Bad and Good Errors: Value-Weighted Skill Scores in Deep Ensemble Learning

Sabrina Guastavino, Michele Piana, and Federico Benvenuto

***Abstract*— Forecast verification is a crucial task for assessing the predictive power of prognostic model forecasts and it is usually implemented by checking quality-based skill scores. In this article, we propose a novel approach to realize forecast verification focusing not just on the forecast quality but rather on its value. Specifically, we introduce a strategy for assessing the severity of forecast errors based on the evidence that, on the one hand, a false alarm just anticipating an occurring event is better than one in the middle of consecutive nonoccurring events, and that, on the other hand, a miss of an isolated event has a worse impact than a miss of a single event, which is part of several consecutive occurrences. Relying on this idea, we introduce a notion of value-weighted skill scores giving greater importance to the value of the prediction rather than to its quality. Then, we introduce an ensemble strategy to maximize quality-based and value-weighted skill scores independently of one another. We test it on the predictions provided by deep learning methods for binary classification in the case of four applications concerned with pollution, space weather, stock price, and IoT data stream forecasting. Our experimental studies show that using the ensemble strategy for maximizing the value-weighted skill scores generally improves both the value and quality of the forecast.**

***Index Terms*— Deep learning, ensemble learning, forecast verification.**

I. Introduction

PREDICTING events over time applies to a number of fields, ranging from weather [1] and space weather forecasting [2], through environment [3] to stock market forecasting [4]. In all these frameworks, the goodness of prediction is usually defined in terms of the correspondence between forecasts and observations and is known in the literature as the forecast quality [5]. For binary predictions, forecast quality is typically measured by using skill scores based on a confusion matrix whose entries count the number of false and true negatives (FNs and TNs) and of false and true positives (FPs and TPs). Specifically, these skill scores rely on simple arithmetic formulas that compute in different ways the imbalance of the diagonal entries (representing the correct prediction) with respect to the off-diagonal entries (representing the incorrect prediction) of the confusion matrix. Typical skill scores for forecast quality are the true skill statistic (TSS) [6], the Heidke skill score (HSS) [7], and the critical success index (CSI) [8]. However, there is a different perspective to evaluate forecast, i.e., in terms of its usefulness to support the user while making a decision. This type of goodness is known in the literature as the forecast value [9] and two examples of scores for the quantitative evaluation of this prediction goodness are the cost value score and the relative value score introduced in [10]. Moreover, the field of cost-sensitive learning is devoted to take the misclassification costs (and possibly other types of cost) into consideration [11]. In this framework, the evaluation of the forecast value is commonly carried out on the basis of preassigned cost of FPs and FNs. However, such a cost depends on problem-related factors and its quantification usually involves the economic cost. In particular, it concerns the actions taken to restore the damage caused by a sudden event or to prevent the damage caused by a possible future event. There are no papers in the literature that we are aware of evaluating the forecast value of the prediction without an additional problem-dependent cost–benefit analysis.

In this article, we focus on binary predictions performed over time and we propose two novelties. First, we introduce a novel approach to evaluate the severity of prediction errors by considering that a false alarm predicting that an event will occur just before its actual occurrence, anyway eases the right decision from the user, while a delayed alarm is of little use. The idea is to exploit the sequential order, naturally given by the time, with which the prediction occurs in order to assign a cost for FPs and FNs. In this way, we classify errors on the basis of their importance and we get a novel notion of confusion matrix and related skill scores. In this new framework, the severity of errors depends on their impact on the decision making process and therefore we refer to these novel confusion matrix and skill scores as value-weighted. Such a strategy allows for evaluating the forecast value independently of a cost–benefit analysis.

The second novelty is the introduction of an ensemble strategy to select, among many different predictions, those that maximize the newly introduced value-weighted skill scores. It can be applied to any set of probabilistic predictions regardless of the method used for generating such predictions. In this work, we test this strategy on the predictions generated by deep learning methods by varying the training epoch.

Manuscript received 4 March 2021; revised 27 May 2022; accepted 19 June 2022. The work of Sabrina Guastavino was supported by a regional grant of the "Fondo Sociale Europeo," Regione Liguria. The work of Michele Piana and Federico Benvenuto was supported by ASI-INAF under Agreement n.2018-16-HH.0. *(Corresponding author: Michele Piana.)*

The authors are with the Dipartimento di Matematica, Università di Genova, 16146 Genoa, Italy (e-mail: guastavino@dima.unige.it; piana@dima.unige.it; benvenuto@dima.unige.it).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3186068.

Digital Object Identifier 10.1109/TNNLS.2022.3186068

This ensemble strategy generally yields better results (in terms of both forecast quality and value) when maximizing the value-weighted skill scores rather than the classical quality-based skill scores. We showed this better effectiveness in the case of four applications. We considered the problem of forecasting the air quality in highly polluted urban environments, with particular reference to the prediction of concentration of fine particles with diameters of 2.5 μm and smaller (PM2.5) [12]–[14]. The second problem was concerned with solar flare forecasting, i.e., the prediction of those explosive solar events that trigger most space weather phenomena [15]–[17]. The third application focuses on stock price forecasting and, in particular, on the prediction of "down" movements in the market [18]–[20]. Eventually, the fourth application concerns with IoT data streams [21].

The plan of the article is as follows. In Section II, we define the classical skill scores and we show that binary predictions of the same quality can have completely different forecast values. In Section III, we introduce a value-weighted confusion matrix and the corresponding value-weighted skill scores to quantify the forecast value of a binary prediction. In Section IV, we show how this value-weighted approach works to assess the performances of standard machine learning. Section V introduces the ensemble learning process and describes its application to pollution, space weather, stock price, and IoT data stream forecasting. Our conclusions are offered in Section VI.

II. Confusion Matrix and Skill Scores

The results of a binary classifier are usually evaluated by computing the confusion matrix, also known as contingency table. Let $\mathbb{M}_{2,2}(\mathbb{N})$ be the set of 2-D matrices with integers elements. Let $\mathbf{y} \in \{0,1\}^n$ be a binary vector representing the true label vector and let $\mathbf{p} \in \{0,1\}^n$ be a binary prediction. Then the confusion matrix $\mathbf{C} \in \mathbb{M}_{2,2}(\mathbb{N})$ is defined as

1) $\mathbf{C}_{1,1} = \#\{i \in \{1,\dots,n\} : y_i = 1, p_i = 1\}$.
2) $\mathbf{C}_{2,2} = \#\{i \in \{1,\dots,n\} : y_i = 0, p_i = 0\}$.
3) $\mathbf{C}_{1,2} = \#\{i \in \{1,\dots,n\} : y_i = 0, p_i = 1\}$.
4) $\mathbf{C}_{2,1} = \#\{i \in \{1,\dots,n\} : y_i = 1, p_i = 0\}$.

This definition implies that $\mathbf{C}_{1,1}$ computes the true positives (TPs), $\mathbf{C}_{2,2}$ computes the true negatives (TNs), $\mathbf{C}_{1,2}$ computes the false positives (FPs) and $\mathbf{C}_{2,1}$ computes the false negatives (FNs). From this confusion matrix several skill scores can be computed in order to evaluate the binary classifier performances. Given a confusion matrix $\mathbf{C}$, we denote with $S : \mathbb{M}_{2,2}(\mathbb{N}) \to \mathbb{R}$ a skill score defined on the matrix $\mathbf{C}$. Four frequently used skill scores are as follows.

1) Accuracy (ACC) [22]

$$\mathrm{ACC}(\mathbf{C}) = \frac{\mathbf{C}_{1,1} + \mathbf{C}_{2,2}}{\mathbf{C}_{1,1} + \mathbf{C}_{1,2} + \mathbf{C}_{2,1} + \mathbf{C}_{2,2}} \tag{1}$$

i.e., the ratio between the number of correct predictions over the total number of predictions. $\mathrm{ACC}(\mathbf{C}) \in [0,1]$ and the optimal value is 1.

