A heterogeneous ensemble approach for the prediction of the remaining useful life of packaging industry machinery

Francesco Cannarile, Piero Baraldi, M. Compare, D. Borghi, L. Capelli, Enrico Zio

A heterogeneous ensemble approach for the prediction of the remaining useful life of packaging industry machinery

F. Cannarile
Energy Department, Politecnico di Milano, Via La Masa 34/3, 20156 Milano, Italy
Aramis Srl, Via Pergolesi 5, 20124, Milano, Italy
P. Baraldi
Energy Department, Politecnico di Milano, Via La Masa 34/3, 20156 Milano, Italy
M. Compare
Energy Department, Politecnico di Milano, Via La Masa 34/3, 20156 Milano, Italy
Aramis Srl, Via Pergolesi 5, 20124, Milano, Italy
D. Borghi & L. Capelli
Tetra Pak Packaging Solutions S.p.A., 41122 Modena, Italy
E. Zio
Energy Department, Politecnico di Milano, Via La Masa 34/3, 20156, Milano, Italy
Chair on Systems Science and the Energetic Challenge, European Foundation for New Energy-Electricité de France, Ecole Centrale Paris and Supelec, France
Aramis Srl, Via Pergolesi 5, 20124, Milano, Italy

ABSTRACT: We present a method based on heterogeneous ensemble learning for the prediction of the Remaining Useful Life (RUL) of cutting tools (knives) used in the packaging industry. Ensemble diversity is achieved by training multiple prognostic models using different learning algorithms. The combination of the outcomes of the models in the ensemble is based on a weighted averaging strategy, which assigns weights proportional to the individual model performances on patterns of a validation set. The proposed heterogeneous ensemble has been applied to real condition monitoring knife data. It has provided more accurate RUL predictions compared to those of each individual base

1 INTRODUCTION

As the digital, physical and human worlds continue to integrate, the 4th industrial revolution, the internet of things and big data, the industrial internet, are changing the way we design, manufacture, deliver products and services. In this fast-pace changing environment, the attributes related to the reliability of components and systems continue to play a fundamental role for industry. On the other hand, the advancements in knowledge, methods and techniques, the increase in information sharing and data availability, offer new opportunities of analysis and assessment for reliability engineering. Based on this increased knowledge, information and data available, we can improve our reliability prediction capability. Particularly, the increased availability of data coming from monitoring the relevant components and systems parameters and the grown ability of treating these data by intelligent algorithms capable of mining out information relevant to the assessment and prediction of their state, has open wide the doors for Prognostics and Health Management (PHM) and predictive maintenance in many industrial sectors, for improved operation and maintenance (Zio, 2016). Approaches for RUL estimation can be generally categorized into model-based and data-driven (Baraldi et al., 2015a). Model-based approaches use physics-based models to describe the degradation behavior of the equipment (Baraldi et al., 2015a). On the other side, data-driven methods are of interest when an explicit model of the degradation process is not available, as they rely on the availability of field data collected during the operation of one or more similar components. Among data-driven methods one can distinguish between (i) degradation-based approaches, modeling the future equipment degradation evolution and (ii) direct RUL prediction approaches, directly predicting the RUL.

Degradation-based approaches are based on statistical models that *learn* the equipment degradation time evolution from time series of the observed degradation (Baraldi et al., 2017). The predicted degradation state is, then, compared with a failure criterion, such as the value of deg-

radation beyond which the equipment fails performing its function (failure threshold). Examples of modeling techniques used in degradation-based approaches are Auto-Regressive models (Gorjian et al., 2009), Relevance Vector Machines (Di Maio et al., 2012) and Semi-Markov Models (Cannarile et al., 2017a) (Cannarile et al., 2018).

Direct RUL predictions approaches, instead, typically resort to machine learning techniques that directly map the relation between the observable parameters and the equipment RUL, without the need of predicting the equipment degradation state evolution towards a failure threshold (Schwabacher et al., 2007). Techniques used in direct RUL prediction approaches are, for example, Artifical Neural Networks (Wang & Vachtsenavos, 2001), Extreme Learning Machines (ELM) (Yang et al., 2017), Gaussian Processes (GP) (Baraldi et al., 2015b), etc.

