Improving an ensemble of neural networks via a novel multi-class decomposition schema

Antonio L. Alfeo[1,2], Mario G.C.A. Cimino [1,2] and Guido Gagliardi[1,3,4]✉
[1]*Department of Information Engineering, University of Pisa, Pisa, Italy.*
[2]*Bioengineering and Robotics Research Center E. Piaggio, University of Pisa, Pisa, Italy.*
[3]*Department of Information Engineering (DINFO), University of Florence, Florence, Italy.*
[4]*Department of Electrical Engineering, KU Leuven, Leuven, Belgium.*
luca.alfeo@ing.unipi.it, mario.cimino@unipi.it, guido.gagliardi@phd.unipi.it

Keywords: Machine Learning, Neural Networks, Multiclass Decomposition, Ensemble learning, Non-competent classifier

Abstract: The need for high recognition performance demands increasingly complex machine learning (ML) architectures, which might be extremely computationally burdensome to be implemented in real-world. This issue can be addressed by using an ensemble learning model to decompose the multi-class classification problem into many simpler binary classification problems, e.g. each binary classification problem can be handled via a simple multi-layer perceptron (MLP). The so-called *one-versus-one* (OVO) is a widely used multi-class decomposition schema in which each classifier is trained to distinguish between two classes. However, with an OVO schema each MLP is non-competent to classify instances of classes that have not been used to train it. This results in classification noise that may degrade the performance of the whole ensemble, especially when the number of classes grows. The proposed architecture employs a weighting mechanism to minimize the contribution of the non-competent MLPs and combine their outcomes to effectively solve the multi-class classification problem. In this work, the robustness to the classification noise introduced by non-competent MLPs is measured to assess in what conditions this translates in better classification accuracy. We test the proposed approach with five different benchmark data sets, outperforming both the baseline and one state-of-the-art approach in multi-class decomposition algorithms.

1 INTRODUCTION

Ensemble-based machine learning approaches provide the classification results as a combination of the outcomes of different machine learning models (base models). This strategy is aimed at improving the classification performance by averaging the error of a single base model in the whole ensemble (Zhang and Ma, 2012). This allows (i) using much simpler base classifiers than a non-ensemble-based ML approach with similar performance, thus resulting in improved computational efficiency (Sagi and Rokach, 2018); and (ii) providing greater robustness against overfitting and local minima by differentiating the training of the base models to better cover the solution space (Zhang and Ma, 2012).

In this context, the differentiation of the base models is key to reduce the sensitivity of the ensemble to the choice of the training set (Belkin et al., 2019) and thus provide greater and more robust classification accuracy. In (Sagi and Rokach, 2018) the authors categorizes the strategies to differentiate base models according to which component of the classification procedure they manipulate: input, learning algorithm, and output. With *input manipulation*, the base models are differentiated by being trained with different partitions of a data set, i.e., using different instances (horizontal partitioning) or different features (vertical partitioning). With *learning algorithm manipulation*, the base models are differentiated by using different algorithms for each one of them, or by using the same algorithm with different hyper-parameters. With *output manipulation*, a single multi-class classification problem is transformed into multiple binary classification problems, and each one of them is assigned to a base model. This strategy is also called multi-class problem decomposition, since it decomposes the problem into different simpler sub-problems (Galar et al., 2011). This *divide and conquer* approach aims at recognizing complex macro-behaviors as a combination of simpler-to-recognize micro-behaviors (Alfeo et al., 2017a), and allows both single (Alfeo et al., 2018)

and multiple abstraction layers for the representation of such behaviors (Alfeo et al., 2017b).

The two main multi-class decomposition schema are named *one-versus-all* (OVA) and *one-versus-one* (OVO) (Goienetxea et al., 2021). Given a multi-class classification problem, the *one-versus-all* (OVA) decomposition scheme, consists of generating a sub-problem for each class, i.e. distinguishing that class from all the others. Via a *OVO* strategy, the problem is decomposed into multiple binary classifications, one for each possible pair of classes. In a C-multi-class classification problem, OVA and OVO scheme results in C and $C(C-1)/2$ base models, respectively. Given its greater recognition performance, OVO scheme is often preferred over OVA (Alfeo et al., 2021).

Still, the robustness of OVO schema may be affected by the number of non-competent base classifiers. In a OVO scheme, a base classifier *BC* is non-competent for the classification of a sample s if the class to which s belongs does not match any of the classes used to train *BC*. Among all base classifiers in an OVO scheme for C-multi-class classification problem, $(C-1)$ base classifiers are considered competent to classify a sample of class C, whereas the others are non-competent (Galar et al., 2011).