2) True skill statistic (TSS) [23]

$$\mathrm{TSS}(\mathbf{C}) = \frac{\mathbf{C}_{1,1}}{\mathbf{C}_{1,1} + \mathbf{C}_{2,1}} - \frac{\mathbf{C}_{1,2}}{\mathbf{C}_{1,2} + \mathbf{C}_{2,2}} \tag{2}$$

i.e., the balance between the true positive rate (or probability of detection) and the false alarm rate. $\mathrm{TSS}(\mathbf{C}) \in [-1,1]$ and it is optimal when it is equal to 1. A negative value means that forecasting behaves in a wrong way, i.e., it mixes the role of the positive events with the role of the negative ones.

3) Heidke skill score (HSS) [24]

$$\mathrm{HSS}(\mathbf{C}) = \frac{2(\mathbf{C}_{1,1}\mathbf{C}_{2,2} - \mathbf{C}_{2,1}\mathbf{C}_{1,2})}{\mathbf{T}_1 + \mathbf{T}_2} \tag{3}$$

where $\mathbf{T}_1 := (\mathbf{C}_{1,1}+\mathbf{C}_{2,1})(\mathbf{C}_{2,1}+\mathbf{C}_{2,2})$ and $\mathbf{T}_2 := (\mathbf{C}_{1,1}+\mathbf{C}_{1,2})(\mathbf{C}_{1,2}+\mathbf{C}_{2,2})$, i.e., a measure of the improvement of forecast over random forecast. $\mathrm{HSS}(\mathbf{C}) \in (-\infty, 1]$. The optimal value is equal to 1, a negative value meaning that forecast is worse than random forecast and the 0 value meaning that the forecast has the same skill of random forecast.

4) Critical Success Index (CSI) [8]

$$\mathrm{CSI}(\mathbf{C}) = \frac{\mathbf{C}_{1,1}}{\mathbf{C}_{1,1} + \mathbf{C}_{1,2} + \mathbf{C}_{2,1}} \tag{4}$$

i.e., the ratio between the number of correct event forecast and the number of events which occurred plus the number of false alarms. $\mathrm{CSI}(\mathbf{C}) \in [0,1]$ and the optimal value is equal to 1.

Given a binary-vector $\mathbf{y} \in \{0,1\}^n$ encoding the binary outcome of an empirical observation, we can define a function $F_{\mathbf{y}}$ such as

$$\begin{aligned} F_{\mathbf{y}} : \{0,1\}^n &\to \mathbb{M}_{2,2}(\mathbb{N}) \\ \mathbf{p} &\mapsto C \end{aligned} \tag{5}$$

which maps a binary prediction $\mathbf{p} \in \{0,1\}^n$ onto the confusion matrix obtained by comparing $\mathbf{y}$ and $\mathbf{p}$. The function $F_{\mathbf{y}}$ is clearly not injective. Fig. 1 illustrates an example in which, given a binary observation $\mathbf{y}$, four different binary predictions lead to the same confusion matrix. In the example, $\mathbf{y}$ has 64 components such that 14 components are equal to 1 and 50 components are equal to zero. The four predictions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \mathbf{p}^{(3)}$ and $\mathbf{p}^{(4)}$ in the four panels of the figure provide the same confusion matrix with entries TP $= 11$, FN $= 3$, FP $= 7$ and TN $= 43$. From a forecasting quality viewpoint, the four predictions are the same, since they lead to the same confusion matrix and, accordingly, to the same skill scores. However, from a forecasting value viewpoint, the four predictions are different. More specifically, prediction $\mathbf{p}^{(4)}$, for which the corresponding FPs closely anticipate the observed outcomes equal to 1 should be preferred, in value terms, than the other predictions that sound alarms after the occurrence of the events.

We now introduce a novel definition of confusion matrix, which is able to distinguish between these ambiguous configurations and therefore to locally restore the injectivity of $F_{\mathbf{y}}$.

III. Value-Weighted Confusion Matrix and Skill Scores

In order to define the value-weighted confusion matrix, we first introduce the error functions $\varepsilon_{1,2} : \{0,1\} \times \{0,1\} \to$

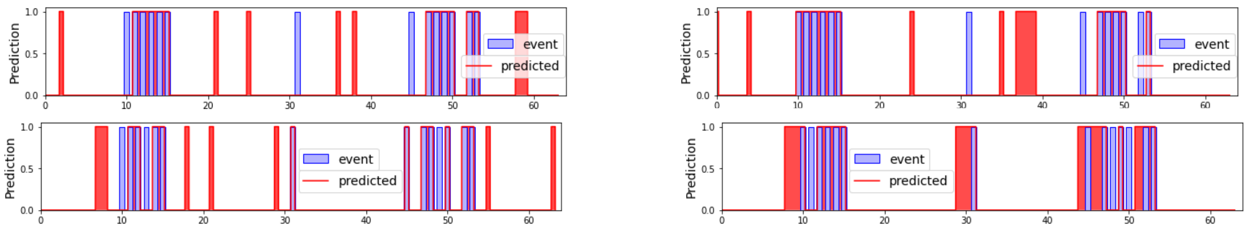


Fig. 1. Four different binary predictions with the same confusion matrix: $\mathbf{C}_{1,1} = 11$ (TPs), $\mathbf{C}_{1,2} = 7$ (FPs), $\mathbf{C}_{2,1} = 3$ (FNs) and $\mathbf{C}_{2,2} = 43$ (TNs).

$\mathbb{R}_+$ and $\varepsilon_{2,1} : \{0, 1\} \times \{0, 1\} \rightarrow \mathbb{R}_+$ such that, given the component y_i of a binary observed outcome $\mathbf{y}$ and the component p_i of a binary prediction $\mathbf{p}$, $\varepsilon_{1,2}(y_i, p_i)$ and $\varepsilon_{2,1}(y_i, p_i)$ measure the error of the incorrect prediction when either an FP or an FN occurs. These error functions allow the generalization of the confusion matrix concept as follows. We denote with $\mathbb{1}_{\{a\}}$ a number equal to 1 when condition a is satisfied and 0 otherwise. Then the weighted confusion matrix $\tilde{\mathbf{C}} \in \mathbb{M}_{2,2}(\mathbb{R}_+)$ is defined as

$$\tilde{\mathbf{C}}_{1,1} = \sum_{i=1}^{n} \mathbb{1}_{\{y_i=1, p_i=1\}}, \quad \tilde{\mathbf{C}}_{2,2} = \sum_{i=1}^{n} \mathbb{1}_{\{y_i=0, p_i=0\}} \tag{6}$$

$$\tilde{\mathbf{C}}_{1,2} = \sum_{i=1}^{n} \varepsilon_{1,2}(y_i, p_i), \quad \tilde{\mathbf{C}}_{2,1} = \sum_{i=1}^{n} \varepsilon_{2,1}(y_i, p_i). \tag{7}$$

On the one hand, quality-based forecasting assumes

$$\varepsilon_{1,2}(y_i, p_i) := \mathbb{1}_{\{y_i=0, p_i=1\}}, \ \ \varepsilon_{2,1}(y_i, p_i) := \mathbb{1}_{\{y_i=1, p_i=0\}}. \tag{8}$$

On the other hand, in the case of value-weighted forecasting we choose

$$\varepsilon_{1,2}(y_i, p_i) = \psi(y_i, p_i)\mathbb{1}_{\{y_i=0, p_i=1\}} \tag{9}$$

with

$$\psi(y_i, p_i) = \begin{cases} 1 - \max\limits_{1\le k\le K}\left(\dfrac{y_{i+k}}{k+1}\right), & \text{if } 1 \in \{y_{i+t}\}_{t=-K}^{K} \\ 2, & \text{otherwise} \end{cases} \tag{10}$$

and

$$\varepsilon_{2,1}(y_i, p_i) = \phi(y_i, p_i)\mathbb{1}_{\{y_i=1, p_i=0\}} \tag{11}$$

with

$$\phi(y_i, p_i) = \begin{cases} 1 - \max\limits_{1\le k\le K}\left(\dfrac{p_{i-k}}{k+1}\right), & \text{if } 1 \in \{p_{i+t}\}_{t=-K}^{K} \\ 2, & \text{otherwise.} \end{cases} \tag{12}$$

In (10) and (12), K is a fixed positive integer number. Analogously to the case of the standard confusion matrix in Section II, we have that $\tilde{\mathbf{C}}_{1,1}$, $\tilde{\mathbf{C}}_{2,2}$, $\tilde{\mathbf{C}}_{1,2}$ and $\tilde{\mathbf{C}}_{2,1}$ compute, respectively, the numbers of value-weighted true positives (wTPs), value-weighted true negatives (wTNs), value-weighted false positives (wFPs), and value-weighted false negatives (wFNs).