When few run-to-failure degradation trajectories are available, direct RUL approaches may overfit, i.e., these algorithms customize themselves too much to learn the relationship between the observable parameters and the corresponding RUL in the training set. Therefore, these methods tend to lose their generalization power, which leads to poor performance on new data. To overcome this, ensemble approaches, based on the aggregation of multiple model outcomes, have been introduced (Baraldi et al., 2013a). The basic idea is that the diverse models in the ensemble complement each other by leveraging their strengths and overcoming their drawbacks.

Thus, the combination of the outcomes of the individual models in the ensemble improves the accuracy of the predictions compared to the performance of a single model (Brown et al., 2005) (Baraldi et al., 2013a). Different methods, such as ANN (Baraldi et al., 2013b), Support Vector Machine (SVM) (Liu et al., 2006) and kernel learning (Liu et al., 2015), have been used with success to build the individual models. For example, an ensemble of feedforward Artificial Neural Networks (ANN) has been embedded into a Particle Filter (PF) for the prediction of crack length evolution (Baraldi et al., 2013b) and an ensemble of data-driven regression models has been exploited for the RUL prediction of lithium-ion batteries (Xing et al., 2013). In (Rigamonti et al., 2017) a local ensemble of Echo State Networks (ESN) has been proposed to improve the RUL prediction accuracy of turbofan engines.

The objective of this work is to predict the RUL of knives installed on Tetra Pak® A3/Flex filling machines used to cut package material. The prognostic task is complicated by the fact that few run-to-failure degradation trajectories are available, and a failure threshold is not available. To cope with these issues, this work proposes an ensemble formed by multiple data-driven direct RUL prediction models, capable of aggregating the RUL predictions for good performance throughout the entire degradation trajectory of a knife. Ensemble diversity is achieved by *heterogeneous* ensemble generation, i.e., by training the models using different prognostics algorithms. Aggregation is obtained by averaging the output of the individual base models with weights proportional to the inverse of their Empirical Generalization Error (EGE) on retrieved patterns in a validation set. The application of the proposed heterogeneous ensemble method to real condition monitoring knife data has shown to provide more accurate RUL prediction compared to that of each individual base learner in the ensemble.

The paper is organized as follows: in Section 2, the objectives of this work and the assumptions are discussed; in Section 3, ensemble learning main concepts for data-driven direct RUL prediction are illustrated; in Section 4, performance metrics to compare different prognostic models are discussed. The application of the methodology to Tetra Pak® A3/Flex filling data is described in Section 5, whereas Section 6 draws the work conclusions.

2 ASSUMPTIONS AND OBJECTIVES

We assume to have available run-to-failure degradation trajectories of N pieces of equipment similar to the one currently monitored (test equipment). Let $\boldsymbol{x}_i(\tau_i) \in \mathbb{R}^m, i = 1,..,N;\ \tau = 1,\dots,n_i$ be the vector of m features extracted from signal measurements performed at time τ_i on the i^{th} equipment, with n_i indicating the total number of data acquisitions performed on the i^{th} equipment before its failure. The ground truth RUL of the i^{th} piece pf equipment at time τ_i will be referred to as $y_i(\tau_i), i = 1,\dots,N; \tau_i = 1,\dots,n_i$. We consider a case in which the failure thresholds for the extracted features are not known. In this setting, fault prognostics is framed as a regression problem: given the historical dataset U formed by N realizations (degradation trajectories) $\{\boldsymbol{x}_i(\tau_i), y_i(\tau_i),\ \tau_i = 1,\dots,n_i\}, i = 1,\dots,N$, of a stochastic process $\big(\boldsymbol{X}(\tau), Y(\tau)\big) \in \mathbb{R}^m x\ (0,+\infty)$, our task is to find a function $\tilde{f}: \mathbb{R}^m \to (0,+\infty)$ such that it associates to a test pattern $x_{test}(\tau_{test}) \in \mathbb{R}^m$, the corresponding output

$y_{test}(\tau_{test})$. In what follows, we refer to $\tilde{f}$ as base model or base learner (Zhou, 2012).

