Physica A 681 (2026) 131083

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Hybrid ensemble machine learning models☆

Paolo Giudici [a,*], Francesca Mariani [b], Gloria Polinesi [b]

[a] University of Pavia, Department of Economics, Via San Felice, 5, Pavia, 27100, Italy
[b] Marche Polytechnic University, Department of Economic and Social Sciences (DISES), Piazzale Martelli Raffaele, 8, Ancona, 60121, Italy

ARTICLE INFO

Keywords:
Ensemble
Model averaging
Predictive accuracy
Mean squared error

ABSTRACT

Machine learning models are usually assessed and compared in terms of predictive performance. Ensemble models, which average the predictions obtained from different models, often improve such performance. In this paper we show how to further improve the predictive accuracy of ensemble models, and allow them to achieve strong performance without retraining. To this aim we leverage the diversity among individual models, expressed by their covariance, computed on a subsample of the data ordered by the best model. We illustrate our proposal with applications to real data.

1. Introduction

We are interested in predictive problems, in which we would like to predict the values of an output variable using a machine learning model trained on a set of input variables. These problems can be of two main different types: classification and regression problems. In classification problems, the output variable is categorical, and the model learns to predict a category of the output, based on a set of input feature values, by assigning a real-valued score to the target units. The score represents the probability that the unit belongs to one of the classes of the target response. In regression problems, the model learns to predict the numerical value of the output, based on a set of input feature values, by directly assigning a real-value prediction to each target unit. In this case, the goal is to predict a continuous numeric value rather than a class label.

Machine learning models are typically assessed in terms of their predictive accuracy, comparing the predicted values with the actual ("true") values. Accuracy requires that the output of the machine learning model is close to the observed (or expected) output. For example, in the U.S. National Institute of Standards and Technology Risk Management Framework [1], accuracy is defined as the closeness of results or estimates to the true values or to the values accepted as being true.

The measurement of predictive accuracy is well known in the statistics and machine learning literature (see, e.g., [2,3]). In most papers and applied works the mean squared error (MSE) is employed as an accuracy metric for a continuous response; the Area Under the Receiver Operating Characteristic curve (AUROC) for a binary response. While the MSE measures how calibrated the model predictions are with respect to the true response values, the AUROC measures how concordant the model-predicted ranks are with respect to the true response ranks, so that units are appropriately classified.

To allow assessing a model independently of the nature of the underlying variable, the Brier score [4] was introduced to assess classification models in terms of the calibration distance underlying the MSE, and the Rank Graduation Accuracy measure [5] was introduced to assess regression problems extending the concordance notion underlying the Area Under the Curve.

In this paper we propose a methodology aimed at improving the accuracy of test-set predictions, in a model-agnostic way, and without retraining. To measure predictive accuracy we employ the MSE.

☆ This article is part of a Special issue entitled: 'Statistical mechanics for Artificial Learning' published in Physica A.

* Corresponding author.
E-mail address: paolo.giudici@unipv.it (P. Giudici).

https://doi.org/10.1016/j.physa.2025.131083

Available online 4 November 2025

Many studies in the literature have found that ensemble models, which average the predictions obtained from different models, often improve predictive accuracy of individual models (see, e.g., [6]). For example, substantial improvements in classification and regression accuracy have been attained through the use of ensembles of decision trees. Among them, the Random Forest algorithm (see, e.g., [7]) constructs multiple decision trees on random subsets of the data and averages their predictions to mitigate overfitting, whereas Gradient Boosting (see, e.g., [8]) builds trees sequentially, with each tree aiming to correct the errors of its predecessors, thereby progressively enhancing predictive performance.

Ensemble machine learning methods have recently found wide applications across different areas of physics, offering improved predictive accuracy compared to single models. For example, in high-energy physics, boosted decision trees and gradient boosting ensembles (e.g., XGBoost, LightGBM) are routinely applied to classify rare events such as Higgs boson production against background processes [9]. In astrophysics, ensembles of random forests and neural networks have enhanced the classification of supernovae and variable stars, as well as the estimation of photometric redshifts from incomplete and noisy survey data [10]. In materials science, stacked ensembles combining random forests, kernel methods, and boosting have achieved state-of-the-art predictions of electronic properties such as band gaps and superconductivity from high-throughput DFT data [11]. In nuclear physics, ensemble models integrating random forests, Gaussian processes, and neural networks have been used to predict nuclear masses and half-lives with improved extrapolation and quantified uncertainty [12]. In climate physics, ensembles of deep networks and random forests have been employed for weather downscaling and extreme event forecasting, improving over conventional single-model approaches [13]. Finally, in quantum physics, ensembles of random forests, boosted trees, and convolutional neural networks have been successfully applied to identify quantum phases of matter from simulation data [14] and to improve quantum error correction in surface codes [15].

