	15.08±0.64	15.02±0.62	15.02±0.62	15.02±0.62	15.02±0.62
hypo	0.43±0.02	0.49±0.02	0.43±0.02	0.43±0.02	0.43±0.02	0.43±0.02	0.43±0.02
ionosphere	8.17±0.18	10.58±0.39	8.18±0.18	8.15±0.18	8.15±0.18	8.15±0.18	8.15±0.18
iris	3.12±0.17	3.89±0.36	3.12±0.17	3.12±0.17	3.12±0.17	3.12±0.17	3.12±0.17
krkp	0.70±0.02	0.83±0.03	0.70±0.02	0.70±0.02	0.70±0.02	0.70±0.02	0.70±0.02
labor	19.31±0.83	21.10±0.95	19.18±0.84	19.09±0.81	19.09±0.81	19.10±0.81	19.09±0.81
letter	13.50±0.06	16.48±0.12	13.50±0.06	13.44±0.06	13.43±0.06	13.43±0.06	13.43±0.06
lymph	17.88±0.34	21.72±0.58	17.79±0.37	17.69±0.36	17.69±0.37	17.70±0.36	17.69±0.37
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
primary	48.31±0.45	52.38±0.36	47.83±0.44	47.79±0.41	47.64±0.43	47.72±0.44	47.63±0.43
segment	3.52±0.08	4.54±0.15	3.53±0.08	3.51±0.08	3.51±0.08	3.51±0.08	3.51±0.08
sick	1.16±0.04	1.40±0.06	1.16±0.04	1.16±0.04	1.16±0.04	1.16±0.04	1.16±0.04
sonar	23.59±0.40	30.40±0.78	23.58±0.39	23.48±0.39	23.48±0.39	23.48±0.39	23.48±0.39
soybean	11.19±0.18	12.87±0.25	11.12±0.18	11.09±0.18	11.08±0.18	11.09±0.18	11.08±0.18
splice	6.41±0.09	7.36±0.10	6.29±0.07	6.27±0.07	6.27±0.07	6.28±0.07	6.27±0.07
vehicle	21.69±0.27	26.60±0.37	21.71±0.27	21.55±0.27	21.54±0.27	21.55±0.27	21.54±0.27
vote	4.53±0.06	4.79±0.10	4.51±0.07	4.50±0.07	4.50±0.07	4.51±0.06	4.50±0.07
vowel	25.20±0.24	31.63±0.49	25.26±0.24	25.12±0.23	25.11±0.23	25.11±0.23	25.11±0.23
waveform	19.90±0.08	25.76±0.19	19.91±0.09	19.82±0.08	19.82±0.08	19.82±0.08	19.82±0.08
zoo	6.38±0.44	8.31±0.70	6.31±0.44	6.26±0.41	6.26±0.41	6.26±0.41	6.26±0.41
Mean	14.33	16.94	14.28	14.22	14.21	14.22	14.21

Table 22: Comparison of \overline{D}_P .