Thus, the prediction of *BC* for sample s can be unreliable and unpredictable, resulting in classification noise that may propagate up to the final decision via the base model's outcomes aggregation (Galar et al., 2015). The management of the classification noise generated by non-competent classifiers is key to prevent the degradation of the classification performance. This work aims at enriching the literature of techniques designed for this purpose.

The paper is structured as follows. In section 2, the literature review is presented. Section 3 details the proposed approach. The experimental results are presented in sections 4. Finally, Section 5 discusses the obtained results and the conclusions.

2 RELATED WORKS

Non-competent base models can be identified exactly only by knowing *apriori* the class of the sample being classified. Of course, at classification time, this information is unknown. For this reason, several approaches have been proposed to aggregate the outcome of the base models while mitigating the effect of the classification noise generated by non-competent ones. According to the survey (Cruz et al., 2018), the main strategies can be grouped as *non-trainable*, *trainable*, and based on *dynamic weighting*. *Non-trainable* approaches combine the outcome of the base models by leveraging some assumptions about the classifiers or the classification problem. For instance, the most used non-trainable approach is the majority voting strategy, which relies on the assumption that all base classifiers are independent (Duin, 2002). *Trainable* approaches employ the outcomes of the base models as input for another machine learning model (Cruz et al., 2010). Despite having a more complex architecture design due to the additional model to tune and train, these approaches usually result in greater classification accuracy when compared to the non-trainable approaches, since the aggregation of base models' outcome is actually data-driven (Cruz et al., 2018). With *dynamic weighting*, the outcomes of all base classifiers are weighted according to their expected local competence and subsequently aggregated to provide the final decision (Zhang et al., 2019). Most of them assess the competence of a base model on the local region of the feature space close to the instance to classify. This region can be defined by employing a k-NN procedure (Cruz et al., 2018). The weighted aggregation is designed to provide the most competent classifiers with higher contribution in the classification outcome (Cruz et al., 2018).

Overall, choosing the best weights to be used in the aggregation of the base classifiers' outcome, is not trivial at all (Costa et al., 2018). Such weights can be identified by leveraging some prior knowledge about the classification problem (Costa et al., 2018), or by searching them via an optimization method, such as a genetic algorithm (GA) (Pintro et al., 2013). However, all the aforementioned approaches apply procedures to define static weights, i.e. not adaptive to a specific sample. With *dynamic weighting* instead, the system weights the outcome of the base classifiers by learning the weights directly from the data and specializing them for each data sample. While non-trainable approaches (e.g., a set of rules) result in the lowest complexity, trainable approaches (e.g. a machine learning algorithm) may result in the best classification performance (Cruz et al., 2018). In this context, *dynamic weighting* approaches can offer the best trade-off.

Examples of dynamic weighting schemes are the local classifier weighting by quadratic programming (Cevikalp and Polikar, 2008), the dynamic integration of classifiers (Jiménez, 1998), and the fuzzy dynamic classifier aggregation (Štefka and Holeňa, 2015). A large number of dynamic weighting schemas are extensively compared in (Zhang et al., 2017). Among the others, the approach known as DRCWOVO (Galar et al., 2015) has proved its effectiveness with a number of different datasets (Zhang et al., 2017). DRCWOVO is specifically designed for OVO multi-class

decomposition, and weights the base classifier outputs classifying a sample s according to its distance to the k nearest neighbors taken from each class (Galar et al., 2015). Given such effectiveness (Zhang et al., 2017), DRCWOVO is employed as a term of comparison w.r.t. the approach proposed in this paper.

3 ENSEMBLE ARCHITECTURE

In the OVO multi-class decomposition scheme proposed in this work, the base models are designed as multi-layer perceptrons (MLP) (Cimino et al., 2009), the simplest yet powerful machine learning architectures based on artificial neural networks (Galatolo et al., 2021).

Being a OVO schema, each base model is trained to distinguish between two classes, and provides as output the class probability associated with these two specific classes.