Beyond the choice of base algorithms used in ensemble methods, techniques such as stacking and performance-weighted averaging have also been widely explored. Stacking combines multiple base models by training a meta-learner to optimally aggregate their predictions, leveraging the strengths of each component model (see [16]). Performance-weighted averaging [17], on the other hand, assigns weights to individual model predictions based on their relative accuracy, thereby allowing better-performing models to have a greater influence on the final output.

In this paper we propose a *hybrid* approach that averages model predictions for some observations, and employs individual predictions for others. We show empirically that such an approach can further improve predictive accuracy. Our proposal is fundamentally different from stacking or performance-weighted averaging methods, as our hybrid approach enhances predictions independently of the original training data. By doing so, it allows for the flexible integration of predictions from multiple models (including those generated by ensemble methods), thereby offering an additional layer of refinement and potential accuracy gains.

2. Methodology

In this section, we compare *complete* ensemble machine learning models (those that use all available predictions) with *reduced* ensemble machine learning models, which use only subsets of the available predictions. We also highlight how the covariance between reduced-ensemble predictions influences this comparison.

Let $S = \{1, 2, \dots, n\}$ denote the set of units, and $\mathcal{M} = \{1, 2, \dots, M\}$ the set of machine learning models. Let $o = \{o_i, i \in S\}$ represent the true outcomes, and $p_j = \{p_{ji}, i \in S\}$ the predictions produced by the jth model. The model error is defined as $e_{ji} = o_i - p_{ji}$ for $j \in \mathcal{M}$ and $i \in S$.

Definition 1. The prediction of a complete ensemble model for the outcome o_i, $i \in S$, is defined as

$$p_{ai} = \sum_{j=1}^{M} w_j p_{ji}, \tag{1}$$

where $w_j \geq 0$ and $\sum_{j=1}^{M} w_j = 1$.

Definition 2. Let $0 < M_1 < M$. The predictions of the reduced ensemble models for the outcome o_i, $i \in S$, are given by

$$p_{a_1 i} = \frac{1}{W} \sum_{j=1}^{M_1} w_j p_{ji}, \tag{2}$$

$$p_{a_2 i} = \frac{1}{1-W} \sum_{j=M_1+1}^{M} w_j p_{ji}, \tag{3}$$

where $W = \sum_{j=1}^{M_1} w_j$ and $W \in [0, 1]$.

Proposition 1. *The mean squared error (MSE) of the complete ensemble can be expressed in terms of the reduced-ensemble MSE as*

$$\mathrm{MSE}_a = W^2 \, \mathrm{MSE}_{a_1} + (1-W)^2 \, \mathrm{MSE}_{a_2} + 2W(1-W)C, \tag{4}$$

where C denotes the covariance between the reduced models.

Proof. See Appendix.

Proposition 1 extends the result of Liao and Moody [18]. For terminology, hereafter we refer to C in Eq. (4) as the *reduced-model covariance*.

Proposition 1 provides a criterion for evaluating the performance of a complete ensemble relative to the reduced models. The key insight is that the reduced-model covariance plays a key role in determining whether it is preferable to use a complete model rather than reduced models.

Following this intuition, we propose a hybrid approach that uses the complete ensemble model for units with low reduced-model covariance and the best reduced model for units with high reduced-model covariance.

Without loss of generality, we consider only two reduced models, a_1 and a_2, and assume that a_1 is the best model, that is: $\text{MSE}_{a_1} \leq \text{MSE}_{a_2}$. The reduced-model covariance, C, is estimated as the sample covariance between the predictions p_{a_1} and p_{a_2}.

Our objective is to partition the units into two subgroups, $\mathcal{A}$ and $\bar{\mathcal{A}} = S \setminus \mathcal{A}$. The set $\mathcal{A}$ contains the units with low reduced-model covariance, best predicted by the complete ensemble p_a. The set $\bar{\mathcal{A}}$, instead, contains the units with high reduced-model covariance, best predicted by the reduced ensemble p_{a_1}.

Empirically, the cumulative reduced-model covariance among the predictions tends to increase for the most accurate (top-ranked) predictions. Hence, cumulative reduced-model covariance is expected to rise as predictions become more accurate.

To identify the two groups, $\mathcal{A}$ and $\bar{\mathcal{A}}$, we sort the units in non-decreasing order according to p_{a_1}, and iteratively compute the cumulative reduced-model covariance across the first k units, for $k = n, n-1, \dots, 1$. The aim is to detect units that, when added, increase the cumulative covariance between p_{a_1} and p_{a_2}. Units that increase cumulative reduced-model covariance are assigned to $\bar{\mathcal{A}}$, while the remaining units belong to $\mathcal{A}$.