We remark that $\tilde{\mathbf{C}}_{1,1}, \tilde{\mathbf{C}}_{2,2} \in \mathbb{N}$ and that they have the same definition of the ones in the classical confusion matrix whereas $\tilde{\mathbf{C}}_{1,2}, \tilde{\mathbf{C}}_{2,1} \in \mathbb{R}_+$ and can be seen as weighted version of $\mathbf{C}_{1,2}$ and $\mathbf{C}_{2,1}$ with weighting functions $\psi(y_i, p_i)$ and $\phi(y_i, p_i)$, respectively. In order to illustrate how these weighting functions work while computing the corresponding prediction error, we introduce the window

$$\mathcal{I}_{i,K} = \{i - K, \ldots, i, \ldots, i + K\} \tag{13}$$

centered in the index i, with size K. Then, for the ith sample we have two possible cases as follows.

1) Case of a false positive ($y_i = 0$, $p_i = 1$): $\psi(y_i, p_i)$ depends on the sequence $\{y_j\}_{j\in\mathcal{I}_{i,K}}$. In fact:

a) If no event occurs in the window, i.e., $y_j = 0$ for each $j \in \mathcal{I}_{i,K}$ then

$$\psi(y_i, p_i) = 2. \tag{14}$$

b) If at least one event occurs in the window, i.e., $1 \in \{y_j\}_{j\in\mathcal{I}_{i,K}}$, then

$$\psi(y_i, p_i) = 1 - \max_{1\le k\le K}\left(\frac{1}{k+1} y_{i+k}\right). \tag{15}$$

Therefore, from (15) we can distinguish two situations as follows.

i) If the event occurs before time i and there is no event in the next times, i.e., $y_j = 0$ for $j \in \mathcal{I}_{i,K}$ and $j \ge i + 1$, then

$$\psi(y_i, p_i) = 1 \tag{16}$$

since in (15) $\max_{1\le k\le K}((1/(k+1))y_{i+k}) = 0$;

ii) If at least one event occurs after time i then

$$\frac{1}{2} \le \psi(y_i, p_i) < 1 \tag{17}$$

and the weighting function decreases according to the inverse of the distance of the next event occurrence, e.g., if $y_{i+1} = 1$ then $\psi(y_i, p_i) = (1/2)$.

2) Case of a false negative ($y_i = 1$, $p_i = 0$): $\phi(y_i, p_i)$ depends on the sequence $\{p_j\}_{j\in\mathcal{I}_{i,K}}$. In fact:

a) If no predicted alarm is in the window, i.e., $p_j = 0$ for each $j \in \mathcal{I}_{i,K}$ then

$$\phi(y_i, p_i) = 2. \tag{18}$$

b) If there is at least one predicted alarm in the window, i.e., $1 \in \{p_j\}_{j\in\mathcal{I}_{i,K}}$, then

$$\phi(y_i, p_i) = 1 - \max_{1\le k\le K}\left(\frac{1}{k+1} p_{i-k}\right). \tag{19}$$

In particular from (19) we can distinguish two situations as follows.

TABLE I

COMPARISON BETWEEN THE VALUE-WEIGHTED AND QUALITY-BASED CONFUSION MATRIX AND CORRESPONDING TSS FOR THE MOTIVATING EXAMPLES IN FIG. 1

Prediction	Value-weighted					
	wFP	wFN	wTSS	FP	FN	TSS
$\mathbf{p}^{(1)}$	14	4	0.4877	7	3	0.6457
$\mathbf{p}^{(2)}$	14	3.67	0.5044	7	3	0.6457
$\mathbf{p}^{(3)}$	8.08	1.67	0.7102	7	3	0.6457
$\mathbf{p}^{(4)}$	3.83	1.5	0.7981	7	3	0.6457

i) If the predicted alarm is after time i and there are no predicted alarms in the previous times, i.e., $p_j = 0$ for $j \in \mathcal{I}_{i,K}$ and $j \leq i-1$, then

$$\phi(y_i, p_i) = 1 \tag{20}$$

since $\max_{1\leq k\leq K}((1/k+1)p_{i-k}) = 0$.

ii) If there is at least one predicted alarm before time i, then

$$\frac{1}{2} \leq \phi(y_i, p_i) < 1 \tag{21}$$

and the weighting function decreases according to the inverse of the distance of the previous predicted alarm, e.g., if $p_{i-1} = 1$ then $\phi(y_i, p_i) = (1/2)$.

We applied the value-weighted confusion matrix $\tilde{\mathbf{C}}$ relying on this error function on the motivating example in Fig. 1. The results in Table I inspires the following comments. First, different from what happens with the classical quality-based confusion matrix $\mathbf{C}$, the off-diagonal terms of $\tilde{\mathbf{C}}$ depend on the prediction vector. Second, the new confusion matrix gives a clearer idea on how the incorrect predictions are distributed while rolling them along the sample index. On the one hand, in the value-weighted approach, wFPs associated with $\mathbf{p}^{(1)}$ notably increase with respect to the quality-based FPs, coherently to the fact that this prediction sounds alarms far from the actual event occurrence. On the other hand, wFPs associated with prediction $\mathbf{p}^{(4)}$ significantly decrease, coherently to the fact that, in this case, many incorrectly predicted alarms anticipate the event occurrence. Similar considerations can be repeated for what concerns the three original samples incorrectly predicted as 0: we notice that, in prediction $\mathbf{p}^{(4)}$, the number of wFNs is small, which means that the three missed events have been predicted in advance.

IV. VALUE-WEIGHTED SKILL SCORES IN ACTION

The aim of this section is to show that the example in Section II (see Fig. 1), in which forecasts with the same quality level (i.e., TSS) have completely different value (i.e., wTSS), is not only theoretical, but it occurs in practice when real datasets and conventional methods are used. Toward this aim we considered a dataset from the University of California at Irvine (UCI), released by the U.S. embassy in Beijing [25]. In detail, this archive includes the following.

TABLE II

RESULTS PROVIDED BY LR, SVM, AND NN IN THE CASE OF THE POLLUTION FORECASTING EXPERIMENT, WHEN THE TEST SET IS THE UCI DATABASE IN THE TIME RANGE FROM 1/7/2013 AT 00:00 THROUGH 1/15/2013 AT 07:00

