Comparing Methods for Generating Diverse Ensembles of Artificial Neural Networks

T. Löfström, U. Johansson and H. Boström

***Abstract*—It is well-known that ensemble performance relies heavily on sufficient diversity among the base classifiers. With this in mind, the strategy used to balance diversity and base classifier accuracy must be considered a key component of any ensemble algorithm. This study evaluates the predictive performance of neural network ensembles, specifically comparing straightforward techniques to more sophisticated. In particular, the sophisticated methods GASEN and NegBagg are compared to more straightforward methods, where each ensemble member is trained independently of the others. In the experimentation, using 31 publicly available data sets, the straightforward methods clearly outperformed the sophisticated methods, thus questioning the use of the more complex algorithms.**

I. Introduction

ONE primary target for predictive data mining is to obtain high accuracy, i.e., making few misclassifications when applying a developed model on novel data. A well-established finding in the machine learning community is that ensembles, i.e., collections of base classifiers that are combined using some combination scheme, all but guarantees a higher accuracy than using any specific base classifier alone. The main reason for this is that the effect of uncorrelated errors made by the ensemble members tends to be neutralized by the other members [1]. For this neutralization to take effect, it is, however, necessary that the base classifiers make their errors more or less independently. Obviously, combining identical models will not improve the performance. The term *diversity* is often used to describe this property of the ensemble, i.e., the degree to which the base classifiers commit their errors on different instances. A taxonomy of methods for creating diversity in ensembles was introduced by Brown et al. [2]. There, the main distinction is made between *explicit methods*, where diversity is optimized, and *implicit methods*, where diversity is not targeted directly. One of the most straightforward implicit methods to introduce diversity in ensembles of artificial neural networks (ANNs) is to use randomized initial weights. This is, however, normally not a very successful method, since most networks tend to converge to the same or very similar local minima. Other, more effective, implicit methods for creating diverse ANN ensembles include varying the ANN architecture or using different sets of data for training, e.g., using *bagging* [3].

In contrast to implicit methods, where each ensemble member is generated independently of the others, explicit methods have to consider multiple ensemble members simultaneously, i.e., the generation or inclusion of each member is dependent on other members in the ensemble. As a consequence, the explicit methods are often more complex, both conceptually and computationally, than the implicit ones.

The main purpose of this study is to compare the predictive performance of several well-known implicit and explicit methods for generating diverse ensembles, specifically investigating whether more complex schemes really outperform straightforward methods, when applied to real-world data sets. In addition, the question of how diversity is affected by the different algorithms, and how diversity may affect the overall ensemble performance, will be analyzed.

In the next section, we briefly describe the implicit and explicit methods for generating diverse ensembles of ANNs that are considered in this study. In section 3, we describe the experimental setup, while the results are shown and discussed in section 4. In section 5, we summarize the main conclusions and discuss future work.

II. Background

Unfortunately, the relationship between ensemble accuracy and diversity is not fully understood, making it hard to utilize diversity for ensemble creation in an effective way. In [4], Krogh and Vedelsby derived the result that the ensemble error depends both on the average base classifier accuracy and on their diversity (ambiguity). More formally, the ensemble error, E, is:

$$E = \bar{E} - \bar{A} \qquad (1)$$

where $\bar{E}$ is the average error of the base classifiers and $\bar{A}$ is the ensemble diversity, measured as the average of the squared differences in the predictions of the base classifiers and the ensemble. Since diversity is always positive, this decomposition shows that the ensemble will always have lower error than the average error obtained by the individual model. The two terms are, however, normally highly correlated, making it necessary to balance them instead of

This work was supported by the Information Fusion Research Program (University of Skövde, Sweden) in partnership with the Swedish Knowledge Foundation under grant 2003/0104 (URL: http://www.infofusion.se).

T. Löfström, CSLABS, School of Business and Informatics, University of Borås, Sweden. SE-50190, Sweden. Phone: +46334354236. (Email: tuve.lofstrom@hb.se).