More precisely, starting from $k = n$, for each unit at position k in the ordered sequence, we compute the difference between the cumulative reduced-model covariances obtained using the first k and the first $k-1$ units, respectively. If the difference is positive, the inclusion of the current (kth) unit increases the cumulative reduced-model covariance, and we assign it to the set $\bar{\mathcal{A}}$. If the difference is non-positive, meaning that the cumulative reduced-model covariance does not increase, the unit is assigned to the set $\mathcal{A}$.

It is important to note that once a unit is assigned to $\bar{\mathcal{A}}$, that unit and all subsequent units in the current ordered sequence are excluded from further consideration in subsequent iterations of the algorithm.

After iterating through all units, the procedure yields two groups. The set $\mathcal{A}$ contains the units that did not increase the cumulative reduced-model covariance during the iterative process. The set $\bar{\mathcal{A}}$ contains the units that did contribute to an increase of the cumulative reduced-model covariance.

In this manner, $\bar{\mathcal{A}}$ consists of the units that strengthen the reduced-model covariance. Conversely, $\mathcal{A}$ contains units that do not strengthen the reduced-model covariance, or that can weaken it.

We expect the reduced-model covariance within $\mathcal{A}$ to be lower, since the units in this set do not contribute to an increase in cumulative reduced-model covariance during the iterative process. Conversely, the reduced-model covariance within $\bar{\mathcal{A}}$ should be higher, as these units reinforce the pattern established by the preceding units in the p_{a_1} ranking (see Fig. 1).

Formally, let π_{a_1} denote the permutation of S that orders p_{a_1} in non-decreasing order: $p_{a_1\pi_{a_1}(1)} \leq p_{a_1\pi_{a_1}(2)} \leq \cdots \leq p_{a_1\pi_{a_1}(n)}$. Note that $p_{a_1\pi_{a_1}(k)}$ is the kth order statistic of the predictions p_{a_1}.

The cumulative reduced-model covariance between p_{a_1} and p_{a_2} for the first k units is:

$$s_{a_1 a_2 \pi_{a_1}(k)} = \frac{1}{k} \sum_{i=1}^{k} \left(p_{a_1 \pi_{a_1}(i)} - \bar{p}_{a_1 \pi_{a_1}(i)}\right)\left(p_{a_2 \pi_{a_1}(i)} - \bar{p}_{a_2 \pi_{a_1}(i)}\right), \tag{5}$$

where $\bar{p}_{a_1\pi_{a_1}(i)}$ and $\bar{p}_{a_2\pi_{a_1}(i)}$ denote the mean predictions of the first i units, for $i = 1, 2, \dots, n$.

The algorithm proceeds iteratively, starting from $k = n$. At each step, we compute the difference between consecutive cumulative reduced-model covariances: $s_{a_1 a_2 \pi_{a_1}(k)} - s_{a_1 a_2 \pi_{a_1}(k-1)}$. If this difference is positive, the kth unit increases covariance and is assigned to $\bar{\mathcal{A}}$; otherwise, it is assigned to $\mathcal{A}$. The process continues until $k = 2$.

Finally, the algorithm assigns to each unit the corresponding prediction according to

$$\hat{p}_i = \begin{cases} p_{ai}, & i \in \mathcal{A}, \\ p_{a_1 i}, & i \in \bar{\mathcal{A}}. \end{cases} \tag{6}$$

We name $\hat{p}$ the *Hybrid Ensemble* prediction. By construction, the reduced-model covariance within $\mathcal{A}$ is expected to be lower than that within $\bar{\mathcal{A}}$.

We can now state the following proposition.

Proposition 2. *Let $C_{\mathcal{A}}$ and $C_{\bar{\mathcal{A}}}$ denote the covariance between the reduced-ensemble model predictions p_{a_1} and p_{a_2} within the subsets $\mathcal{A}$ and $\bar{\mathcal{A}}$, respectively. Similarly, let $\text{MSE}_{a_1}|_{\mathcal{A}}$ and $\text{MSE}_{a_1}|_{\bar{\mathcal{A}}}$ represent the mean squared error of p_{a_1} computed over the sets $\mathcal{A}$ and $\bar{\mathcal{A}}$, respectively.*

If $C_{\mathcal{A}} < C_{\bar{\mathcal{A}}}$ and $\text{MSE}_{a_1}|_{\mathcal{A}} > \text{MSE}_{a_1}|_{\bar{\mathcal{A}}}$, the mean squared error of the hybrid prediction, $\widehat{\text{MSE}}$, satisfies the following inequality:

$$\widehat{\text{MSE}} \leq \min(\text{MSE}_a, \text{MSE}_{a_1}). \tag{7}$$

Proof. See Appendix.