3 ENSEMBLE LEARNING FOR FAULT PROGNOSTICS

In contrast to ordinary learning approaches which try to construct one base learner from training data, ensemble methods try to construct a set of learners $\widetilde{f_1}, \dots, \widetilde{f_H}$ and combine them to obtain an ensemble learner $\widetilde{f_{ens}}$. In this work, we consider combination of base learners based on weighted averaging (Zhou, 2012), i.e., the combined output $\widetilde{f_{ens}}$ is obtained by averaging the output of the individual learners with different weights α_h, which implies that the different learners have different importance

$$\widetilde{f_{ens}}(\boldsymbol{x}(\tau)) = \sum_{h=1}^{H} \alpha_h \widetilde{f_h}(\boldsymbol{x}(\tau)) \quad (1)$$

where

$$\sum_{h=1}^{H} \alpha_h = 1; \quad \alpha_h \geq 0; \quad h = 1, \dots, H \quad (2)$$

3.1 *Error ambiguity decomposition*

In this Subsection, we motivate the use of ensemble learning to enhance RUL predictions of a test equipment. Referring to the ensemble generalization error as $GE(\widetilde{f_{ens}})$, one can show that the following error-ambiguity decomposition holds (for more details, see the Appendix):

$$GE(\widetilde{f_{ens}}) = \overline{GE}(h) - \overline{ambi}(h) \quad (3)$$

where $\overline{GE}(h) = \sum_{h=1}^{H} \alpha_h GE(\widetilde{f_h})$ is the weighted average of the h^{th} individual base learner generalization error $GE(\widetilde{f_h})$; and $\overline{ambi}(h) = \sum_{h=1}^{H} \alpha_h ambi(\widetilde{f_h})$ is the weighted average of the h^{th} individual base learner ambiguity $ambi(\widetilde{f_h})$ defined in Appendix. The quantity $ambi(\widetilde{f_h})$ quantifies how much the h^{th} base learner predictions ,$\widetilde{f_h}$, differ from the ensemble predictions. On the right-hand of Eq. (3), the first term $\overline{GE}(h)$ represents the individual learner average error, which depends on the generalization ability of individual base learners whereas the second term $\overline{ambi}(h)$ represents the ambiguity, which depends on the ensemble diversity. Since the second term is always positive, and it is subtracted from the first term, it is clear that the error of the ensemble will never be larger than the average error of the individual base learners. Further, Eq. (11) shows that the more accurate and the more diverse the individual learners, the better the ensemble.

3.2 *Ensemble Generation*

According to the error-ambiguity decomposition discussed in Subsection 3.1, ensemble diversity, i.e., the difference among the individual base learners is a fundamental issue in ensemble learning. Therefore, since complementarity is more important than pure accuracy (Zhou, 2012), an ensemble formed by only very accurate learners can provide worse performances than one formed by also some relatively weak learners. Two approaches are typically used to generate diverse base learners:

- *Homogeneous ensemble generation*: different base learners are generated using the same prognostic algorithm and diversity is achieved by manipulating data in different ways: subsampling from the training set (e.g., bagging ((Zhou, 2012))) or using different subsets of features.
- *Heterogeneous ensemble generation*: different base models are generated using different prognostic algorithms.

In this work, we have resorted to heterogeneous ensemble generation since it has been shown able to provide better performance than homogenous ensemble methods in cases of few low-dimensional data (Rathore & Kumal, 2017).