U. Johansson, CSLABS, School of Business and Informatics, University of Borås, Sweden. (Email: ulf.johansson@hb.se).

H.Boström, Dept. of Computer and Systems Sciences, Stockholm University, Sweden. (Email: henrik.bostrom@dsv.su.se).

just maximizing the diversity term.

For classification, where a zero-one loss function is used, it is, however, not possible to decompose the ensemble error into error rates of the individual classifiers and a diversity term. Instead, algorithms for ensemble creation typically use heuristic expressions trying to approximate the unknown diversity term. Naturally, the goal is to find a diversity measure highly correlated with majority vote test accuracy.

Until now, many diversity measures have been proposed for classification. Specifically, ten different diversity measures have been extensively evaluated and analyzed, both theoretically and empirically, throughout the years. These ten measures can be divided into pair-wise and non pair-wise measures. Saitta has showed theoretically why several of these measures cannot work in practice [5]. In addition, several empirical studies have confirmed that the correlation between most of the proposed diversity measures and ensemble test accuracy is low to non-existent, see e.g. [6] and [7]. However, Johansson et al. showed in [7], that a few diversity measures, when measured on a validation set, could be useful. Specifically, ranking ensembles based on results on a validation set, using the two diversity measures *double faul*t [8] and *difficulty* [9], turned out to have some merit. As a matter of fact, the correlations between these diversity measures, obtained on the validation set, and the ensemble accuracy on the test set were comparable to the correlations between ensemble validation and ensemble test accuracies. Consequently, even though no diversity measure in itself is the golden key to optimal ensemble performance, it can in practice be useful when combined with performance measures such as base classifier accuracy, as illustrated in e.g. [10].

Many ensemble techniques divide the available data and train each ensemble member using partly different training sets. These sets can be generated in different ways, but in this study we will only consider the standard *bagging* procedure [3]. More specifically, in bagging, diversity is achieved by training each base model using different training sets obtained through sampling. Each training set (called *bootstrap*) consists of the same number of instances as the entire set of data available for training, and it is created using sampling with replacement according to a uniform distribution. Instances may hence appear more than once in a bootstrap, since they are drawn with replacement, with the result that approximately 63% of available instances are included in each bootstrap.

In addition to using bagging, implicit diversity is in this study also achieved by varying the number of hidden nodes in the ANNs, similar to Johansson et al. [11]. The number of hidden layers is fixed to one for all ANNs, and the exact number of units in the hidden layer is randomized, based on the number of inputs and classes in the current data set (see next section for details).

Two sophisticated ensemble algorithms, are also evaluated in the study: GASEN, which is a well-known algorithm for searching for optimal ensembles by means of genetic algorithms [12], and the recently proposed explicit ensemble learning algorithm NegBagg [13]. GASEN is actually a post-processing technique, since several ANNs are trained before a genetic algorithm is used to select an optimal subset of the individual networks. The optimization criterion (the fitness) boils down to accuracy on a holdout (validation) set. The number of ANNs in the ensemble can vary, since all ANNs with strength values higher than a specific threshold (which is a pre-set parameter) will be included in the ensemble.

The second sophisticated method, NegBagg, incrementally trains different individual ANNs in an ensemble using the negative correlation learning algorithm [14]. Bagging is used in NegBagg to create different training sets for different ANNs in the ensemble. The idea behind using negative correlation learning in conjunction with bagging is to facilitate interaction and cooperation among ANNs during their training. NegBagg uses a constructive approach to automatically determine the number of hidden neurons for ANNs. The constructive approach uses a test on a validation set to find out whether any particular network should be expanded. This is tested repeatedly throughout the simultaneous training of the ANNs. Furthermore, the degree to which negative correlation learning is used can be regulated by a parameter, even to the extent that NegBagg is trained without any negative correlation learning at all.

III. METHOD

The purpose of this study is to compare the predictive performance, here measured by accuracy, of sophisticated ensemble algorithms to more straightforwardly built ensembles. In addition, the question of how diversity is affected by these different algorithms, and how diversity may affect the overall ensemble performance, will be analyzed.