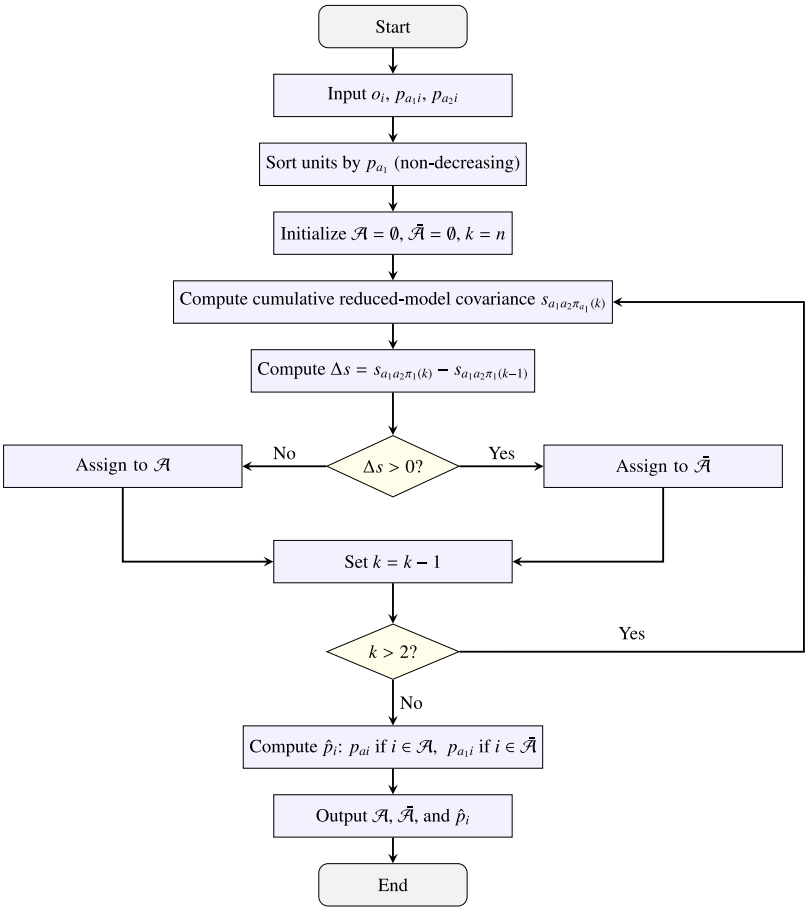

Fig. 1. Flowchart of the Hybrid Ensemble Model Selection Algorithm.

So far we have considered regression problems, in which we aim to predict a continuous response variable. However, the same reasoning can be extended to classification tasks, such as binary or categorical forecasts. In such cases, the predicted classes can be replaced with their corresponding predicted probabilities, pr, leading to a formulation based on the Brier score:

Definition 3. The Brier score associated with the probabilities pr_i, $i \in S$, is defined as

$$\text{BS}(pr) = \frac{1}{n}\sum_{i=1}^{n}(pr_i - o_i)^2. \tag{8}$$

Comparing the definition of the MSE with Definition 3 it becomes clear that the two formulations are identical when the predicted classes p_i are replaced by the corresponding predicted probabilities (or scores), pr_i. Consequently, the results established for the mean squared error directly extend to the Brier score.

3. Application

In this section we illustrate, by means of three experiments, the performance of the proposed hybrid ensemble model, as defined in Eq. (6).

As previously discussed, the approach is model-agnostic and does not require retraining; instead, it relies only on averaging model predictions over the test set. Without loss of generality, in 2 we choose $M_1 = 1$ (i.e., $p_{a_1} = p_1$). For the complete ensemble defined in Definition 1, we choose equal weights ($w_j = 1/M$, $j = 1, 2, \ldots, M$). This ensures a neutral combination, avoiding any prior assumptions about the relative performance of individual models. It allows the comparison to focus on the role of the covariance structure among model predictions rather than on weight optimization.

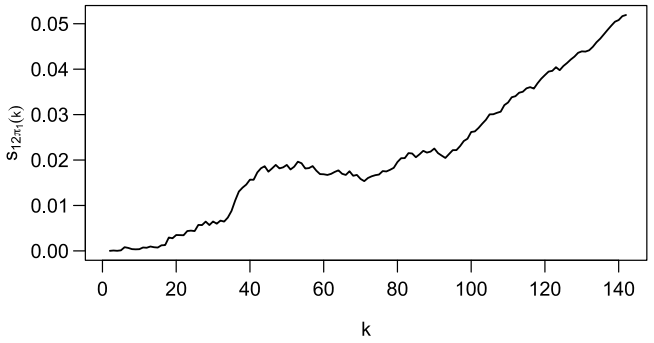

Fig. 2. Cumulative covariance, $s_{12\pi_1(k)}$, between p_1 and p_2 according to the order determined by the best reduced model, for $k = 1, 2, \ldots, n$..