3.3 *Setting the ensemble base model weights* α_h

The data extracted from the available N run-to-failure degradation trajectories of similar components are divided into training, validation and test subsets, formed by P_{train}, P_{valid} and P_{test} instances, respectively. The training subset is used to build the H individual base models, the validation subset to assign them weights to be used for the aggregation of the individual model outcomes (Eq. (1)) and the test subset to verify the final ensemble performance. The weight α_h associate to the h^{th} base learner is calculated based on its performance in predicting the RUL

of the validation set patterns. Performance is measured resorting to the Empirical Generalization Error (EGE), which for the h^{th} base learner is defined as the mean squared error on validation set patterns:

$$\widehat{GE}(\tilde{f}_h) = \frac{1}{P_{valid}} \sum_{p=1}^{P_{valid}} \frac{1}{n_p} \sum_{\tau_p=1}^{n_p} \left(y_p(\tau_p) - \tilde{f}_h\left(\boldsymbol{x_p}(\tau_p)\right) \right)^2 \quad (4)$$

In this work, we have considered weights proportional to the inverse of the EGE, i.e.,

$$\alpha_h = \frac{\frac{1}{\widehat{GE}(\tilde{f}_h)}}{\sum_{l=1}^{H} \frac{1}{\widehat{GE}(\tilde{f}_l)}} \quad h = 1, \dots, H \quad (5)$$

4 PROGNOSTIC PERFORMANCE METRICS

In addition to EGE, we have considered other performance metrics, which are typically considered (Rigamonti et al., 2017) for quantitatively assessing and comparing the point prediction performance of different prognostic algorithms (Saxena et al., 2009). A brief description of the implemented metrics is given hereafter considering a generic test trajectory $\left(\boldsymbol{x}(\tau), y(\tau)\right), \tau = 1, \dots, n$ and a general base learner $\tilde{f}$.

- *Relative Accuracy (RA)*:

$$R(\tilde{f}) = \sum_{\tau}^{n} \exp\left(-\frac{\left|\tilde{f}(x(\tau) - y(\tau)\right|}{y(\tau)} \right) \quad (6)$$

Notice that $R(\tilde{f})$ is in the range [0,1] and the larger the relative accuracy the more accurate is the model.

- *Precision:*

$$P = \sqrt{\frac{\sum_{\tau=1}^{n} (e(\tau) - \bar{e})^2}{n}} \quad (7)$$

$$e(\tau) = \tilde{f}(x(\tau)) - y(\tau) \quad (8)$$

$$\bar{e} = \frac{1}{n} \sum_{\tau=1}^{n} e(\tau) \quad (9)$$

This measure quantifies the dispersion (stability) of the prediction error around its mean. Closer to zero is the precision, more stable is the model.

5 CASE STUDY

This Section presents the results of the application of the proposed method to Tetra Pak® A3/Flex filling knife condition monitoring data.

We have available run-to failure-degradation trajectories from N = 10 different knives. For each knife, we have available $m = 2$ health indicators which have been extracted using the procedure presented in (Cannarile et al., 2017b).

In this work, a heterogeneous ensemble generation has been developed considering $H = 4$ prognostic algorithms:

- Gaussian Process Regression with Squared Exponential (GPRSE) covariance function;
- GRP with Matern 3/2 (GRPM) covariance function;
- Support Vector Regression with Gaussian Kernel (SVRGK);
- SVR with Quadratic Polynomial Kernel (SVRQPK).

These algorithms have been selected, since they have proved to be effective also when few training data with no clear patterns of regularity are available for training (Domingos, 2012). To properly compare the performance of the ensemble model with that of each base model, we have resorted to a twice nested Leave-One-Out-Cross-Validation (LOOCV) approach. The outer loop is to assess the performance of the ensemble and the single base learners, whereas the inner loop allows setting the weights $\alpha_h, h = 1,..,4$. In practice, the weights associates to the base learners are computed on each outer-validation set (using the inner LOOCV loop) and the final performance is measured on the corresponding outer-testing set (see Figure 1).