By weighting such probabilities and aggregating them, a membership score $MS(C_i)$ is computed for each class. Finally, for sample s, the class predicted by the proposed approach corresponds to $C(s)$ computed as $argmax(MS^s(C_i))$ among the classes.

$$MS^s(C_i) = \sum_{j=1, j\neq i}^{n} P^s_{[C_i|C_j]}(C_i) * \frac{P^s_{k-NN}(C_i) + P^s_{k-NN}(C_j)}{2} \tag{1}$$

As introduced in Section 1, due to the classification noise generated by non-competent classifiers, different classes may result in similar MS values. In this case, given the final *argmax* operation used with OVO schemas, small fluctuations in the MS values might result in classification errors. To minimize the MS of the wrong classes and thus increase the robustness of the proposed OVO schema, the base models' output undergoes a weighting operation (Eq. 1) to mitigate the contribution of the base models that are most likely to be non-competent.

In Eq. 1, $P^s_{[C_i|C_j]}(C_i)$ is the probability that a given sample s belongs to class C_i, provided by the base model trained on C_i and C_j. The weighting factor (the fraction in Eq. 1) exploits N_s, i.e., the close neighborhood of s from the training set. The competence of the base model trained on C_i and C_j is evaluated as the average class probability for classes C_i and C_j in N_s, obtained via the k-NN classifier. Specifically, the base model trained with the most (less) frequent classes in the close neighborhood, N_s, of a sample s is considered the most competent to classify s, and hence, their outcomes are overweighted (underweighted). Such a weighting operation propagates up to MS.

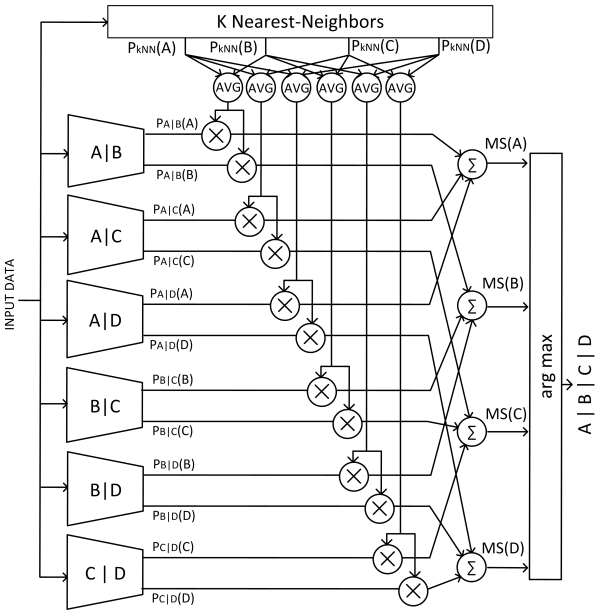


Figure 1: Proposed ensemble architecture with an exemplary classification problem consisting of 4 hypothetical classes (A, B, C, D).

4 EXPERIMENTAL SETUP

The Multi-Layer Perceptrons (MLP) and the k-NN procedure are provided via *sklearn*, a well-known machine learning library in Python. The class probability is obtained via the method *predict_proba*() provided by *sklearn* (Feng et al., 2018). Each MLP features the following hyperparameters: *relu* as activation function, 1e-3 as alpha value, two hidden layers of sizes 120 and 70, *lbfgs* as the solver, 10000 as the maximum number of iterations, 1e-6 as tolerance value. The k-NN procedure features the cosine distance as distance measure, and the number of nearest neighbors equal to 5.

We test the proposed approach by comparing it with the standard weighting-based OVO approach, and the state-of-the-art weighting-based OVO approaches, e.g. DRCWOVO (Galar et al., 2015) featuring the same configuration for the k-NN procedure. For each experiment a 5-folds cross validation is employed as experimental framework.

To prove the generality of the proposed approach it is tested on a number of different data sets. Those are characterized by a different number of attributes, classes and instances. Specifically, the data sets used in this research work are listed below, and they are obtained via the publicly available and well-known UCI repository (Bay et al., 2000).

- *ecoli*: the data addresses 6 different protein localization sites. The classes are not uniformly distributed.

- *glass*: the data, provided by the USA Forensic Science Service, defines the oxide content (i.e. Na, Fe, K, etc.) of 6 types of glass. The attributes of each instance are real values.
- *page-blocks*: the data describes the blocks of the page layout of a document obtained via a segmentation process. The attributes of each instance are real values.
- *shuttle*: the data describes different shuttles' type of control via numerical attributes. Approximately 80% of the data belongs to one single class.
- *zoo*: the database describes via 17 boolean-valued attributes 7 groups of animal species.

We summarize the main characteristics of these data sets in Table 1.

Table 1: Characteristics of the experimental datasets.