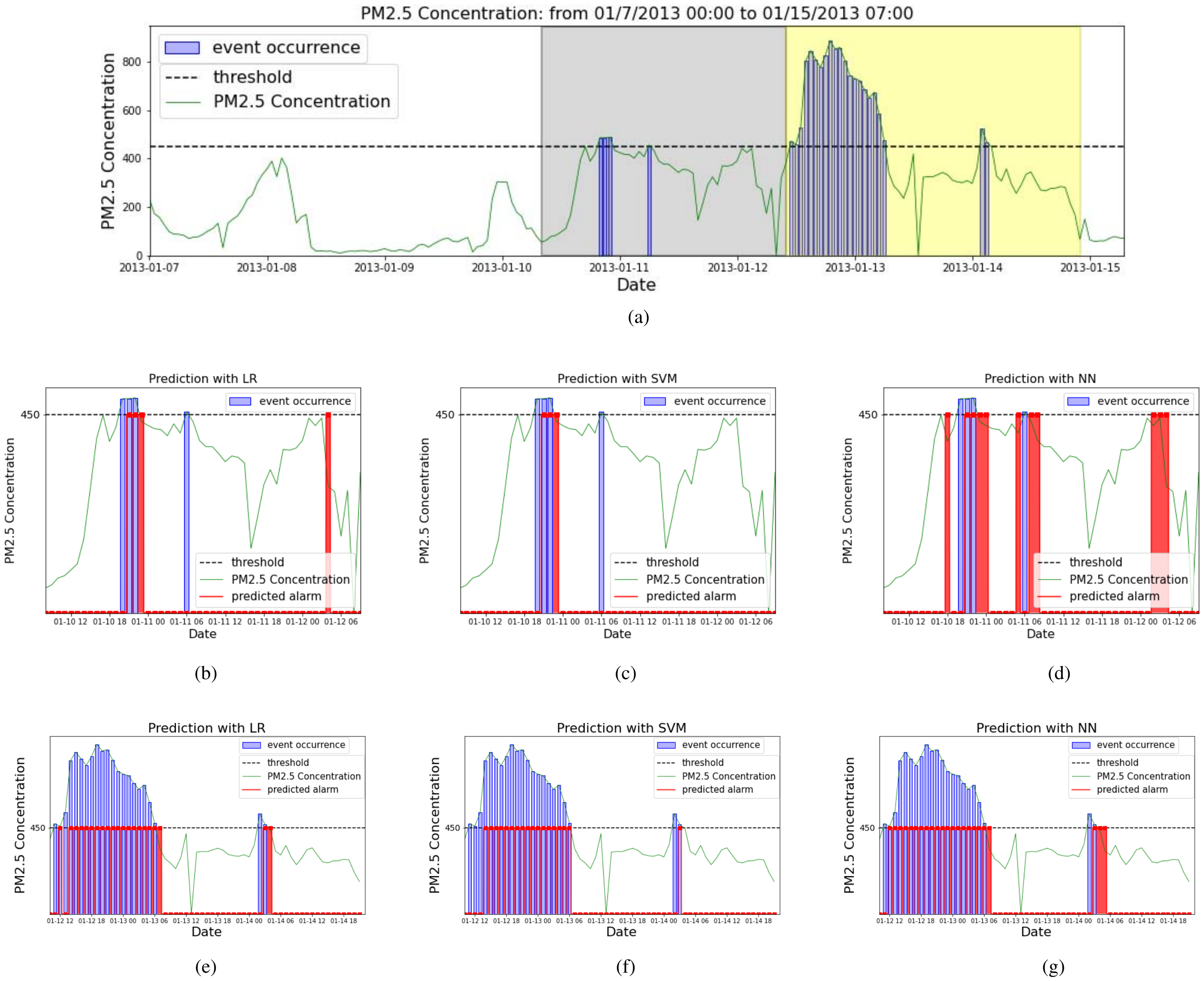
	Method		
	LR	SVM	NN
TP	21	20	22
FP	4	1	12
FN	5	6	4
TN	170	173	162
TSS	**0.7847**	0.7635	0.7772
wFN	5.5	7	3.17
wFP	5	1	14.17
wTSS	0.7639	0.7350	**0.7938**

1) Hourly weather information (dew point, temperature, pressure, wind direction and speed, cumulative number of snowing and raining hours).
2) Hourly concentration of PM2.5.

The forecasting problem we considered is the one to predict whether PM2.5 concentration at time $T+1$ will exceed a fixed threshold associated with a condition of severely polluted air, given weather conditions and PM2.5 concentration at time T. The machine learning methods used to address this forecasting problem were three standard supervised algorithms: logistic regression (LR) [26], support vector machine (SVM) [27], and a standard neural network (NN) [28]. We trained and validated the three algorithms using the dataset in the time range between 01/01/2010 and 12/28/2011 so that the training and validation sets had 17 424 samples with just 122 samples labeled with 1 (corresponding to over-threshold pollution). Fig. 2 shows the forecasting provided by the three algorithms in the case of a test set in the archive corresponding to the time range between 1/7/2013 at 00:00 UT and 1/15/2013 at 07:00 UT. Table II allows a quantitative assessment of these performances by means of both the standard quality-based confusion matrix and corresponding TSS, and the value-weighted confusion matrix and corresponding wTSS. We focused on TSS because we considered a significantly imbalanced training set and in this case this score is more appropriate [29]. A comparison between Fig. 2 and Table II shows how our value-weighted skill score works. Indeed, looking at the table for the following methods.

1) NN has significantly more FPs than the other two methods and slightly less FNs. As a consequence, its TSS is smaller than the one associated with LR (which produces a smaller number of FPs).
2) NN has the highest wTSS, as a consequence of the smallest number of wFNs among the three methods.

Coherently with these values, Fig. 2 shows that NN sounds either timely alarms or alarms in advance with respect to the actual event occurrence (this is particularly true in the case of the window highlighted by the gray box). Furthermore, it provides more FPs, but these are either sounded in a close neighborhood of the actual event occurrence or corresponds to a high PM2.5 concentration level.



(a)

(b) (c) (d)

(e) (f) (g)

Fig. 2. Top panel: PM2.5 concentration distributed on the period from 1/7/2013 at 00:00 through 1/15/2013 at 07:00. Middle panels (from left to right): with reference to the gray box in the top panel, predictions enrolled along time from 1/10/2013 at 08:00 through 1/12/2013 at 09:00 provided by LR, SVM, and NN, respectively. Bottom panels (from left to right): with reference to the yellow box in the top panel, predictions enrolled along time from 1/12/2013 at 10:00 through 1/14/2013 at 22:00 provided by LR, SVM, and NN, respectively. (a) PM2.5 concentration on period from 1/7/2013 00:00 to 1/15/2013 07:00. (b) LR: prediction in gray box. (c) SVM: prediction in gray box. (d) NN: prediction in gray box. (e) LR: prediction in yellow box. (f) SVM: prediction in yellow box. (f) NN: prediction in yellow box.

V. Deep Ensemble Classifiers and Applications

In deep learning, automatic classifiers can be constructed by applying thresholding procedures to NNs with probability outcomes. In this approach, a NN can be formally interpreted as the map

$$\theta(\mathbb{V},\cdot):\mathbb{R}^p \to [0,1] \tag{22}$$

where $\mathbb{V}$ represents the space of weights and the ensemble learning process implements the following steps.

1) Train $\theta(\mathbb{V},\cdot)$ on the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ using an iterative optimization scheme that stops after N epochs.
2) For each epoch j and given $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$, choose the classification threshold as the real number that maximizes a specific skill score S. Therefore, if

$$\hat{\mathbf{y}}_{\mathbf{X}}^{\mathbf{w}_j} := (\theta(\mathbf{w}_j,\mathbf{x}_1),\ldots,\theta(\mathbf{w}_j,\mathbf{x}_n))^T \tag{23}$$

then the optimum threshold is the solution of the optimization problem

$$\tau_j^* = \arg\max_{\tau\in[a,b]} S\left(F_{\mathbf{y}}\left(I_\tau\left(\hat{\mathbf{y}}_{\mathbf{X}}^{\mathbf{w}_j}\right)\right)\right) \tag{24}$$

where $[a,b]$ is a suitable interval with $0 \le a < b \le 1$, $I_\tau : \mathbb{R}^n \to \{0,1\}^n$ is the indicator function $I_\tau(\hat{\mathbf{y}}) = (\mathbb{1}_{\{\hat{y}_i>\tau\}},\ldots,\mathbb{1}_{\{\hat{y}_n>\tau\}})^T$, $F_{\mathbf{y}}$, defined as in (5), maps the binary prediction to the associated confusion matrix computed with respect to the true label vector $\mathbf{y}$, and S is the skill score computed on the confusion matrix, as defined in Section II.
3) Consider a validation set $\{(\tilde{\mathbf{x}}_i, \tilde{y}_i)\}_{i=1}^m$, and the matrix $\tilde{\mathbf{X}}$ such that $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_m)$. For each epoch j, compute

$$\hat{\mathbf{y}}_{\tilde{\mathbf{X}}}^{\mathbf{w}_j} := (\theta(\mathbf{w}_j,\tilde{\mathbf{x}}_1),\ldots,\theta(\mathbf{w}_j,\tilde{\mathbf{x}}_m))^T. \tag{25}$$

4) Given a level α, select just the epochs for which the skill score S computed on the validation set is bigger than α.