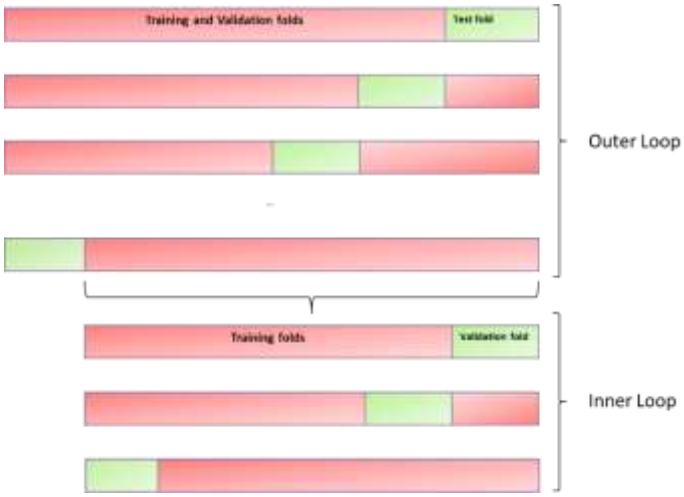

Figure 1: Twice nested LOOCV

Table 1 compares the performances of the developed ensemble model with that of the GRPM model, which has resulted to be the best performing individual model.

	Ensemble	GRPM
Empirical Generalization Error (EGE) (best value 0)	3.2991	3.7127
Relative Accuracy (RA) (best value 1)	0.8149	0.7804
Precision (best value 0)	0.0569	0.0633

Table 1: Comparison between the ensemble and the GRPM performances

Notice that the ensemble model performs better than GRPM in all the considered metrics. In particular, the average EGE is 11.14% lower (more satisfactory) than that of GRPM, the relative accuracy of the ensemble is 3.34% larger (more satisfactory) than that of GRPM, whereas, the two methods are comparable from the point of view of the precision. Finally, Figs. 2 and 3 show the RUL predicted by the ensemble and GRPM for two representative test trajectories.

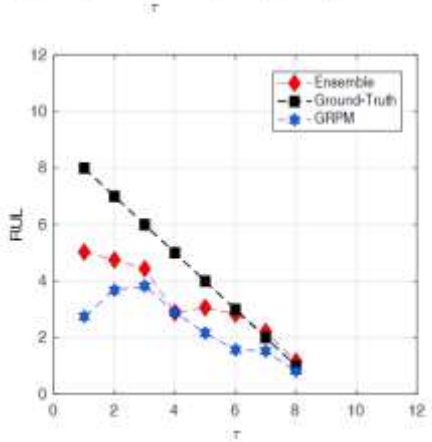

Figure 2: Predicted RUL by the ensemble (diamonds) and GRPM (exagon) for a test trajectory.

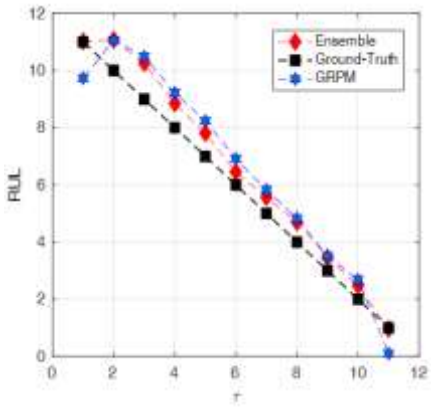

Figure 3: Predicted RUL by the ensemble (diamonds) and GRPM (exagon) for a test trajectory.

The most satisfactory ensemble predictions tend to be at the begininning of the life of the test knife. This is reflected by the great improvement of the EGE metric, which is more sensible to errors at the beginning of the run to failure trajectory than the relative accuracy.

6 CONCLUSIONS

In this work, we have developed a heterogeneous ensemble model for enhancing the accuracy of the RUL prediction of knives used in the packaging industry. Thanks to the diversity of the base learner algorithms, the proposed approach has been shown capable of reducing the generalization error and providing more accurate RUL predictions compared to that of each individual base learner in the ensemble.

7 REFERENCES

Baraldi P., Cadini F., Mangili F., Zio E., 2013a. Model-based and data-driven prognostics under different available information. Probabilistic Engineering Mechanics, 32, pp. 66-79.

Baraldi, P., Compare, M., Sauco, S., & Zio, E., 2013b. Ensemble neural network-based particle filtering for prognostics. Mechanical Systems and Signal Processing, 41(1), 288-300.