Table 1
Brier scores for the RF (p_1), the LR (p_2), their average (p_a) and our proposal ($\hat{p}$), in predicting whether salary doubles.

Model	Brier score	AUC
p_1	0.1971	0.7708
p_2	0.2060	0.7507
p_a	0.1953	0.7745
$\hat{p}$	**0.1920**	**0.7783**

Experiment 1: Binary classification on the employee dataset

We address a classification problem with binary outcomes using the publicly available `Employee` dataset from the `stima` R package. The dataset contains demographic variables relative to 473 bank employees. We aim to predict whether the employee's salaries have doubled in the considered time period. We consider two alternative machine learning models: a Random Forest (RF, p_1) and a Logistic Regression (LR, p_2), and we compare them against the complete-ensemble predictions and against our hybrid ensemble proposal, defined in Eq. (6). More details on the data and on the chosen machine learning models are contained in [5].

The predictions are ordered according to the best-performing model (p_1), from lowest to highest accuracy, in non-decreasing order.

Fig. 2 shows the cumulative covariance, $s_{12\,\pi_1(k)}$, between p_1 and p_2 for the first k units which are ordered according to the best model's prediction (p_1).

Looking at Fig. 2, note that the cumulative reduced-model covariance increases as k increases. Our improvement is obtained by averaging the two predictions for the units with small k values (i.e., in the set $\mathcal{A}$) and using the best model for the remaining values of k (in $\bar{\mathcal{A}}$).

To assess the advantage of our proposed hybrid ensemble model from an empirical viewpoint, we compare its performance against the complete ensemble model in Table 1 based on Brier scores. For completeness, AUC values are also reported.

Table 1 shows the performance of our proposal: it reduces the Brier score in comparison with both the complete ensemble model p_a and the best reduced model (p_1). Consistently, it achieves the highest AUC.

Experiment 2: Regression for bitcoin daily prices

We consider a regression problem focused on predicting the daily Bitcoin prices from January 1, 2018 to April 30, 2018. We use as training set the prices from May 18, 2016 to December 31, 2017, as well as the prices of other assets (Gold and Oil), the S&P 500 index, and the exchange rates Dollar/Eur and Dollar/Yuan. To generate the predicted values, we employ five different neural network models: Gated Recurrent Unit (GRU) (p_1), Long Short-Term Memory (LSTM) (p_2), Radial Basis Function network (RBF) (p_3), Multilayer Perceptron (MLP) (p_4), Neural Network Autoregression (NNAR) (p_5). More details are contained in [19].

As in the previous case, we sort the predictions in non-decreasing order based on the best model (p_1), ranging from those with the lowest accuracy to those with the highest accuracy. In Fig. 3 we plot the cumulative covariances, $s_{1h\,\pi_1(k)}$, between p_1 and p_h, for $h = 1, 2, \ldots, 5$, and $k = 1, 2, \ldots, n$, according to the order determined by the best model (p_1).

Observing Fig. 3, we note that all cumulative covariances, with the exception of Model 5 (NNAR), increase as k increases. This means that, as we move towards the Bitcoin prices most accurately predicted by p_1, the cumulative covariance between p_1 and the predictions from models 2, 3 and 4 increases. This behavior suggests that our hybrid ensemble model will improve the predictions when applied to an ensemble constructed from the first four models.

To empirically verify this, Table 2 shows the values of the mean squared error for the five individual models (p_1–p_5, ranked by accuracy). Table 2 reveals that the GRU model (p_1) achieves the lowest MSE, followed by the LSTM model (p_2).

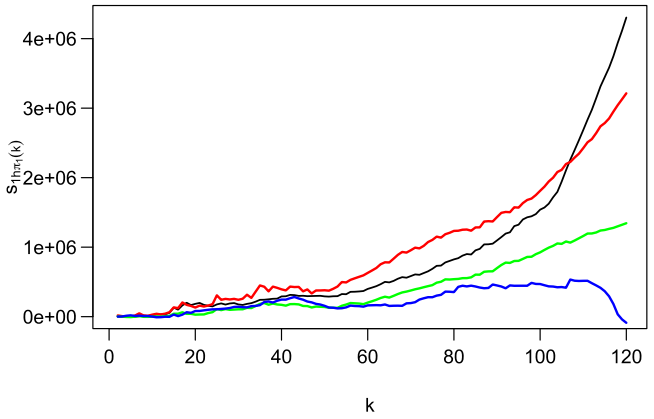

Fig. 3. Cumulative covariances, $s_{1h\pi_1(k)}$, between p_1 and p_h according to the order determined by the best model, for $k = 1, 2, \ldots, n$, and $h = 2$ (LSTM) (in black) $h = 3$ (RBF) (in red) $h = 4$ (MLP) (in green) $h = 5$ (NNAR) (in blue).