This allows the selection of the set of epochs

$$\mathcal{J}^* := \left\{ j \in \{1, \dots, N\} : S\left(F_{\tilde{\mathbf{y}}}\left(I_{\tau_j^*}\left(\hat{\mathbf{y}}_{\tilde{\mathbf{X}}}^{\mathbf{w}_j}\right)\right)\right) > \alpha \right\}. \tag{26}$$

5) Define the result of the ensemble learning process as the binary value corresponding to the median value m among all binary predictions associated with $\mathcal{J}^*$, i.e., given a new sample $\mathbf{x}$ the output is defined as

$$\hat{p} = m(\{I_{\tau_j^*}(\theta(\mathbf{w}_j, \mathbf{x})) : j \in \mathcal{J}^*\}). \tag{27}$$

In the case where the number of zeros is equal to the number of ones, we assume $\hat{p} = 1$.

We now show how this process works in the case of four applications concerning pollution, space weather, stock market, and IoT datastream forecasting. In these four applications, we considered θ to be a NN trained over $N = 100$ epochs using the Adam optimizer [30] with default values of the learning rate (0.001) and mini-batch size equal to 72. For each application, we estimated the values of the NN parameters by an empirical trial-and-error optimization process on several experiments. Our focus will be on the assessment of results when a value-weighted skill score is used in (24) and (26). To reflect the practical application of the proposed ensemble strategy, we show that it generally promotes predictions with higher values of the wTSS with respect to the early stopping strategy which is commonly used in conventional forecasting methods based on NNs. The source code of the experiments is available at https://github.com/SabrinaGuastavino/Value-weighted-skill-scores-in-deep-learning.

A. Pollution Forecasting

We consider the same data and the same forecasting problem discussed in Section IV, i.e., the prediction of over-threshold occurrences of PM2.5 concentration at time $T + 1$, having at disposal measures of this concentration and of eight features associated with weather conditions at time samples from 0 through T. The ensemble learning procedure is applied on the same training and validation sets as in Section IV. The NN θ is now a long short term memory (LSTM) NN available in the Keras library [31], in which the sigmoid function and the binary cross-entropy are used as activation function and loss function, respectively.

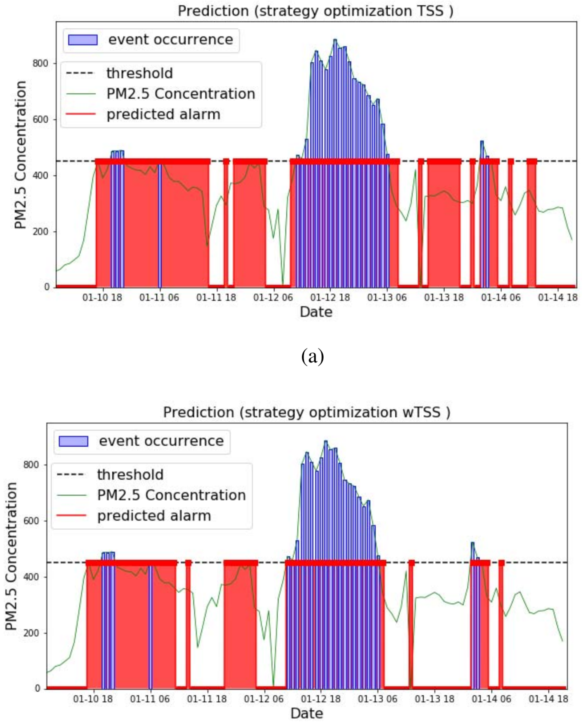
Table III shows the performances on the test set of the ensemble and early stopping strategies both optimized with respect to TSS and wTSS, respectively.

The prediction of the ensemble strategy is based on a shortlist of valuable predictions over the epochs depending on the parameter α [see (26)]. In this experiment, α is fixed equal to 0.9. We note that using the ensemble strategy optimized with respect to the TSS and the wTSS, respectively realizes different predictions. Fig. 3 offers a visualization enrolled over time of these predictions on the test period considered in Section IV. The strategy based on the wTSS optimization systematically leads to (sometimes even significantly) higher quality and value-weighted scores. For what concern the early stopping strategy, we trained the NN over 200 epochs and we stop when the validation loss does not improve for at least 10, 20, 30, 40, and 50 epochs. Then we select the stopping epoch which provides the best TSS (or wTSS) validation score. We note that it provides the same prediction regardless the optimized skill score. Moreover, this is the only application in which the early stopping strategy performs slightly better than the ensemble strategy based on the wTSS optimization.

TABLE III

PM2.5 POLLUTION FORECASTING. RESULTS ON THE TEST PERIOD FROM 12/29/2011 00:00 TO 12/31/2014 23:00. WE REPORT THE RESULT OF THE ENSEMBLE AND THE EARLY STOPPING STRATEGIES

	Ensemble		Early stopping
	TSS opt	wTSS opt	TSS/wTSS opt
TP	143	143	143
FP	785	400	396
FN	5	5	5
TN	25442	25827	25831
TSS	0.9363	0.9510	**0.9511**
HSS	0.2586	0.4087	0.4111
CSI	0.1533	0.2609	0.2629
wFN	4.5	4.5	4.5
wFP	1401.42	650.42	641.42
wTSS	0.9173	0.9449	**0.9453**
wHSS	0.1607	0.2974	0.3004
wCSI	0.0923	0.1792	0.1813



(a)

(b)

Fig. 3. PM2.5 pollution forecasting. Predictions provided by the ensemble method with the TSS optimization strategy (left panel) and the wTSS optimization strategy (right panel) on the period test from 1/10/2013 08:00 to 1/14/2013 21:00. (a) TSS optimization strategy. (b) wTSS optimization strategy.

B. Solar Flare Forecasting

Solar flares are the main trigger of space weather events, including coronal mass ejections and solar wind [32]. Their prediction may rely on features extracted from magnetogram images of active regions (ARs) like the ones recorded by the Helioseismic and Magnetic Imager (HMI) on-board the Solar Dynamics Observatory (SDO) [33]. We considered images from the HMI archives in order to set up an experiment in conditions analogous to the ones considered in [34] and [35].

1) We first grouped the images according to their issuing times and selected, in particular, just the images recorded at issuing time 00:00 UT.
2) 23 features were extracted by means of the algorithms illustrated in [36].
3) The labeling process utilized the flare occurrence alarm sounded by the Geostationary Operational Environment Satellites (GOES) cluster and associated label 1 to each feature vector where a flare was recorded within 24 h after the issuing time. GOES also computes the flare energetic class and, in particular, in the experiment we considered events of GOES class C1 and above (C1+ flares), M1 and above (M1+ flares) and X1 and above (X1+ flares).

The training process exploited the HMI archive in the time range from 09/15/2012 through 10/02/2015 and the validation process from 09/29/2015 through 01/11/2015 for the prediction of C1+ and M1+ solar flares. For the prediction of X1+ solar flares, we considered a different splitting in order to have a reasonable number of positive samples in training and validation: we trained on the period from 09/15/2012 through 10/09/2014 and we validated on the period from 10/10/2014 through 06/14/2015. For all three cases, the test phase focused on the time range from 01/13/2017 through 09/07/2017. We point out that, in this way, the validation set contains just two X1+ events, both associated with AR 12 673 [34], [37]. In order to implement the ensemble learning approach, we used a deep multilayer perceptron with 7 hidden layers. The rectified linear unit (ReLU) function was used to activate the hidden layers, the sigmoid activation function was applied to activate the output and the binary cross-entropy was used as loss function. The model was trained over 100 epochs using the Adam optimizer with learning rate equal to 0.001, with a mini-batch size of 72. In order to prevent overfitting, an L^2 regularization constraint was set as 0.01 in the first two layers. The quality level α in (26) was fixed equal to a percentage of the maximum value of the skill score in the validation step: this rate was set equal to 90% for prediction of C1+ flares, 95% for prediction of M1+ flares and 80% for prediction of X1+ flares. We compared results provided by the ensemble strategy with the early stopping strategy as in Section V-A.