NegBagg and GASEN are sophisticated ensemble algorithms incorporating several interesting and advanced features. These will be compared to two more straightforward ways of generating ANN ensembles, by which diversity is targeted implicitly by varying the number of hidden nodes in the networks. The only difference between the two straightforward methods is whether bagging is used or not. The use of a varied architecture could be more effective than using bagging, according to [11]. In this study, the architecture is varied with respect to the number of hidden nodes only, while the number of hidden layers is fixed to one for all ANNs. The exact number of units in the hidden layer is slightly randomized, and is based on the number of inputs and classes in the current data set. The number of hidden units is determined from (2) below.

$$h = \lfloor 2 \cdot rand \cdot \sqrt{v \cdot c} \rfloor \quad (2)$$

where v is the number of input variables and c is the number of classes. rand is a random number in the interval [0, 1].

All originally generated ensembles consist of 15 ANNs, but it should be noted that the result of using GASEN for

post-processing often results in smaller ensembles. During experimentation, GASEN was applied as a post-processing technique on the two straightforwardly generated ensembles (using implicit diversity only), thus resulting in ensembles consisting of subsets of the 15 available ANNs. In summary, the evaluated algorithms are:

1) 15 single layered MLPs with randomized number of units, where each network is trained using all available data (Ens)
2) GASEN applied to Ens, i.e., selecting a subset of the members in Ens (GAS)
3) 15 single layered MLPs with randomized number of units, where each network is trained on a bootstrap (Ens B)
4) GASEN applied to Ens B (GAS B).
5) NegBagg with 15 ANNs, without negative correlation learning (NB-NC)
6) NegBagg with 15 ANNs, with negative correlation learning (NB+NC)

Ens and Ens B will be referred to as the *ordinary ensembles*.

All networks used are feed-forward neural networks. The implementation used in the experiments is an extended version of networks using the backpropagation algorithm as implemented in [15]. The extensions include support for negative correlation learning as described in [14]. The parameters used in the experiments are listed in Table I. See the original papers for explanations on the parameters specific to GASEN and NegBagg.

TABLE I
PARAMETERS USED IN EXPERIMENTS

Parameter	Experiment					
	Ens	GAS	Ens B	GAS B	NB-NC	NB+NC
Learning Rate	0.5	0.5	0.5	0.5	0.5	0.5
Momentum	0.9	0.9	0.9	0.9	0.9	0.9
Max epochs	1000	1000	1000	1000	-	-
Bagging	No	No	Yes	Yes	Yes	Yes
Lambda	0	0	0	0	0	1
Train epochs	-	-	-	-	10	10
Threshold	-	1/15	-	1/15	-	-
Pop Size	-	100	-	100	-	-
Generations	-	100	-	100	-	-

In this study, eight different chemoinformatics data sets from the study of Bruce et al. [16] are used. The same data sets were originally used by Sutherland et al. [17]. It should be noted that Bruce et al. used two separate feature sets, *2.5D descriptors* (2.5D) and *linear fragment descriptors* (Frags.) to characterize the chemical structures in the data sets. Following Bruce et al., we too evaluate both these feature sets in this study. Table II below summarizes the 16 binary data sets.

TABLE II.
CHARACTERISTICS OF MEDICINAL CHEMISTRY DATA SETS USED

Data set	Instances	Attributes	
		2.5D	Frags
ACE	114	56	1024
AchE	111	63	774
BZR	163	75	832
COX2	322	74	660
DHFR	397	70	951
GPB	66	70	692
THER	76	64	575
THR	88	66	527

Another set of data sets are collected from the UCI repository [18], and represents problems from various domains. The characteristics of these data sets are presented in Table III.