Table 2
Values of the mean squared error for all individual models.

Model	p_1	p_2	p_3	p_4	p_5
MSE	**712678**	1279435	3538016	4037740	6805592

Table 3
Values of the mean squared error for different complete ensemble models and corresponding hybrid models (in bold).

Model	$p_a(5)$	$\hat{p}(5)$	$p_a(4)$	$\hat{p}(4)$	$p_a(3)$	$\hat{p}(3)$	$p_a(2)$	$\hat{p}(2)$
MSE	1934680	**685310**	1464435	**684724**	1215839	**704291**	793755	**699863**

Table 4
Brier scores for the RF (p_1), the LR (p_2) and Stacking (p_3), their average (p_a) and our proposal ($\hat{p}$), in predicting whether salary doubles.

Model	Brier score	AUC
p_1	0.1971	0.7708
p_2	0.2060	0.7507
p_3	0.1985	0.7717
p_a	0.1960	0.7737
$\hat{p}$	**0.1920**	**0.7785**

Experiment 3: Adding a stacking model

The third experiment uses the same dataset as the first experiment but incorporates the predictions of an additional machine learning model: a Stacking Ensemble. This ensemble combines the Random Forest (p_1) and Logistic Regression (p_2) models by training a meta-learner to optimally aggregate their individual predictions (see Table 3). . To implement the stacking, we split the dataset randomly into a 70% training set and a 30% test set. The training set is then divided into five folds to obtain out-of-sample predictions from the base models (Random Forest and Logistic Regression). For each fold, we train the base models and collect their out-of-sample predictions. These predictions form the features of the stacking dataset, with the original response variable (salary doubling) as the target. Using this dataset, we construct a Logit meta-learner to generate the final stacking predictions (p_3).

Based on the Brier score, the stacking model exhibits performance inferior to the individual Random Forest model; hence Random Forest remains the best individual model. We therefore use p_1 as the best model ($p_{a_1} = p_1$) and, analogous to the previous experiments, compute the predictions for the complete ensemble model (p_a) by taking the simple arithmetic mean of the individual predictions, and implement our hybrid model (see Table 4).

4. Conclusions

We have investigated whether ensemble predictions can improve individual predictions, and found that the answer depends on the covariance between the predictions of the individual models. For data points where the covariance among model predictions is high, it is preferable to rely on the best model alone, as averaging highly divergent predictions can lead to poorer results. Conversely,

when the same covariance is low, the model predictions are more consistent with each other, making model averaging a more effective strategy.

These observations have led to the proposal of a hybrid ensemble model that averages models only when the reduced-model covariance is low, and otherwise selects the best model. Applications to two real machine-learning problems reveal its effectiveness in reducing the mean squared error and/or the Brier score, not only relative to a classic ensemble model, but also relative to the best model.

More generally, our paper demonstrates the advantages of a hybrid ensemble machine learning framework that selectively combines model predictions based on reduced covariance structures. Our empirical findings show that this strategy can consistently outperform both traditional ensemble methods and best-performing individual models, across both classification and regression tasks.

Looking forward, potential directions for future research include extending the approach to other metrics, such as F1 scores for unbalanced data, or Kullback–Leibler divergences and entropy measures for multiclass responses. Future work could also include extending the approach to a dynamic setting, where prediction accuracy and model covariances evolve over time. Concerning the covariance structure, which plays a key role in our proposal, future research could explore varying covariance matrices conditional on regime switches, or the application of random matrix theory to filter the matrix itself. Another relevant avenue is to address the explainability of the output from ensemble models using methods such as feature importance and Shapley values. From a robustness viewpoint, the approach could be extended within a Bayesian framework to capture model uncertainty. Finally, the proposal could substantially benefit from applications in physics, energy, finance and health care, where domain-specific loss functions, constraints, or priors could be incorporated to further improve performance and reliability.

CRediT authorship contribution statement

Paolo Giudici: Writing – review & editing, Supervision, Conceptualization. **Francesca Mariani:** Writing – original draft, Software, Methodology, Formal analysis. **Gloria Polinesi:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded by the European Union – NextGenerationEU, in the framework of GRINS – Growing Resilient, INclusive and Sustainable (GRINS PE00000018). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union. The paper is the result of a close collaboration between the authors; however, F.M. elaborated and wrote Sections 2, 3 and 5; P.G. elaborated and wrote Sections 1 and 4 and supervised the work; G.P. contributed by conducting a literature review and proofreading the paper.