Baraldi, P., Mangili, F., Zio, E., 2015a. A belief function theory based approach to combining different representation of uncertainty in prognostics. Information Sciences, 303, pp. 134-149.

Baraldi, P., Mangili, F., Zio, E., 2015B. A prognostics approach to nuclear component degradation modeling based on Gaussian Process Regression. Progress in Nuclear Energy, 78, pp. 141-154.

Baraldi, P., Di Maio, F., Al-Dahidi, S., Zio, E., Mangili, F., 2017. Prediction of industrial equipment Remaining Useful Life by fuzzy similarity and belief function theory. Expert Systems with Applications, 83, pp. 226-241.

Brown, G., Wyatt, J., Harris, R., & Yao, X., 2005. Diversity creation methods: a survey and categorisation. Information Fusion, 6(1), 5-20.

Cannarile, F., Compare, M., Rossi, E., Zio, E., 2017a. A fuzzy expectation maximization based method for estimating the parameters of a multi-state degradation model from imprecise maintenance outcomes. Annals of Nuclear Energy, 110, pp. 739-752.

Cannarile, F., Baraldi., P., Compare, M., Borghi, D., Capelli, L., Cocconcelli, M., Lahrac, A., Zio, E., 2017. An unsupervised clustering method for assessing the degradation state of cutting tools in the packaging industry. Safety and Reliability: Theory and Application- Proceedings of the European Safety and Reliability Conference, ESREL 2017.

Cannarile, F., Compare, M., Baraldi, P., Di Maio, F., Zio, E., 2018. Homogeneous continuous-time, finite-state, hidden semi-Markov modelling for enhancing Empirical Classification System diagnostics of industrial components. Probabilistic Engineering Mechanics, under review.

Di Maio, F., Tsui, K.L., Zio, E., 2012. Combining Relevance Vector Machines and exponential regression for bearing residual life estimation. Mechanical Systems and Signal Processing, 31, pp. 405-427.

Domingos, P., 2012. A few useful things to know about machine learning, Communications of the ACM, 55 (10), pp. 78-87.

Gorjian, N., Ma, L., Mittinty, M., Yarlagadda, P., Sun, Y., 2009. A review on degradation models in reliability analysis. Engineering Asset Lifecycle Management - Proceedings of the 4th World Congress on Engineering Asset Management, WCEAM 2009, pp. 369-384.

Liu, Y., An, A., Huang, X., 2006. Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. In PAKDD 6, pp. 107-118.

Liu, Y., Zhang, Z., & Chen, J., 2015. Ensemble local kernel learning for online prediction of distributed product outputs in chemical processes. Chemical Engineering Science, 137, 140-151.

Rathore, S.S., Kumar, S., 2017. Linear and non-linear heterogeneous ensemble methods to predict the number of faults in software systems. Knowledge-Based Systems, 119, pp. 232-256.

Rigamonti, M., Baraldi, P., Zio, E., Roychoudhury, I., Goebel, K., Poll, S., 2017. Ensemble of optimized echo state networks for remaining useful life prediction. Neurocomputing, Article in Press.

Saxena, A., Celaya, J., Saha, B., Saha, S., Goebel, K., 2009. Evaluating algorithm performance metrics tailored for prognostics. In Aerospace conference, 2009 IEEE, pp. 1-13.

Wang, P., Vachtsevanos, G., 2001. Fault prognostics using dynamic wavelet neural networks. Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM, 15 (4), pp. 349-365.

Xing, Y., Ma, E. W., Tsui, K. L., Pecht, M., 2013. An ensemble model for predicting the remaining useful performance of lithium-ion batteries. Microelectronics Reliability, 53(6), 811-820.

Yang, Z., Baraldi, P., Zio, E., 2017. A comparison between extreme learning machine and artificial neural network for remaining useful life prediction. Proceedings of 2016 Prognostics and System Health Management Conference, PHM-Chengdu 2016.

Zhou, Z., 2012. Ensemble Methods: Foundations and Algorithms, Chapman Hall/CRC.