TABLE III.
CHARACTERISTICS OF UCI DATA SETS USED

Data set	Classes	Instances	Attributes
breast-cancer	2	286	9
breast-w	2	699	9
cmc	3	1473	9
credit-a	2	690	15
cylinder	2	540	39
diabetes	2	768	8
haberman	2	306	3
heart-cleve	2	303	13
heart-statlog	2	270	13
ionosphere	2	351	34
liver-bupa	2	345	6
lymph	4	148	18
tae	3	151	5
tic-tac-toe	2	958	9
vehicle	4	846	18
wine	3	178	13

For all experiments, 10-fold cross validation has been employed, and the mean results over the ten folds are reported for all methods and datasets. The instances not used as test set were divided into a training set ($^2/_3$) and a validation set ($^1/_3$). Majority vote was used as combination strategy for the ensembles.

IV. RESULTS

The accuracy results are shown in Table IV, where bold face is used to indicate the best result achieved by any technique on a specific data set. The number of wins and the mean rank for each technique on all datasets are shown at the bottom of the table.

The most obvious observation is that the ordinary ensembles (Ens and Ens B), which are created using implicit diversity only, clearly outperformed the other methods. The differences in accuracy between the ordinary ensembles and the others are significant in all cases except when compared to GAS, using the evaluation procedure recommended in [19], i.e., a Friedman test followed by a Nemenyi post-hoc test (for which the critical difference in mean rank is 1.29). A one-tailed sign test between the ordinary ensembles and the corresponding GASEN ensembles results for Ens and GAS in a *p-value* of 0.052, while Ens B and GAS B results

in a *p-value* of 0.004. The ordinary ensembles are consequently significantly, or at least clearly better when all base classifiers are used than after GASEN has selected a subset of them. When investigating if there is any difference between the two versions of the NegBagg algorithm (with or without negative correlation learning) using a one-tailed sign test, the resulting *p-value* is 0.557. It hence seems that negative correlation learning only marginally impacts the performance of the algorithm.

Some of the data sets evaluated in this study have previously been evaluated with NegBagg in [13], where very different levels of accuracy were reported. In the earlier study, a completely different experimental setup was used. Instead of using cross validation, the algorithms were evaluated by repeatedly splitting each data set into training and test sets of equal size 30 times. That the observed differences in performance are indeed due to the difference in experimental setup, and not due to the implementation, was confirmed by mimicking the experiment in [13], giving results similar to the ones previously reported.

TABLE IV.
ACCURACY

Data set	Ens	GAS	Ens B	GAS B	NB-NC	NB+NC
ACE Fr	0.815	0.806	0.807	0.805	**0.823**	0.798
AchE Fr	**0.709**	0.691	0.664	0.645	**0.709**	0.700
BZR Fr	0.767	0.779	0.786	**0.787**	0.774	0.785
COX2 Fr	**0.717**	0.707	0.710	0.698	0.682	0.701
DHFR Fr	**0.853**	**0.853**	0.838	0.836	0.841	0.848
GPB Fr	**0.786**	0.755	0.755	0.755	0.712	0.726
THER Fr	0.750	0.698	**0.763**	0.736	0.752	0.752
THR Fr	**0.693**	0.692	0.678	0.608	0.608	0.507
ACE 2.5D	0.868	0.868	0.868	0.868	0.833	**0.877**
AchE 2.5D	0.700	0.691	**0.709**	0.645	0.673	0.645
BZR 2.5D	0.786	**0.792**	0.792	0.762	0.718	0.767
COX2 2.5D	**0.754**	**0.754**	0.748	0.735	0.713	0.741
DHFR 2.5D	0.831	**0.838**	0.818	0.811	0.818	0.798
GPB 2.5D	**0.776**	0.762	0.750	0.736	0.679	0.695
THER 2.5D	0.698	0.738	**0.754**	0.689	0.657	0.684
THR 2.5D	0.732	0.721	0.732	**0.743**	0.710	0.733
breast-cancer	**0.741**	0.734	0.727	**0.741**	0.720	0.734
breast-w	0.967	0.966	**0.968**	0.967	0.966	0.966
cmc	**0.535**	0.521	0.520	0.499	0.518	0.518
credit-a	**0.864**	0.859	0.859	0.862	0.862	**0.864**
cylinder	**0.752**	0.737	0.743	0.724	0.726	0.713
diabetes	0.761	0.759	0.763	**0.764**	0.756	0.747
haberman	0.722	0.729	**0.745**	0.735	0.735	0.735
heart-cleve	0.828	0.832	**0.835**	0.825	0.812	0.828
heart-statlog	0.830	0.830	**0.841**	0.837	0.833	0.830
ionosphere	**0.923**	0.920	0.906	0.906	0.906	0.897
liver-bupa	0.719	0.722	**0.728**	0.696	0.681	0.713
lymph	0.831	0.825	0.831	0.838	0.845	**0.859**
tae	0.527	0.507	**0.547**	0.520	0.520	0.493
tic-tac-toe	**0.983**	0.981	**0.983**	**0.983**	**0.983**	0.982
vehicle	0.812	**0.814**	0.806	0.793	0.793	0.793
wine	0.978	0.978	0.978	0.966	**0.983**	0.983
Wins	**13**	**5**	**10**	**5**	**4**	**3**
Mean Rank	**2.55**	**3.38**	**2.67**	**3.97**	**4.33**	**4.11**