The results of this analysis are shown in Table IV and imply once again that the ensemble strategy based on the maximization of wTSS leads to higher scores (and more diagonal confusion matrices) with respect to the results provided by the maximization of TSS. This is particularly significant in the case of the prediction of M1+ and X1+ flares, i.e., in the case when the training set is significantly more imbalanced. Furthermore, we note that maximizing the wTSS provides better predictions even when the early stopping strategy is used (only in the case of X1+ flare prediction the results are the same). However, the scores provided by the early stopping strategy are lower than the ones provided by the ensemble strategy (particularly lower in the case of M1+ flares prediction). Fig. 4 enrolls over time the prediction associated with AR 12 671 [38], [39], which originated many C1+ flares but no M1+ and no X1+ events. The figure clearly shows that the use of the ensemble strategy based on wTSS leads to a significantly smaller number of FPs in the prediction of both M1+ and X1+ events.

C. Stock Prize Forecasting

We considered the problem of predicting "down" movements in stock prizes relying on information concerned with the daily closure prizes. More specifically, the feature utilized as input of the forecasting algorithm is a time series of five days of daily percentage change defined as [40]

$$\eta = \frac{P_N - P_{N-1}}{P_{N-1}} \cdot 100 \tag{28}$$

where P_{N-1} is the closure prize at day $N-1$ and P_N is the closure prize at day N. We used as label for this feature the condition

$$\eta < L \tag{29}$$

where $L = -1$ corresponds to the "down" movement. We trained an LSTM NN on the training set in the time range from 10/01/2001 through 11/26/2007 in the database put at disposal by Yahoo Finance; the validation set is made of the same data, but in the time range from 11/27/2007 through 11/24/2009; the test set includes data from 11/25/2009 through 12/31/2010. Again we compared results provided by the ensemble strategy with the early stopping as in Section V-A. We report in Table V confusion matrices and skill scores corresponding to ensemble and early stopping strategies using both the quality- and value-weighted approaches. These numbers show that, when we choose the wTSS optimization strategy, the ensemble learning method leads to predictions with lower TSS but higher wTSS. Furthermore, the ensemble strategy leads to better scores than those obtained by applying the early stopping strategy. We further point out that, in stock index forecasting applications, the accuracy is often used for performance evaluation. Therefore, in Table V we also report both the quality- and value-weighted accuracy, although in our experiments the datasets are imbalanced, so that accuracy-type scores are less reliable than other skill scores like, for example, the TSS-type ones. Both accuracy and weighted value accuracy are slightly better for the strategy based on wTSS optimization.

In order to assess the value-weighted approach in an operational framework, we simulated the following investment strategy, starting from an initial asset of ten stocks.

1) If at day $N-1$ a "down" movement is predicted for day N, then we sell two stocks.
2) If either at day N, or day $N+1$, or day $N+2$ the "down" movement occurs, we use all the money earned at step 1 to buy stocks. At day $N+3$ we buy in any case.

This strategy is applied on the test set and the results of this analysis, illustrated in Fig. 5, show that, in a long-term perspective, the asset value provided by the value-weighted strategy overtakes the one provided by a standard quality-based strategy.

TABLE IV

SOLAR FLARE FORECASTING. RESULTS ON THE TEST SET PROVIDED BY THE ENSEMBLE STRATEGY OBTAINED BY REALIZING CLASSIFICATION VIA TSS OPTIMIZATION (SECOND, SIXTH, AND TENTH COLUMNS) AND VIA wTSS OPTIMIZATION (THIRD, SEVENTH, AND ELEVENTH COLUMNS) AND BY THE EARLY STOPPING STRATEGY BY REALIZING CLASSIFICATION VIA TSS OPTIMIZATION (FOURTH, EIGHTH, AND TWELFTH COLUMNS) AND VIA wTSS OPTIMIZATION (FIFTH, NINTH, AND TWELFTH COLUMNS). THE BEST TSS AND wTSS VALUES OBTAINED IN EACH CLASS FLARE PREDICTION ARE IN BOLD FACE

	Prediction C1+ flares				Prediction M1+ flares				Prediction X1+ flares		
	Ensemble		Early stopping		Ensemble		Early stopping		Ensemble		Early stopping
	TSS opt	wTSS opt	TSS opt	wTSS opt	TSS opt	wTSS opt	TSS opt	wTSS opt	TSS opt	wTSS opt	TSS/wTSS opt
TP	31	32	28	32	5	5	4	4	2	2	2
FP	30	32	29	35	21	19	24	22	31	20	23
FN	3	2	6	2	1	1	2	2	0	0	0
TN	198	196	199	193	235	237	232	234	229	240	237
TSS	0.7802	**0.8008**	0.6963	0.7877	0.7513	**0.7591**	0.5729	0.5807	0.8808	**0.9231**	0.9115
HSS	0.5832	0.5823	0.5407	0.5575	0.2859	0.3080	0.2053	0.2210	0.1013	0.1548	0.1359
CSI	0.4844	0.4848	0.4444	0.4638	0.1852	0.2	0.1333	0.1429	0.0606	0.0909	0.0800
wFN	2.5	1.5	6	1.67	1	1	1.75	1.75	0	0	0
wFP	29.92	33.92	31.75	42.42	36.25	32.25	44.75	40.75	60.5	40	44.5
wTSS	0.7941	**0.8077**	0.6859	0.7703	0.6997	**0.7136**	0.5340	0.5473	0.7910	**0.8571**	0.8419
wHSS	0.5886	0.5715	0.5185	0.5072	0.1807	0.2012	0.1145	0.1267	0.0494	0.0784	0.0699
wCSI	0.4888	0.4747	0.4259	0.4206	0.1183	0.1307	0.0792	0.0860	0.032	0.0476	0.0430

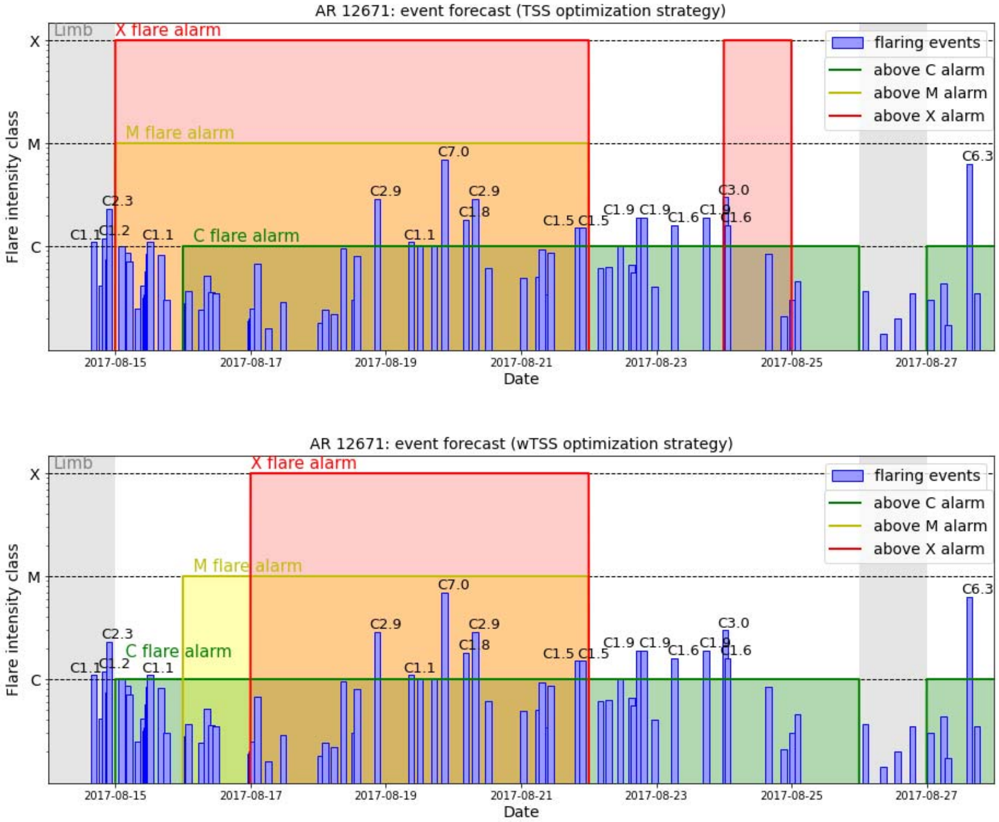


Fig. 4. Predictions enrolled along time by the ensemble method when the TSS and wTSS optimization strategies are adopted (top and bottom panels, respectively). Green alarms correspond to C1+ flares, yellow alarms to M1+ flares, and red alarms to X1+ flares. The blue bar plots correspond to the actual flaring events recorded by GOES, the y-label representing the corresponding GOES flare classes. Note that the gray boxes correspond to time period where the input data are missing.