Zio, E., 2016. Some challenges and opportunities in reliability engineering. IEEE Transactions on Reliability, 65 (4), pp. 1769-1782.

APPENDIX

Given an instance $\boldsymbol{x} = \boldsymbol{x}(\tau)$, the ambiguity of the individual base learner $\widetilde{f_h}$ is defined as

$$ambi\left(\widetilde{f_h}|\boldsymbol{x}\right) = \left(\widetilde{f_h}(\boldsymbol{x}) - \widetilde{f_{ens}}(\boldsymbol{x})\right)^2 \quad h = 1, \dots, H \tag{10}$$

and the ambiguity of the ensemble is

$$\begin{aligned} \overline{ambi}\left(\widetilde{f_{ens}}|\boldsymbol{x}\right) &= \sum_{h=1}^{H} \alpha_h ambi\left(\widetilde{f_h}|\boldsymbol{x}\right) = \\ &= \sum_{h=1}^{H} \alpha_h \left(\widetilde{f_h}(\boldsymbol{x}) - \widetilde{f_{ens}}(\boldsymbol{x})\right)^2 \end{aligned} \tag{11}$$

The ambiguity term measures the disagreement among the individual base learners on instance $\boldsymbol{x}$. If we use the Squared Error (SE) to measure the performance, then, the error of the individual base learner $\widetilde{f_h}$ and the ensemble $\widetilde{f_{ens}}$ are, respectively,

$$SE(\widetilde{f_h}|\boldsymbol{x}) = \left(\widetilde{f_h}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \quad h = 1, \dots, H \tag{12}$$

$$SE(\widetilde{f_{ens}}|\boldsymbol{x}) = \left(\widetilde{f_{ens}}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \tag{13}$$

Then, one can show that (Zhou, 2012)

$$\overline{ambi}(\widetilde{f_{ens}}|\boldsymbol{x}) = \overline{SE}(\tilde{h}|\boldsymbol{x}) - SE(\widetilde{f_{ens}}|\boldsymbol{x}) \tag{14}$$

where $\overline{SE}(\tilde{h}|\boldsymbol{x}) = \sum_{h=1}^{H} \alpha_h \, SE(\widetilde{f_h}|\boldsymbol{x})$ is the weighted average of the individual base learner errors. Since Eq. (14), holds for every instance $\boldsymbol{x}$, after averaging over the input distribution $p(\boldsymbol{x})$ from which the instances are sampled, it still holds that

$$\begin{aligned} \sum_{h=1}^{H} \alpha_h \int ambi(\widetilde{f_h}|\boldsymbol{x}) p(\boldsymbol{x}) \, d\boldsymbol{x} = \\ = \sum_{h=1}^{H} \alpha_h \int SE(\widetilde{f_h}|\boldsymbol{x}) p(\boldsymbol{x}) \, d\boldsymbol{x} \\ - \int SE(\widetilde{f_{ens}}|\boldsymbol{x}) p(\boldsymbol{x}) \, d\boldsymbol{x} \end{aligned} \tag{15}$$

The generalization error and the ambiguity of the individual base learner $\widetilde{f_h}$, can be written as, respectively,

$$GE(\widetilde{f_h}) = \int SE(\widetilde{f_h}|\boldsymbol{x}) p(\boldsymbol{x}) \, d\boldsymbol{x} \quad h = 1, \dots, H \tag{16}$$

$$ambi(\widetilde{f_h}) = \int_{h = 1, \dots, H} ambi(\widetilde{f_h}|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \tag{17}$$

Similarly, the generalization error of the ensemble reads

$$GE(\widetilde{f_{ens}}) = \int SE(\widetilde{f_{ens}}|\boldsymbol{x}) p(\boldsymbol{x}) \, d\boldsymbol{x} \quad h = 1, \dots, H \tag{18}$$

Based on the notation just introduced and Eq. (14), we obtain the error-ambiguity decomposition (Zhou, 2012):

$$GE(\widetilde{f_{ens}}) = \overline{GE}(h) - \overline{ambi}(h) \tag{19}$$