In order to analyze whether the relative performances can be explained by the base classifier accuracy or the diversity alone, we investigate this further for the different methods. The mean base classifier accuracy (measured on the test folds) for each technique and data set is presented in Table V.

The mean base classifier accuracy of ensembles trained on the whole training set is clearly much higher than for ensembles trained on bootstraps, which should be no surprise. This shows that the overall ensemble accuracy cannot be explained by high base classifier accuracy alone, since neither Ens nor GAS outperformed Ens B and GAS B, respectively, when considering ensemble accuracy. The results in Table V also show that GAS has a preference for individual ANNs with higher accuracy. But, even though GAS often selects ANNs from among the more individually accurate members in the ensemble, it cannot compete when considering ensemble accuracy. NegBagg in general obtains lower base classifier accuracy than the ordinary ensembles and GASEN. This is hence one possible reason for why NegBagg does not achieve as high overall ensemble accuracy as the other methods.

TABLE V.
AVERAGE BASE CLASSIFIER ACCURACY

Data set	Ens	GAS	Ens B	GAS B	NB-NC	NB+NC
ACE Fr	0.788	**0.790**	0.773	0.758	0.737	0.723
AchE Fr	**0.684**	0.665	0.638	0.617	0.652	0.641
BZR Fr	0.743	**0.748**	0.716	0.722	0.707	0.705
COX2 Fr	0.691	**0.693**	0.673	0.661	0.616	0.623
DHFR Fr	0.830	**0.834**	0.815	0.814	0.778	0.786
GPB Fr	0.748	**0.750**	0.719	0.731	0.666	0.682
THER Fr	**0.723**	0.721	0.708	0.701	0.676	0.657
THR Fr	0.861	**0.864**	0.855	0.850	0.833	0.852
ACE 2.5D	0.676	**0.682**	0.653	0.651	0.632	0.609
AchE 2.5D	0.761	**0.762**	0.744	0.733	0.721	0.733
BZR 2.5D	**0.743**	0.741	0.716	0.707	0.683	0.683
COX2 2.5D	**0.817**	0.816	0.794	0.784	0.785	0.774
DHFR 2.5D	**0.743**	0.734	0.688	0.684	0.623	0.649
GPB 2.5D	**0.727**	0.725	0.707	0.694	0.659	0.665
THER 2.5D	0.612	**0.624**	0.574	0.562	0.541	0.530
THR 2.5D	**0.751**	0.747	0.702	0.692	0.696	0.664
breast-cancer	0.719	**0.720**	0.712	0.718	0.693	0.699
breast-w	0.966	**0.967**	0.962	0.961	0.963	0.964
cmc	**0.516**	0.506	0.484	0.479	0.489	0.496
credit-a	**0.857**	0.857	0.853	0.853	0.831	0.836
cylinder	0.728	**0.731**	0.705	0.689	0.662	0.668
diabetes	0.756	0.755	0.756	**0.757**	0.747	0.747
haberman	0.727	**0.729**	0.725	0.722	0.722	0.721
heart-cleve	**0.822**	0.819	0.814	0.811	0.798	0.813
heart-statlog	**0.823**	0.819	0.812	0.819	0.801	0.810
ionosphere	0.910	**0.911**	0.884	0.878	0.878	0.883
liver-bupa	0.703	**0.704**	0.675	0.663	0.661	0.681
lymph	**0.817**	0.817	0.805	0.809	0.810	0.806
tae	**0.507**	0.506	0.488	0.495	0.445	0.423
tic-tac-toe	0.971	0.967	0.975	0.971	0.978	**0.978**
vehicle	**0.805**	0.804	0.784	0.778	0.779	0.773
wine	**0.976**	0.975	0.972	0.971	0.970	0.970
Wins	**15**	**15**	**0**	**1**	**0**	**1**
Mean Rank	**1.66**	**1.70**	**3.47**	**4.19**	**5.06**	**4.92**