Dataset	DEM	ARIST	PROB	PR-SQ	PR-EXP	PR-SQN	PR-EXPN
anneal	1.23±0.06	1.30±0.15	1.10±0.06	1.06±0.07	1.06±0.07	1.08±0.07	1.05±0.07
audiology	20.01±0.84	22.44±1.08	19.44±1.02	19.12±0.84	19.08±0.87	19.65±0.83	19.08±0.87
autos	20.65±0.78	26.30±0.89	19.49±1.02	19.62±1.12	19.43±1.08	19.77±0.95	19.48±1.08
balance	17.44±0.48	19.37±0.65	17.16±0.51	16.76±0.33	16.93±0.34	17.11±0.33	16.98±0.31
breastc	26.55±0.54	30.27±1.12	26.40±0.55	26.72±0.29	26.75±0.31	26.37±0.41	26.68±0.31
breastw	4.15±0.14	5.15±0.21	4.02±0.21	4.01±0.24	4.01±0.24	3.99±0.24	4.01±0.24
colic	14.67±0.17	15.95±0.48	14.89±0.27	14.97±0.24	14.97±0.24	14.86±0.21	14.97±0.23
credita	13.55±0.37	15.55±0.49	13.72±0.44	13.74±0.43	13.72±0.41	13.61±0.41	13.70±0.45
creditg	25.90±0.40	30.05±0.84	26.42±0.34	26.40±0.42	26.36±0.41	26.24±0.56	26.38±0.44
diabetes	24.86±0.57	28.54±0.88	25.25±0.61	24.97±0.48	24.97±0.49	24.85±0.54	24.92±0.52
glass	27.41±1.06	32.46±1.60	27.27±0.84	26.61±1.16	26.61±1.23	26.61±1.06	26.52±1.17
heartc	20.79±0.61	25.70±1.10	21.15±0.64	21.33±0.62	21.33±0.62	21.23±0.58	21.33±0.62
hearth	19.80±0.70	21.84±0.83	20.33±0.74	20.37±0.68	20.37±0.64	19.96±0.62	20.37±0.68
hearts	18.96±0.73	22.70±0.89	19.30±0.97	19.15±0.92	19.15±0.92	19.04±0.87	19.15±0.92
hepatitis	19.04±0.62	19.61±1.28	19.23±0.76	19.18±0.71	19.18±0.71	19.24±0.80	19.18±0.71
hypo	0.43±0.02	0.46±0.03	0.45±0.02	0.46±0.01	0.45±0.01	0.44±0.02	0.46±0.02
ionosphere	7.78±0.36	10.31±0.85	8.46±0.37	8.04±0.38	8.04±0.38	8.01±0.34	8.10±0.36
iris	5.93±0.46	5.67±0.59	5.67±0.50	5.73±0.42	5.73±0.42	5.73±0.42	5.73±0.42
krkcp	0.62±0.03	0.72±0.07	0.59±0.03	0.59±0.03	0.59±0.03	0.59±0.03	0.59±0.03
labor	14.07±1.83	17.00±1.76	13.30±1.28	13.73±1.84	13.73±1.84	13.73±1.84	13.73±1.84
letter	7.63±0.08	12.67±0.11	7.55±0.07	7.44±0.09	7.39±0.09	7.42±0.08	7.39±0.09
lymph	20.66±0.85	24.03±1.16	20.67±0.75	21.08±0.97	21.00±0.92	21.13±0.76	21.00±0.92
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
primary	58.17±0.70	62.09±0.80	58.55±0.84	58.07±0.77	58.10±0.82	58.20±0.62	58.16±0.79
segment	2.69±0.08	4.10±0.23	2.69±0.11	2.72±0.09	2.72±0.09	2.72±0.09	2.72±0.09
sick	1.16±0.04	1.38±0.09	1.14±0.04	1.15±0.04	1.15±0.05	1.14±0.05	1.15±0.04
sonar	22.42±1.31	30.23±1.39	22.19±1.19	22.34±1.08	22.34±1.08	22.29±1.02	22.29±1.04
soybean	8.05±0.32	10.04±0.63	7.57±0.34	7.54±0.31	7.49±0.31	7.61±0.33	7.48±0.31
splice	6.15±0.17	6.93±0.16	5.73±0.11	5.64±0.13	5.65±0.14	5.69±0.13	5.65±0.13
vehicle	26.25±0.38	29.60±0.52	25.74±0.43	26.00±0.37	25.91±0.37	25.95±0.37	25.91±0.37
vote	3.70±0.14	4.07±0.27	3.72±0.20	3.81±0.15	3.81±0.15	3.88±0.19	3.81±0.15
vowel	10.97±0.43	22.56±0.74	10.77±0.46	10.27±0.45	10.15±0.41	10.22±0.41	10.14±0.42
waveform	18.72±0.10	25.28±0.24	18.59±0.16	18.38±0.19	18.39±0.19	18.37±0.20	18.37±0.19
zoo	6.56±0.81	8.54±1.15	6.26±0.85	6.26±0.73	6.26±0.73	6.26±0.73	6.36±0.74
Mean	14.62	17.44	14.55	14.51	14.49	14.50	14.49

Table 23: Comparison of L .