D. IoT Data Stream Forecasting

We considered the problem of predicting the usage of light from IoT data stream. We analyze a public dataset available at https://www.kaggle.com/garystafford/environmentalsensor-data132k containing measurements of carbon monoxide, humidity, gas, smoke, temperature, taken by an environmental sensor collected in one-minute time series to predict the light in the next 30 seconds. The dataset ranges in the time period from 2020-07-12 00:01:34 to 2020-07-20 00:03:37 with a cadence of about 1.33 s. We split the dataset in such a way training, validation and test sets contain 1.78%, 1.71% and 1.73% of positive labeled samples, respectively. The forecast algorithm we implemented is a 1-D convolutional NN (1D CNN) followed by two stacked LSTM networks: the 1D CNN is designed with 32 kernels and a max-pooling

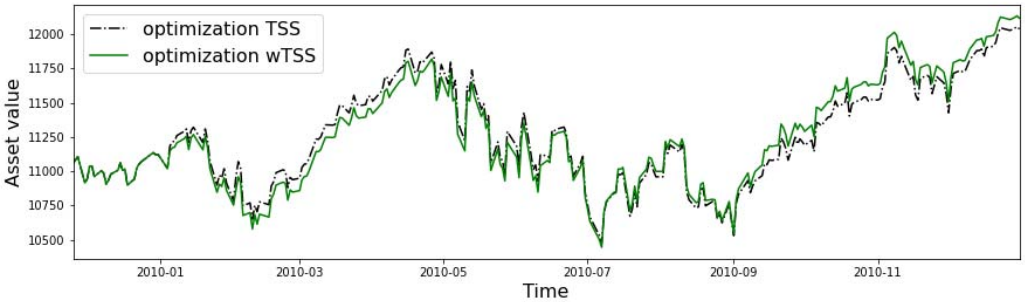


Fig. 5. Asset profile versus time, associated with the quality-based and value-weighted optimization strategies, starting from an initial asset made of ten stocks. Dashed black line: the impact of the TSS optimization strategy. Green solid line: the impact of the wTSS optimization strategy.

TABLE V

Stock Prize Forecasting. Results on the Test Set Extracted From the Yahoo Finance Database, Obtained by Using the Ensemble Strategy Based on the Optimization of TSS and wTSS (Second and Third Columns, Respectively) and the Early Stopping Strategy Based on the Optimization of TSS and wTSS (Third and Fourth Columns, Respectively)

	Ensemble		Early stopping	
	TSS opt	wTSS opt	TSS opt	wTSS opt
TP	22	20	27	25
FP	78	70	118	102
FN	19	21	14	16
TN	158	166	118	134
TSS	**0.2061**	0.1912	0.1585	0.1776
ACC	0.6498	0.6715	0.5235	0.5740
CSI	0.1849	0.1802	0.1698	0.1748
wFN	17	17.42	10.67	13.83
wFP	71.83	62.33	123.75	107.08
wTSS	0.2516	**0.2615**	0.2049	0.1996
wACC	0.67	0.7	0.5189	0.5680
wCSI	0.1985	0.2005	0.1673	0.1713

TABLE VI

IoT Data Stream Forecasting. Results on the Test Set Obtained by Using the Ensemble Strategy Based on the Optimization of TSS and wTSS (Second and Third Columns, Respectively) and the Early Stopping Strategy Based on the Optimization of TSS and wTSS (Third and Fourth Columns, Respectively)

	Ensemble		Early stopping	
	TSS opt	wTSS opt	TSS opt	wTSS opt
TP	72	72	72	72
FP	958	916	1006	943
FN	0	0	0	0
TN	3135	3177	3087	3150
TSS	0.7659	**0.7762**	0.7542	0.7696
HSS	0.1016	0.1071	0.0959	0.1035
CSI	0.0699	0.0729	0.0668	0.0709
wFN	0	0	0	0
wFP	1913	1829	2009	1883
wTSS	0.621	**0.6346**	0.6058	0.6259
wHSS	0.0441	0.0469	0.0411	0.0451
wCSI	0.0699	0.0379	0.0346	0.0368

layer to extract features from the 1D time series sequence data and the extracted features are then processed by the two LSTM layers with 32 and 16 hidden neurons. Finally, a dense layer of 16 neurons is applied. In LSTM layers a fraction of 0.5 neurons is randomly dropped during training to prevent overfitting [41]. We compared predictions provided by the ensemble strategy with the early stopping. In Table VI, we report the obtained results. As in the case of the prediction of X1+ flares, the learning method furnishes pretty good TSS (and wTSS) values and very low HSS values. This is due to the fact that the dataset is strongly unbalanced. Also, in this case, the ensemble strategy based on the wTSS optimization provides higher scores than all the other strategies.

VI. Comments and Conclusion

This study introduces two novelties for binary prediction problems over time. The first one is the definition of value-weighted skill scores that evaluate the forecasting performances in a way which is more appropriate for decision-making processes. According to this definition false positives anticipating the actual event occurrence are weighed less than the ones associated with alarms sounded behind schedule. Value-weighted skill score can be applied whenever we deal with a binary prediction over time and the decision-making process critically depends on the time a decision is taken. The second novelty is an ensemble strategy to provide a forecast optimized in terms of quality-based or value-weighted skill scores starting from the probabilistic predictions provided at each epoch by a deep learning algorithm. The ensemble strategy furnishes different predictions when optimized with respect to quality-based and value-weighted skill scores, respectively. Moreover, it generally yields forecasts with higher quality and value, when optimized with respect to the value-weighted skill score. The next step in this investigation will be the encoding of these value-weighted newly introduced skill scores in the loss functions utilized as part of the ensemble learning algorithm, in such a way to implement a forecasting approach in which the value-weighted strategy is *a priori* introduced in the optimization process.

References

[1] T. Gneiting and A. E. Raftery, “Weather forecasting with ensemble methods,” *Science*, vol. 310, no. 5746, pp. 248–249, 2005.

[2] E. Camporeale, S. Wing, and J. Johnson, *Machine Learning Techniques for Space Weather*. Amsterdam, The Netherlands: Elsevier, 2018.

[3] A. Kurt, B. Gulbagci, F. Karaca, and O. Alagha, “An online air pollution forecasting system using neural networks,” *Environ. Int.*, vol. 34, no. 5, pp. 592–598, Jul. 2008.

[4] R. S. Deepak, S. I. Uday, and D. Malathi, “Machine learning approach in stock market prediction,” *Int. J. Pure Appl. Math.*, vol. 115, no. 8, pp. 71–77, 2017.

[5] A. H. Murphy, “What is a good forecast? An essay on the nature of goodness in weather forecasting,” *Weather Forecasting*, vol. 8, no. 2, pp. 281–293, Jun. 1993.

[6] O. Allouche, A. Tsoar, and R. Kadmon, “Assessing the accuracy of species distribution models: Prevalence, Kappa and the true skill statistic (TSS),” *J. Appl. Ecol.*, vol. 43, no. 6, pp. 1223–1232, Sep. 2006.

[7] O. Hyvärinen, “A probabilistic derivation of Heidke skill score,” *Weather Forecasting*, vol. 29, no. 1, pp. 177–181, Feb. 2014.

[8] J. T. Schaefer, “The critical success index as an indicator of warning skill,” *Weather Forecasting*, vol. 5, no. 4, pp. 570–575, 1990.