When considering ensemble accuracy there was a clear difference between the ordinary ensemble and the corresponding GASEN ensemble, in favor of the former. When making a similar analysis regarding base classifier accuracy, it turns out that there is no difference at all when considering ensembles trained without bagging. For ensembles trained with bagging the picture is quite different, and it turns out that the base classifier accuracy of the ordinary ensemble is significantly higher than for the corresponding GASEN ensemble (p = 0.003). This implies

that the way base classifiers are trained clearly affects the performance of GASEN.

Since base classifier accuracy alone cannot explain the performance of the ensemble, it is interesting to make an analysis of the diversity, to see to what extent diversity might explain the results.

The diversity of the generated ensembles (according to the diversity measure *double fault*) is presented in Table VI. *Double fault* is a pair-wise diversity measure which indicates the average proportion of instances misclassified by both classifiers [8]. A lower *double fault* diversity value indicates more diversity.

TABLE VI.
DOUBLE FAULT DIVERSITY

Data set	Ens	GAS	Ens B	GAS B	NB-NC	NB+NC
ACE Fr	0.141	0.144	0.131	**0.116**	0.126	0.148
AchE Fr	0.186	0.190	0.195	0.208	0.186	**0.179**
BZR Fr	0.170	0.164	**0.160**	0.168	0.163	0.164
COX2 Fr	0.212	0.209	0.199	0.199	**0.184**	0.186
DHFR Fr	0.114	0.117	0.114	0.112	0.104	**0.102**
GPB Fr	0.181	0.182	0.162	**0.162**	0.177	0.179
THER Fr	0.187	0.181	0.159	**0.151**	0.179	0.167
THR Fr	0.112	0.111	0.098	0.105	0.111	**0.096**
ACE 2.5D	0.224	0.220	**0.200**	0.204	0.221	0.230
AchE 2.5D	0.168	0.169	**0.159**	0.173	0.163	0.166
BZR 2.5D	0.188	0.186	0.183	0.183	0.188	**0.180**
COX2 2.5D	0.135	0.135	**0.126**	0.127	0.133	0.142
DHFR 2.5D	0.168	0.171	0.157	**0.141**	0.218	0.213
GPB 2.5D	0.210	0.217	**0.178**	0.189	0.185	0.189
THER 2.5D	0.255	0.249	0.234	**0.228**	0.268	0.272
THR 2.5D	0.206	0.203	**0.187**	0.187	0.190	0.189
breast-cancer	0.206	0.207	0.198	**0.198**	0.214	0.203
breast-w	0.030	0.028	0.028	**0.027**	0.029	0.028
cmc	0.384	0.386	**0.361**	0.372	0.369	0.367
credit-a	0.118	0.117	0.111	0.110	**0.105**	0.109
cylinder	0.175	0.176	0.177	0.199	0.168	**0.166**
diabetes	0.218	0.219	0.197	**0.194**	0.197	0.196
haberman	0.247	0.242	0.213	0.215	0.211	**0.211**
heart-cleve	0.130	0.135	0.126	0.131	0.129	**0.120**
heart-statlog	0.140	0.146	0.124	**0.118**	0.135	0.129
ionosphere	0.067	**0.065**	0.076	0.080	0.076	0.074
liver-bupa	0.268	0.266	0.225	0.243	0.238	**0.223**
lymph	0.137	0.135	0.130	**0.124**	0.130	0.125
tae	0.408	0.411	**0.367**	0.371	0.385	0.406
tic-tac-toe	0.016	0.017	0.015	**0.014**	0.016	0.016
vehicle	0.151	0.150	**0.144**	0.151	0.154	0.159
wine	0.020	0.019	0.017	0.018	0.015	**0.013**
Most diversity	**0**	**1**	**9**	**11**	**2**	**9**
Mean Rank	**4.80**	**4.67**	**2.39**	**2.86**	**3.33**	**2.95**