GENERALIZED UNIFIED DECOMPOSITION OF ENSEMBLE LOSS

Dataset	DEM	ARIST	PROB	PR-SQ	PR-EXP	PR-SQN	PR-EXPN
anneal	1.09±0.06	1.05±0.15	0.99±0.06	0.96±0.07	0.96±0.07	0.98±0.07	0.95± 0.07
audiology	20.14±0.84	22.42±1.08	19.62±1.02	19.33±0.84	19.23±0.87	19.74±0.83	19.23± 0.87
autos	20.96±0.78	26.43±0.89	19.50±1.02	19.63±1.12	19.50±1.08	19.97±0.95	19.53± 1.08
balance	17.47±0.48	19.32±0.65	17.17±0.51	16.78±0.33	16.96±0.34	17.12±0.33	17.00± 0.31
breastc	26.56±0.54	29.89±1.12	26.41±0.55	26.78±0.29	26.81±0.31	26.42±0.41	26.74± 0.31
breastw	4.12±0.14	5.01±0.21	4.00±0.21	3.96±0.24	3.96±0.24	3.95±0.24	3.96± 0.24
colic	14.74±0.17	15.95±0.48	14.95±0.27	15.05±0.24	15.05±0.24	14.95±0.21	15.05± 0.23
credita	13.64±0.37	15.55±0.49	13.80±0.44	13.84±0.43	13.82±0.41	13.70±0.41	13.79± 0.45
creditg	25.95±0.40	30.10±0.84	26.47±0.34	26.46±0.42	26.42±0.41	26.29±0.56	26.44± 0.44
diabetes	24.96±0.57	28.60±0.88	25.33±0.61	25.06±0.48	25.05±0.49	24.95±0.54	25.00± 0.52
glass	27.88±1.06	32.79±1.60	27.76±0.84	27.06±1.16	27.05±1.23	27.01±1.06	26.94± 1.17
heartc	20.63±0.61	25.74±1.10	21.07±0.64	21.27±0.62	21.27±0.62	21.16±0.58	21.27± 0.62
hearth	19.85±0.70	21.81±0.83	20.46±0.74	20.44±0.68	20.44±0.64	20.03±0.62	20.44± 0.68
hearts	19.13±0.73	22.65±0.89	19.39±0.97	19.25±0.92	19.25±0.92	19.11±0.87	19.25± 0.92
hepatitis	18.89±0.62	19.53±1.28	19.28±0.76	19.14±0.71	19.14±0.71	19.17±0.80	19.14± 0.71
hypo	0.41±0.02	0.43±0.03	0.43±0.02	0.44±0.01	0.44±0.01	0.42±0.02	0.44± 0.02
ionosphere	7.71±0.36	10.30±0.85	8.26±0.37	7.93±0.38	7.93±0.38	7.86±0.34	7.98± 0.36
iris	5.53±0.46	5.07±0.59	5.31±0.50	5.35±0.42	5.35±0.42	5.35±0.42	5.35± 0.42
krkcp	0.60±0.03	0.66±0.07	0.58±0.03	0.58±0.03	0.58±0.03	0.58±0.03	0.58± 0.03
labor	11.66±1.83	13.20±1.76	11.19±1.28	11.72±1.84	11.72±1.84	11.64±1.84	11.72± 1.84
letter	7.63±0.08	12.67±0.11	7.55±0.07	7.44±0.09	7.39±0.09	7.42±0.08	7.39± 0.09
lymph	20.27±0.85	23.90±1.16	20.48±0.75	20.91±0.97	20.84±0.92	21.00±0.76	20.84± 0.92
mushroom	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00± 0.00
primary	58.17±0.70	62.03±0.80	58.54±0.84	58.09±0.77	58.10±0.82	58.21±0.62	58.16± 0.79
segment	2.69±0.08	4.05±0.23	2.70±0.11	2.73±0.09	2.73±0.09	2.73±0.09	2.73± 0.09
sick	1.17±0.04	1.37±0.09	1.14±0.04	1.15±0.04	1.15±0.05	1.14±0.05	1.15± 0.04
sonar	22.59±1.31	30.31±1.39	22.22±1.19	22.52±1.08	22.52±1.08	22.47±1.02	22.47± 1.04
soybean	8.08±0.32	9.96±0.63	7.59±0.34	7.59±0.31	7.54±0.31	7.64±0.33	7.52± 0.31
splice	6.15±0.17	6.91±0.16	5.74±0.11	5.65±0.13	5.65±0.14	5.70±0.13	5.66± 0.13
vehicle	26.31±0.38	29.68±0.52	25.79±0.43	26.06±0.37	25.97±0.37	26.00±0.37	25.97± 0.37
vote	3.57±0.14	3.84±0.27	3.56±0.20	3.64±0.15	3.64±0.15	3.68±0.19	3.64± 0.15
vowel	10.95±0.43	22.54±0.74	10.77±0.46	10.29±0.45	10.16±0.41	10.23±0.41	10.15± 0.42
waveform	18.72±0.10	25.25±0.24	18.59±0.16	18.38±0.19	18.39±0.19	18.37±0.20	18.38± 0.19
zoo	5.57±0.81	6.73±1.15	5.37±0.85	5.38±0.73	5.38±0.73	5.38±0.73	5.47± 0.74
Mean	14.52	17.23	14.47	14.44	14.42	14.42	14.42