Appendix

Proof of Proposition 1. The MSE of the complete ensemble is

$$\mathrm{MSE}_a = \frac{1}{n}\sum_{i=1}^{n}(p_{ai} - o_i)^2. \tag{9}$$

The complete-ensemble prediction can be rewritten as

$$p_{ai} = W p_{a_1 i} + (1 - W) p_{a_2 i}, \quad i \in S. \tag{10}$$

Substituting (10) into Eq. (9) we obtain

$$\mathrm{MSE}_a = \frac{1}{n}\sum_{i=1}^{n}\left[W(p_{a_1 i} - o_i) + (1 - W)(p_{a_2 i} - o_i)\right]^2 \tag{11}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big[\, W^2(p_{a_1 i} - o_i)^2 + (1 - W)^2(p_{a_2 i} - o_i)^2 \tag{12}$$

$$+ 2W(1 - W)(p_{a_1 i} - o_i)(p_{a_2 i} - o_i)\,\Big] \tag{13}$$

$$= W^2\frac{1}{n}\sum_{i=1}^{n} e_{a_1 i}^2 + (1 - W)^2\frac{1}{n}\sum_{i=1}^{n} e_{a_2 i}^2 + 2W(1 - W)\frac{1}{n}\sum_{i=1}^{n} e_{a_1 i} e_{a_2 i} \tag{14}$$

$$= W^2\,\mathrm{MSE}_{a_1} + (1 - W)^2\,\mathrm{MSE}_{a_2} + 2W(1 - W)\frac{1}{n}\sum_{i=1}^{n} e_{a_1 i} e_{a_2 i}. \tag{15}$$

Now, substituting (10), (2), (3) into Eq. (11) we obtain

$$\mathrm{MSE}_a = W^2\,\mathrm{MSE}_{a_1} + (1-W)^2\,\mathrm{MSE}_{a_2} \tag{16}$$

$$+ 2W(1-W)\frac{1}{nW(1-W)}\sum_{h=1}^{M_1}\sum_{k=M_1+1}^{M}\sum_{i=1}^{n} w_h w_k e_{hi} e_{ki} \tag{17}$$

$$= W^2\,\mathrm{MSE}_{a_1} + (1-W)^2\,\mathrm{MSE}_{a_2} + 2W(1-W)C. \tag{18}$$

This concludes the proof. □

Proof of Proposition 2. First we prove that $C_{\mathcal{A}} \le C \le C_{\bar{\mathcal{A}}}$ and $\mathrm{MSE}_{a_1}|_{\bar{\mathcal{A}}} \le \mathrm{MSE}_{a_1} \le \mathrm{MSE}_{a_1}|_{\mathcal{A}}$.

We write the ensemble-model covariance C in terms of $C_{\mathcal{A}}$ and $C_{\bar{\mathcal{A}}}$ as follows:

$$C = n_{\mathcal{A}} C_{\mathcal{A}} + \left(1 - n_{\mathcal{A}}\right) C_{\bar{\mathcal{A}}}, \tag{19}$$

where $n_{\mathcal{A}} = \frac{|\mathcal{A}|}{n}$ is the ratio between the cardinality of $\mathcal{A}$ and n.

If $C_{\mathcal{A}} \le C_{\bar{\mathcal{A}}}$, Eq. (19) implies that

$$\frac{C}{C_{\mathcal{A}}} = n_{\mathcal{A}} + (1-n_{\mathcal{A}})\frac{C_{\bar{\mathcal{A}}}}{C_{\mathcal{A}}} \ge n_{\mathcal{A}} + (1-n_{\mathcal{A}}) = 1. \tag{20}$$

Analogously,

$$\frac{C}{C_{\bar{\mathcal{A}}}} = n_{\mathcal{A}}\frac{C_{\mathcal{A}}}{C_{\bar{\mathcal{A}}}} + (1-n_{\mathcal{A}}) \le n_{\mathcal{A}} + (1-n_{\mathcal{A}}) = 1, \tag{21}$$

so $C_{\mathcal{A}} \le C \le C_{\bar{\mathcal{A}}}$. Using a similar argument one can prove that $\mathrm{MSE}_{a_1}|_{\bar{\mathcal{A}}} \le \mathrm{MSE}_{a_1} \le \mathrm{MSE}_{a_1}|_{\mathcal{A}}$.

If $\mathrm{MSE}_a \le \mathrm{MSE}_{a_1}$, Proposition 1 implies that

$$C \le \frac{(1+W)}{2W}\,\mathrm{MSE}_{a_1} - \frac{(1-W)}{2W}\,\mathrm{MSE}_{a_2}. \tag{22}$$

Bearing in mind that $C_{\mathcal{A}} \le C$ and $\mathrm{MSE}_{a_1} \le \mathrm{MSE}_{a_1}|_{\mathcal{A}}$, from Eq. (22) it follows that $\mathrm{MSE}_a|_{\mathcal{A}} \le \mathrm{MSE}_{a_1}|_{\mathcal{A}}$. As a consequence,

$$\widehat{\mathrm{MSE}} = n_{\mathcal{A}}\,\mathrm{MSE}_a|_{\mathcal{A}} + (1-n_{\mathcal{A}})\,\mathrm{MSE}_{a_1}|_{\bar{\mathcal{A}}} \le \mathrm{MSE}_{a_1}. \tag{23}$$