Here, the picture is quite different from that of comparing base classifier accuracy, when summarizing the results over all the data sets. There are several interesting observations that can be made in Table VI. First of all, Ens and GAS, which achieved high base classifier accuracies, were not very diverse. This could be expected, since they are the only ensembles not utilizing bagging; i.e. training on all available data. When considering mean ranks, the ordinary ensembles trained with bagging are, on the other hand, the most diverse. Furthermore, GASEN resulted in the most diverse ensembles quite a few times. Unfortunately, it does not change the fact that GASEN still is significantly worse than the full ensemble when considering ensemble accuracy, which, in the end, is all that matters.

Finally, the results seem to indicate that using negative correlation learning in NegBagg does indeed result in more diversity, but the effect is rather small.

Only one diversity measure, *double fault*, has been evaluated in this study. It should be noted, however, that several theoretical and empirical studies have shown that most other diversity measures have a very weak to non-existent correlation with ensemble test accuracy. The only exception, in addition to *double fault*, is the non pair-wise diversity measure *difficulty*. However, the distribution of the *difficulty* measure is not necessarily similar to *double fault*, since they measure diversity in different ways.

To conclude the analysis of the results, it is clear that the straightforward methods obtained higher ensemble accuracy than their sophisticated counterparts. In the case of NegBagg, this can most probably be attributed to the lower base classifier accuracy achieved by that method. The GASEN ensembles were generally outperformed by their ordinary ensemble counterparts when considering ensemble accuracy. For ensembles trained without bagging, this was despite the fact that there was no detectable difference when comparing both base classifier accuracies and diversities. When considering ensembles trained with bagging, on the other hand, GASEN seems to have overemphasized diversity, at the cost of base classifier accuracy. Finally, there was no significant difference in ensemble accuracy between the ordinary ensembles. Nevertheless, there are different reasons for their high performance. While bagging leads to less accurate base classifiers, it results in more diverse ensembles. When not using bagging, on the other hand, the base classifiers were not very diverse, but very accurate individually.

V. CONCLUSION

We have compared the predictive performance of ensembles of artificial neural networks, generated by straightforward and more sophisticated methods. In this study, the straightforward methods generated each ensemble member independently of the others as opposed to the sophisticated methods. In particular, we have investigated whether the more complex methods, as exemplified by GASEN and NegBagg, outperform straightforward implicit methods. An empirical investigation has been presented that provides strong evidence in favor of the straightforward methods, hence questioning the use of the more complex methods, at least for the type of ensembles considered in this study.

The analysis of mean accuracy and diversity among the ordinary ensembles shows that lower mean base classifier accuracy can be compensated for by more diversity. More specifically, even though the ordinary ensembles trained without bagging had more accurate but less diverse base classifiers, and the opposite was true for ensembles trained with bagging, there was no significant difference in ensemble accuracy between them.

VI. Discussion

It should be noted that the overall purpose of this study was not to discredit GASEN and NegBagg. On the contrary, we believe that many recognized algorithms would suffer similar results if evaluated in the same way. Nevertheless, the results presented here clearly cast some doubts on sophisticated ensemble techniques in general.