Dataset	#Classes	#Samples	#Attributes
ecoli	8	336	7
glass	6	214	9
page-blocks	5	5473	10
shuttle	6	58000	9
zoo	7	101	17

The classification accuracy is employed as the main classification performance metric (Eq. 2), which is defined as the ration between the correct classified predictions and the total number of predictions.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (2)$$

The $CompetenceReinforcement$ (CR) measure is proposed (Eq. 3) to evaluate the effectiveness of the weighting operation in minimizing such classification noise.

$$CompetenceReinforcement_s = \frac{MS^s(C_s)}{\sum_{j=1}^{\#classes} MS^s(C_j)} \quad (3)$$

The CR for sample s is defined as the ratio of the MS^s of the true class, C_s (i.e., $MS^s(C_s)$) and the sum of all membership score (MS) values associated with s (i.e., with all classes). CR ranges between one (best case) and zero (worst case). The higher the CR value the lower the MS associated with the wrong classes, which may be due to the classification noise provided by non-competent classifiers.

5 RESULTS AND DISCUSSION

The table 2 shows the results obtained in terms of average accuracy and standard deviation with a 5 cross fold validation approach on the data sets considered.

When considering the *ecoli*, *glass*, *page-blocks* and *shuttle* data sets, it can be seen that there is a progression in the improvement of the results obtained with the OVO, DRCWOVO and the proposed approach, respectively.

Considering the data sets with the greatest number of samples: *page-blocks* and *shuttle*, all the three approaches perform very well, proving how the proposed decomposition scheme handles effectively these classification problem despite the different the number of classes, their balance and the number of attributes in the data set. *Page-blocks* is the only dataset with real values attributes and so it may result in a more complex classification problem than the others. In this case, the performance gain obtained by using the proposed approach is significant, reaching 96.7% accuracy (about 10% more of the other approaches).

With the *zoo* data set, which features the smallest number of samples, and the largest number of attributes (all boolean values), the OVO scheme results in 94.1% accuracy, i.e. on average is 7% more accurate than DRCWOVO. Again, the introduction of the proposed approach resulted in even better recognition performances by reaching 95% and reducing the standard deviation from OVO's 6.4% to 3.5%.

Table 2: Average % accuracy ± standard deviation over 5 repeated trials. The result with the best average accuracy is in bold.

Dataset	OVO	DRCWOVO	Proposed Approach
ecoli	71.4 ± 7.3	82.7 ± 5.0	**86.0 ± 2.5**
glass	58.3 ± 9.8	64.0 ± 5.2	**70.0 ± 4.7**
pageblocks	84.6 ± 3.8	86.5 ± 1.6	**96.7 ± 0.3**
shuttle	99.6 ± 0.1	99.7 ± 0.1	**99.8 ± 0.01**
zoo	94.1 ± 6.4	87.1 ± 2.6	**95.0 ± 3.5**

As mentioned in Section 1, OVO decomposition scheme performances can be affected by the classification noise generated by the non-competent classifiers. Such a noise may propagate up to the final prediction, i.e. increasing the MS value for the wrong classes, and increasing the sensitivity of the schema to small MS fluctuations (Section 3). Table 3 reports the CR values obtained with the proposed approach, standard weighting-based OVO, and DRCWOVO.

When compared to the OVO approach, both k-NN-based weighting approaches (DRCWOVO and the proposed approach) are able to effectively reduce

Table 3: Average % CR ± standard deviation over 5 repeated trials. The result with the best average % CR is in bold.

Dataset	OVO	DRCWOVO	Proposed Approach
ecoli	15.7 ± 0.4	18.4 ± 0.6	**26.3 ± 1.2**
glass	19.3 ± 1.0	20.3 ± 0.5	**24.2 ± 1.1**
page-blocks	34.6 ± 3.0	44.9 ± 0.6	**72.7 ± 0.6**
shuttle	19.1 ± 0.2	30.4 ± 0.2	**64.4 ± 0.01**
zoo	14.6 ± 0.5	16.7 ± 0.5	**24.2 ± 0.02**

classification noise (Table 3). Moreover, passing from OVO to DRCWOVO to the proposed approach corresponds to a progressive improvement of the CR values (Table 3) and this corresponds to a gain in terms of recognition accuracy (Table 2). This confirms that the improvement in the recognition performance is actually correlated to the capability of reducing the classification noise.

The number of classes in the recognition problem and the values of competence reinforcement exhibit a negative correlation. This is explained by considering that the number of classifiers increases quadratically with respect to the number of classes; this corresponds to a greater number of non-competent classifiers, and thus to a greater number of terms in the denominator of the competence reinforcement formula.