Otherwise, if $\mathrm{MSE}_{a_1} \le \mathrm{MSE}_a$, Proposition 1 implies that

$$C \ge \frac{(1+W)}{2W}\,\mathrm{MSE}_{a_1} - \frac{(1-W)}{2W}\,\mathrm{MSE}_{a_2}. \tag{24}$$

Now, using $C \le C_{\bar{\mathcal{A}}}$ and $\mathrm{MSE}_{a_1}|_{\bar{\mathcal{A}}} \le \mathrm{MSE}_{a_1}$, from Eq. (24) it follows that $\mathrm{MSE}_{a_1}|_{\bar{\mathcal{A}}} \le \mathrm{MSE}_a|_{\bar{\mathcal{A}}}$. As a consequence,

$$\widehat{\mathrm{MSE}} = n_{\mathcal{A}}\,\mathrm{MSE}_a|_{\mathcal{A}} + (1-n_{\mathcal{A}})\,\mathrm{MSE}_{a_1}|_{\bar{\mathcal{A}}} \le \mathrm{MSE}_a. \tag{25}$$

This concludes the proof. □

References

[1] N.I. for Standards Technology, AI risk management framework, 2023, see also. URL https://www.nist.gov/artificial-intelligence.

[2] T. Gneiting, Making and evaluating point forecasts, JOurnal the AMerican STatistical ASsociation 106 (494) (2011) 746–762, http://dx.doi.org/10.1198/jasa.2011.r10138.

[3] D. Hand, R. Till, A simple generalisation of the area Under the ROC curve for multiple class classification problem, MAchine LEarn. 45 (2001) 171–186, http://dx.doi.org/10.1023/A:1010920819831.

[4] G. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1) (1950) 1–3.

[5] P. Giudici, E. Raffinetti, Rga: a unified measure of predictive accuracy, Adv. Data Anal. Classif. (2024) http://dx.doi.org/10.1007/s11634-023-00574-2.0.

[6] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference and Prediction, Springer-Verlag, 2009.

[7] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[8] J.H. Friedman, Stochastic gradient boosting, Comput. Statist. Data Anal. 38 (4) (2002) 367–378.

[9] K. Albertsson, et al., Machine learning in high energy physics community white paper, J. Phys. Conf. Ser. 1085 (2018) 022008, http://dx.doi.org/10.1088/1742-6596/1085/2/022008.

[10] M. Lochner, et al., Photometric supernova classification with machine learning, Astrophys. J. Suppl. 225 (2) (2016) 31, http://dx.doi.org/10.3847/0067-0049/225/2/31.

[11] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (2018) 145301, http://dx.doi.org/10.1103/PhysRevLett.120.145301.

[12] R. Utama, J. Piekarewicz, H. Prosper, Nuclear mass predictions for the crustal composition of neutron stars: A bayesian neural network approach, Phys. Rev. C 93 (1) (2016) 014311, http://dx.doi.org/10.1103/PhysRevC.93.014311.

[13] S. Rasp, N. Thuerey, Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench, J. Adv. Model. Earth Syst. 13 (2) (2021) http://dx.doi.org/10.1029/2020MS002405.

[14] J. Carrasquilla, R.G. Melko, Machine learning phases of matter, Nat. Phys. 13 (2017) 431–434, http://dx.doi.org/10.1038/nphys4035.

[15] P. Baireuther, T.E. O'Brien, B. Tarasinski, C. Beenakker, Machine-learning-assisted correction of correlated qubit errors in a topological code, Quantum 2 (2018) 48, http://dx.doi.org/10.22331/q-2018-01-29-48.

[16] M.A. Muslim, Y. Dasril, H. Javed, W.F. Abror, D.A.A. Pertiwi, T. Mustaqim, et al., An ensemble stacking algorithm to improve model accuracy in bankruptcy prediction, J. Data Sci. Intell. Syst. 2 (2) (2024) 79–86.

[17] H. Zhu, G. Zou, Stability and l2-penalty in model averaging, J. Mach. Learn. Res. 25 (322) (2024) 1–59.
[18] Y. Liao, J. Moody, Constructing heterogeneous committes using input feature grouping: Application to economic forecasting, in: K.-R. Mller S. A. Solla (Ed.), Advances in Neural Information Processing Systems, M. Press, Cambridge, 2000, 12.
[19] P. Giudici, A. Piergallini, M.C. Recchioni, Explainable artificial intelligence methods for financial time series, physica a: Statistical mechanics and its applications, 2024, https://www.sciencedirect.com/science/article/pii/S037843712400685X.