One direction for future work concerns investigating if the main findings also hold for other types of ensembles. This includes exploring larger sets of ensembles as well as networks with larger variations in architecture. Another important task for future work is to evaluate other diversity measures for the classification context. One set of diversity measures, applicable only to multiclass problems, are the so called *diversity of errors measures*, see e.g. [20], which only measures diversity on instances where base classifiers err. Recently, Ko et al., in [21], proposed a method for combining mean accuracy with an arbitrary diversity measure, thus creating what they called *compound diversity*, for which they reported some promising results.

Another direction for future work concerns the identification of conditions under which explicit methods may outperform the implicit ones. This direction would obviously also include meta-learning, concerning characteristics of different problems and how different ensemble techniques apply to them. One question to address is how the size of the ensembles affects performance and diversity for different ensemble algorithms.

References

[1] T.G. Dietterich, "Machine-learning research: Four current directions," *AI magazine*, vol. 18, 1997, pp. 97-136.

[2] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *Journal of Information Fusion*, vol. 6, 2005, pp. 5-20.

[3] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, 1996, pp. 123-140.

[4] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *in Advances in Neural Information Processing Systems (NIPS)*, vol. 7, 1995, pp. 231-238.

[5] L. Saitta, "Hypothesis Diversity in Ensemble Classification," *Foundations of Intelligent Systems*, Springer, 2006, p. 662–670.

[6] L.I. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble,accuracy" *Machine Learning*, vol. 51, 2003, pp. 181-207.

[7] U. Johansson, T. Löfström, and L. Niklasson, "The Importance of Diversity in Neural Network Ensembles - An Empirical Investigation," *The International Joint Conference on Neural Networks*, 2007.

[8] G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing*, vol. 19, 2001, pp. 699-707.

[9] L.K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine*, vol. 12, 1990, pp. 993-1001.

[10] T. Löfström, U. Johansson, and H. Boström, "On the Use of Accuracy and Diversity Measures for Evaluating and Selecting Ensembles of Classifiers," *Seventh International Conference on Machine Learning and Applications*, 2008, pp. 127-132.

[11] U. Johansson, T. Löfström, and L. Niklasson, "Evaluating Standard Techniques for Implicit Diversity," *Advances in Knowledge Discovery and Data Mining*, 2008, pp. 592-599.

[12] Z.H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artificial Intelligence*, vol. 137, 2002, pp. 239-263.

[13] M.M. Islam, X. Yao, S. Shahriar Nirjon, M. Islam, and K. Murase, "Bagging and boosting negatively correlated neural networks," *IEEE Transactions on Systems Man and Cybernetics-Part B-Cybernetics*, vol. 38, 2008, p. 771–784.

[14] Y. Liu, "Negative correlation learning and evolutionary neural network ensembles," *Ph.D. thesis,University College, The University of New South Wales, Australian Defence Force Academy,Canberra, Australia,* 1998.

[15] D. Patterson, (2010, Jan 10). "a per-epoch backpropagation training for a multilayer feedforward neural network," [Source code], [Online].
Avalable: http://www.csee.umbc.edu/~dpatte3/nn/res/bbackprop.m.

[16] C. Bruce, J. Melville, S. Pickett, and J. Hirst, "Contemporary QSAR classifiers compared," *Journal of chemical information and modeling*, vol. 47, 2007, pp. 219-227.

[17] J. Sutherland and L. O'Brien, "A Comparison of Methods for Modeling Quantitative Structure− Activity Relationships," *Journal of Medicinal Chemistry*, vol. 47, 2004, pp. 5541-5554.

[18] A. Asuncion and D.J. Newman, "UCI machine learning repository," School of Information and Computer Sciences. University of California, Irvine, California, USA, 2007

[19] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, 2006, pp. 1-30.

[20] M. Aksela and J. Laaksonen, "Using diversity of errors for selecting members of a committee classifier," *Pattern Recognition*, vol. 39, 2006, pp. 608-623.

[21] A. Ko, R. Sabourin, and A.D. Britto JR, "Compound Diversity Functions for Ensemble Selection," *Int. J. Patt. Recog. Art. Intel.*, vol. 23, 2009, pp. 659-686.