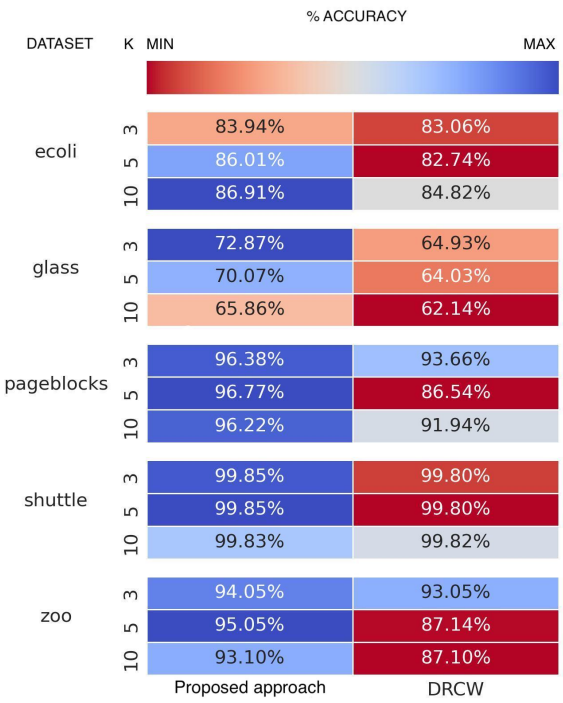
This is evident considering the lower CR values with the data sets corresponding to the highest number of classes: *ecoli*, *zoo*, and *glass*. On the other hand, with the *page-blocks* data set (which features the lowest number of classes) the proposed approach reaches CR values of about 72.7%.

Finally, it is provided an evaluation of the impact of varying the k-values within the k-NN procedure for the proposed approach and DRCWOVO, as this is the main hyper-parameter of the two weighting strategies. The figure 2 shows the results in terms of average accuracy over 5 cross-fold validation on the data sets analysed with the proposed approach and DRCWOVO with three different k-values in 3, 5, 10.

As k varies, the proposed approach results to be more accurate than DRCWOVO in all the data sets analyzed.

Considering the proposed approach, the variation of the hyper-parameter k introduces a negligible variation on the average accuracy, especially in the *page-blocks*, *shuttle*, and *zoo* data sets. On the other hand, the choice of k results in a difference between minimum and maximum accuracy value of approximately 2.97% for *ecoli* and 7.01% for *glass* data set. Instead, DRCWOVO approach results in a maximum accuracy variation due to different values of k on the *page-blocks* dataset, 7.12%, and the *zoo* dataset, 5.95%.

Figure 2: Accuracy scores on the different UCI data sets of the proposed model and DRCWOVO varying the k-value of the k-NN classifier.



Overall, these results suggest that DRCWOVO is more sensitive the right choice of k with respect to the proposed approach.

The usage of the proposed weighting approach positively impacts the classification performance as can be seen by comparing the proposed approach's and the OVO's results. As k k varies our approach consistently outperforms OVO. Considering DRCWOVO instead there is a loss of accuracy in the *zoo* dataset with k values different than 3. These results in lower performance than the simple OVO approach.

6 CONCLUSION

In this paper, a new MLP ensemble scheme is proposed. The scheme features a number of MLPs trained in an OVO fashion and a k-NN procedure to weight their outcomes and obtain the final classification.

The proposed scheme was tested with 5 benchmark data sets: ecoli, glass, page-block, shuttle, and zoo, through a 5 cross-fold validation methodology and compared with two other decomposition schema: OVO and DRCWOVO. The comparison metrics considered were the recognition accuracy, and compe-

tence reinforcement to measure the ability of these OVO decomposition schemes of handling the classification noise provided by non-competent classifiers.

The results confirm that the proposed approach achieves better results in terms of accuracy than other state-of-the-art methods. Furthermore, considering the results obtained by measuring competence reinforcement, the proposed scheme is less affected by classification noise due to non-competent classifiers as it produces better CR results than state-of-the-art approaches.

Future developments of the architecture will include testing on real-world data sets, to prove its effectiveness also in real-world applications. Especially in this context, it would be useful to have an explainable artificial intelligence architecture, and since the proposed approach leverages a number of the binary base classifiers that are specialized on the decision boundary between a pair of classes, those can be fruitfully employed to generate counterfactual explanations, i.e. motivate why a given prediction belongs to a given class rather than the other one.