Table 24: Comparison of L^* .

6. Conclusions

A generalized unified decomposition of ensemble loss for predicting ensemble performance was presented that generalizes the unified decomposition presented in UDEL by allowing for classifier weights and probabilistic voting schemes. While UDEL was limited to uniform weights and democratic voting schemes, where the various base classifiers each can vote for a single class once only, this article extends the result to more general voting schemes. Experimental results suggest that democratic voting can be outperformed by probabilistic and progressive voting schemes, hence a GUEDEL is worth exploring. Specifically probabilistic voting schemes are shown to provide a real advantage over democratic voting schemes in many datasets. So they should be tried more often in real world problem solving situations. In addition the formalization of ensembles presented gave insight into why the voting schemes provided such differing results.

References

- [1] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: bagging boostin and variants. *Machine learning*, 36(1-2): 105-139, 1999.
- [2] L. Breiman. Bagging predictors. *Machine Learning* 24(2): 123-140, 1996.
- [3] L. Breiman, L. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California at Berkeley, 1996.
- [4] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems, First International Workshop, MCS 2000, June 2000, Cagliari, Italy, 2000*.
- [5] P. Domingos. A Unified Bias-Variance Decomposition and its Applications. Technical report. Department of Computer Science and Engineering, University of Washington, 2000.
- [6] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119-139, 1997.
- [7] M. Goebel, P. Riddle, M. Barley. A unified decomposition of ensemble loss for predicting ensemble performance. *ICML 2002*.
- [8] L. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. Pattern analysis and Machine Intelligence*, 12: 993–1003, 1990.
- [9] R. Kohavi and D. Wolpert. Bias plus variance decomposition for zero-one loss functions. *ICML*, 275–283, 1996.
- [10] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. *NIPS*, 7: 231–238, 1995.
- [11] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11: 169–198, August 1999.
- [12] Fabio Roli, Giorgio Giacinto, and Gianni Vernazza. Methods for designing multiple classifier systems. *Proc. Multiple Classifier Systems*, 78–87, 2001.
- [13] R. Tibshirani. and K. Knight. Model search and inference by bootstrap “bumping”. University of Toronto technical report. 1995.
- [14] I.H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, 2000.
- [15] D.H. Wolpert. Stacked generalization. *Neural Networks*, 5:241-259, 1992.