[9] K. R. Mylne, “Decision-making from probability forecasts based on forecast value,” *Meteorolog. Appl.*, vol. 9, no. 3, pp. 307–315, Sep. 2002.

[10] D. Richardson, “Skill and relative economic value of the ECMWF ensemble prediction system,” *Quart. J. Roy. Meteorolog. Soc.*, vol. 126, no. 563, pp. 649–667, Jan. 2000. [Online]. Available: https://www.ecmwf.int/node/11903

[11] C. Elkan, “The foundations of cost-sensitive learning,” in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 17, no. 1. Mahwah, NJ, USA: Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[12] P. Pérez, A. Trier, and J. Reyes, “Prediction of $\mathrm{PM}_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile,” *Atmos. Environ.*, vol. 34, no. 8, pp. 1189–1196, Jan. 2000.

[13] M. Z. Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, “$\mathrm{PM}_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data,” *Atmosphere*, vol. 10, no. 7, p. 373, Jul. 2019.

[14] U. Pak *et al.*, “Deep learning-based $\mathrm{PM}_{2.5}$ prediction considering the spatiotemporal correlations: A case study of Beijing, China,” *Sci. Total Environ.*, vol. 699, Jan. 2020, Art. no. 133561.

[15] M. G. Bobra and S. Couvidat, “Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm,” *Astrophys. J.*, vol. 798, no. 2, p. 135, Jan. 2015.

[16] F. Benvenuto, M. Piana, C. Campi, and A. M. Massone, “A hybrid supervised/unsupervised machine learning approach to solar flare prediction,” *Astrophys. J.*, vol. 853, no. 1, p. 90, Jan. 2018.

[17] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, and M. Ishii, “Deep flare net (DeFN) model for solar flare prediction,” *Astrophys. J.*, vol. 858, no. 2, p. 113, May 2018.

[18] O. Bustos and A. Pomares-Quimbaya, “Stock market movement forecast: A systematic review,” *Expert Syst. Appl.*, vol. 156, Oct. 2020, Art. no. 113464.

[19] S. Liu, X. Zhang, Y. Wang, and G. Feng, “Recurrent convolutional neural kernel model for stock price movement prediction,” *PLoS ONE*, vol. 15, no. 6, Jun. 2020, Art. no. e0234206.

[20] Y. Shi, Y. Zheng, K. Guo, and X. Ren, “Stock movement prediction with sentiment analysis based on deep learning networks,” *Concurrency Comput., Pract. Exper.*, vol. 33, no. 6, p. e6076, Mar. 2021.

[21] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, “Deep learning for IoT big data and streaming analytics: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2923–2960, 4th Quart., 2018.

[22] M. Sokolova, N. Japkowicz, and S. Szpakowicz, “Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation,” in *Proc. Australas. Joint Conf. Artif. Intell.* Berlin, Germany: Springer, 2006, pp. 1015–1021.

[23] A. Hanssen and W. Kuipers, *On the Relationship Between the Frequency of Rain and Various Meteorological Parameters: With Reference to the Problem OB Objective Forecasting*. Koninkl, The Netherlands: Meterologisch Institut. Mededelingen en Verhandelingen. Staatsdrukkerij-en Uitgeverijbedrijf, 1965. [Online]. Available: https://books.google.it/books?id=nTZ8OgAACAAJ

[24] P. Heidke, “Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst,” *Geografiska Annaler*, vol. 8, pp. 301–349, Aug. 1926.

[25] X. Liang *et al.*, “Assessing Beijing’s $\mathrm{PM}_{2.5}$ pollution: Severity, weather impact, APEC and winter heating,” *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 471, no. 2182, Oct. 2015, Art. no. 20150257.

[26] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.

[27] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[28] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford Univ. Press, 1995.

[29] D. S. Bloomfield, P. A. Higgins, R. T. J. McAteer, and P. T. Gallagher, “Toward reliable benchmarking of solar flare forecasting methods,” *Astrophys. J.*, vol. 747, no. 2, p. L41, Mar. 2012.

[30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[31] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[32] E. Tandberg-Hanssen and A. G. Emslie, *The Physics of Solar Flares*, vol. 14. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[33] P. H. Scherrer *et al.*, “The helioseismic and magnetic imager (HMI) investigation for the solar dynamics observatory (SDO),” *Sol. Phys.*, vol. 275, nos. 1–2, pp. 207–227, Jan. 2012.

[34] F. Benvenuto, C. Campi, A. M. Massone, and M. Piana, “Machine learning as a flaring storm warning machine: Was a warning machine for the 2017 September solar flaring storm possible?” *Astrophys. J.*, vol. 904, no. 1, p. L7, Nov. 2020.

[35] C. Campi, F. Benvenuto, A. M. Massone, D. S. Bloomfield, M. K. Georgoulis, and M. Piana, “Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence,” *Astrophys. J.*, vol. 883, no. 2, p. 150, Sep. 2019.

[36] J. A. Guerra, S.-H. Park, P. T. Gallagher, I. Kontogiannis, M. K. Georgoulis, and D. S. Bloomfield, “Active region photospheric magnetic properties derived from line-of-sight and radial fields,” *Sol. Phys.*, vol. 293, no. 1, pp. 1–25, Jan. 2018.

[37] S. Guastavino, M. Piana, A. Massone, R. Schwartz, and F. Benvenuto, “Desaturating SDO/AIA observations of solar flaring storms,” *Astrophy. J.*, vol. 882, p. 109, Sep. 2019.

[38] B. Boe, S. Habbal, M. Druckmüller, A. Ding, J. Hodérova, and P. Štarha, “CME-induced thermodynamic changes in the corona as inferred from Fe XI and fe XIV emission observations during the 2017 August 21 total solar eclipse,” *Astrophys. J.*, vol. 888, no. 2, p. 100, Jan. 2020.

[39] J. Duncan *et al.*, “NuSTAR observation of energy release in 11 solar microflares,” *Astrophys. J.*, vol. 908, no. 1, p. 29, Feb. 2021.

[40] C. Chatfield and H. Xing, *The Analysis of Time Series: An Introduction With R*. Boca Raton, FL, USA: CRC Press, 2019.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdi, “DropOut: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Nov. 2014.

Sabrina Guastavino received the B.Sc. and M.Sc. degrees *(cum laude)* in mathematics and the Ph.D. degree in mathematics and applications from University of Genoa, Genoa, Italy, in 2014, 2016, and 2020, respectively.

She is currently an Assistant Professor in numerical analysis with the University of Genoa. Her research interests include statistical learning, inverse problems, and regularization theory. Her research is primarily concerned with machine/deep learning, statistical, and variational methods for inverse problems with particular focus on solar physics applications.

Michele Piana received the Laurea *(cum laude)* and Ph.D. degrees in physics from the Università di Genova, Genoa, Italy, in 1992 and 1996, respectively.

From 1997 to 1998, he held a post-doctoral position at the Department of Mathematical Sciences, University of Delaware, Newark, DE, USA. From 1999 to 2001, he was a Researcher with the Istituto Nazionale di Fisica della Materia, Genoa; and from 2001 to 2005, he was a Research Associate of numerical analysis with the University of Genoa; and then an Associate Professor of computer science at Verona University, Verona, Italy. He is currently a Full Professor of numerical analysis with the Mathematical Department, Genoa University. His research interests include regularization theory for inverse problems, inverse scattering, medical imaging, computational neuroscience, and computational astrophysics.

Federico Benvenuto received the Ph.D. degree in mathematics and applications from the University of Genoa, Genoa, Italy, in 2010, jointly with the University of Nice-Sophia Antipolis, Nice, France.

He is currently an Associate Professor in numerical analysis with the University of Genoa. Before joining Genoa University, he was a Marie Curie Post-Doctoral Research Associate with the Centre de Mathematique Appliqueé, École Polytechnique in Paris, Palaiseau, France. His research is primarily concerned with inverse problems and their applications, with particular emphasis on regularization theory. His interest focuses on statistical and variational approaches to discrete inverse problems, maximum likelihood and maximum a posteriori estimation, and iterative reconstruction methods.