ACKNOWLEDGMENT

Work partially supported by (i) the Tuscany Region in the framework of the "SecureB2C" project, POR FESR 2014-2020, Law Decree 7429 31.05.2017; (ii) the Italian Ministry of University and Research (MUR), in the framework of the "Reasoning" project, PRIN 2020 LS Programme, Law Decree 2493 04-11-2021; and (iii) the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence). The authors thank Mirco Quintavalla for his work on the subject during his master thesis.

REFERENCES

Alfeo, A. L., Barsocchi, P., Cimino, M. G., La Rosa, D., Palumbo, F., and Vaglini, G. (2018). Sleep behavior assessment via smartwatch and stigmergic receptive fields. *Personal and ubiquitous computing*, 22(2):227–243.

Alfeo, A. L., Catrambone, V., Cimino, M. G., Vaglini, G., and Valenza, G. (2021). Recognizing motor imagery tasks from eeg oscillations through a novel ensemble-based neural network architecture. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5983–5986. IEEE.

Alfeo, A. L., Cimino, M. G. C., Egidi, S., Lepri, B., Pentland, A., and Vaglini, G. (2017a). Stigmergy-based modeling to discover urban activity patterns from positioning data. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 292–301. Springer.

Alfeo, A. L., Cimino, M. G. C. A., and Vaglini, G. (2017b). Measuring physical activity of older adults via smartwatch and stigmergic receptive fields. In *The 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 724–730. PRT.

Bay, S. D., Kibler, D., Pazzani, M. J., and Smyth, P. (2000). The uci kdd archive of large data sets for data mining research and experimentation. *ACM SIGKDD explorations newsletter*, 2(2):81–85.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Cevikalp, H. and Polikar, R. (2008). Local classifier weighting by quadratic programming. *IEEE Transactions on Neural Networks*, 19(10):1832–1838.

Cimino, M. G., Pedrycz, W., Lazzerini, B., and Marcelloni, F. (2009). Using multilayer perceptrons as receptive fields in the design of neural networks. *Neurocomputing*, 72(10-12):2536–2548.

Costa, V. S., Farias, A. D. S., Bedregal, B., Santiago, R. H., and Canuto, A. M. d. P. (2018). Combining multiple algorithms in classifier ensembles using generalized mixture functions. *Neurocomputing*, 313:402–414.

Cruz, R. M., Cavalcanti, G. D., and Ren, T. I. (2010). An ensemble classifier for offline cursive character recognition using multiple feature extraction techniques. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. Ieee.

Cruz, R. M., Sabourin, R., and Cavalcanti, G. D. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216.

Duin, R. P. (2002). The combining classifier: to train or not to train? In *Object recognition supported by user interaction for service robots*, volume 2, pages 765–770. IEEE.

Feng, P., Ma, J., Sun, C., Xu, X., and Ma, Y. (2018). A novel dynamic android malware detection system with ensemble learning. *IEEE Access*, 6:30996–31011.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761–1776.

Galar, M., Fernández, A., Barrenechea, E., and Herrera, F. (2015). Drcw-ovo: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems. *Pattern recognition*, 48(1):28–42.

Galatolo, F. A., Cimino, M. G., Marincioni, A., and Vaglini, G. (2021). Noise boosted neural receptive fields. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6. IEEE.

Goienetxea, I., Mendialdua, I., Rodríguez, I., and Sierra, B. (2021). Problems selection under dynamic selection of the best base classifier in one versus one: Pseudovo. *International Journal of Machine Learning and Cybernetics*, 12(6):1721–1735.

Jiménez, D. (1998). Dynamically weighted ensemble neural networks for classification. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, volume 1, pages 753–756. IEEE.

Pintro, F., Canuto, A. M., and Fairhurst, M. (2013). Using genetic algorithms and ensemble systems in online cancellable signature recognition. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Štefka, D. and Holeňa, M. (2015). Dynamic classifier aggregation using interaction-sensitive fuzzy measures. *Fuzzy Sets and Systems*, 270:25–52.

Zhang, C. and Ma, Y. (2012). *Ensemble machine learning: methods and applications*. Springer.

Zhang, Z.-L., Chen, Y.-Y., Li, J., and Luo, X.-G. (2019). A distance-based weighting framework for boosting the performance of dynamic ensemble selection. *Information Processing & Management*, 56(4):1300–1316.

Zhang, Z.-L., Luo, X.-G., García, S., Tang, J.-F., and Herrera, F. (2017). Exploring the effectiveness of dynamic ensemble selection in the one-versus-one scheme. *Knowledge-Based Systems*, 125:53–63.