Predicting Corporate Failure Using Ensemble Extreme Learning Machine

David Veganzones

Abstract Corporate failure prediction has become a major topic in the accounting and finance literature. Effective prediction models are essential for banks and financial institutions to solve financial decision-making problems. In general, artificial intelligence and machine learning techniques have been mainly employed to develop corporate failure models due to their prediction superiority in comparison to the traditional statistical method. Extreme learning machine is a newly developed artificial intelligence technique with an extremely fast learning speed. Nonetheless, its performance instability may be a major constraint for its practical application. The literature documents that the ensemble is one of the widely used methods to improve the generalization performance of weak classifiers. Therefore, we propose in this study an ensemble of extreme learning machine for improving the prediction performance on corporate failure task. In particular, we compare four benchmark ensemble methods (multiple classifiers, bagging, boosting, and random subspace) to evaluate which is best suited for extreme learning machine. Experimental results on French firms indicated that bagged and boosted extreme learning machine showed the best-improved performance.

Keywords Forecasting · Corporate failure · Machine learning · Extreme learning machine · Ensemble

1 Introduction

The global economic developments of recent decades have put corporate failure and their consequences for economic well-being under the spotlight, to the extent that bankruptcy or business failure has become a crucial task in finance. This, in turn, has emphasized that financial institutions need effective prediction mechanisms in order to make an appropriate lending decision.

D. Veganzones (✉)
ESCE International Business School, OMNES Education, Paris La Défense, France

M. Z. Abedin, P. Hajek (eds.), *Novel Financial Applications of Machine Learning and Deep Learning*, International Series in Operations Research & Management Science 336, https://doi.org/10.1007/978-3-031-18552-6_7

In general, the objective of corporate failure prediction is to forecast the likelihood that a firm will survive or fail with the minimum possible classification error. That is why corporate failure research aims at binary classification (Séverin & Veganzones, 2021; Ouenniche & Tone, 2017). From the binary classification point of view, the model's output is a dichotomous variable that takes the value of 1 when the firm follows a bankruptcy procedure and is set to 0 when the firm survives. The explanatory variables to design corporate failure prediction models are often financial ratios, which measure the relationship between any two items on financial statements.

Since the pioneer studies of Beaver (1966) and Altman (1968) who documented the predictive power of ratio analysis, many prediction techniques have been employed to develop corporate failure prediction models, including statistical and artificial intelligence methods (Veganzones & Severin, 2020; Kumar & Ravi, 2007; Moula et al., 2017). On the one hand, researchers still employ well-known statistical methods, notably linear discriminant analysis and logistic regression, due to their simplicity and capacity to interpret the data, even though they are clearly outperformed by machine learning techniques. On the other hand, artificial intelligence techniques (i.e., support vector machine, decision trees, neural networks, fuzzy set theory, self-organizing map) have become indispensable tools in the field of corporate failure prediction, especially in this era of advanced informatics and computing technology (Abedin et al., 2021). Their superiority relies on the fact that they learn directly from the data, which makes it possible to test complex data using nonlinear approaches, and therefore, their predictions are more reliable. Nonetheless, these mentioned methods are not free of drawbacks: low learning rate, slow computational time, converge in local minima, etc. (Yu et al., 2014; Abedin et al., 2018), which could make corporate failure prediction time consuming and arduous.

To overcome these, we consider a novel prediction method, Extreme Learning Machine (ELM) (Huang et al., 2006a) to predict corporate failure. There are several reasons behind choosing ELM as the classifier for the prediction of corporate failures. Firstly, despite many existing methodologies for predicting corporate failure, new methods of research should be continually explored by researchers and practitioners. Secondly, the main concept behind ELM is the random initialization of the Single Layer Feed-Forward Neural Network (SLFN), which replaces the computationally cost procedure of training the hidden layer performed by other artificial intelligence techniques. Unlike the AI techniques, it does not need to calibrate parameters, such as the learning rate. For this reason, ELM has good performance with an extremely fast learning speed (Akusok et al., 2015) and it is proven to be a universal approximator given enough hidden neurons (Huang et al., 2006b).

However, as other techniques, ELM possesses a main drawback: the random initialization that allows ELM to be an extremely fast algorithm, it becomes ELM a highly unstable classifier as well. In ELM, even if we train the same training sample several times, it performs differently due to the random initialization of bias and weights between the input and hidden nodes. Although the reliance on a single ELM may be misguided, the ensemble of predictions might improve the generalization performance of the ELM. Indeed, ensemble methods are usually used as an

instrument for improving the accuracy of the learning algorithm by constructing and combining a set of weak classifiers (Kim & Kang, 2010; Abedin et al., 2022). This rationale motivates our specific study of the performance of the ensemble extreme learning machine to predict corporate failure.

Consequently, the aim of this current work is to fully examine which is the best ensemble procedure to improve the performance of ELM for corporate failure prediction. This is of significant importance because the diversity generation method is key in the process of creating an ensemble of classifiers. According to Rokach (2010), diversity creation can be obtained in several ways: by manipulating the training sample, by manipulating the inducer, by varying the representation of the target attribute and by changing the search space. Of all possible ensemble techniques, we selected 4 based on their popularity in the literature (Verikas et al., 2010): Multiple classifiers, Bagging, Boosting, and Random Subspace. The fact that the chosen techniques rely on different ensemble procedures might provide further insight into the general characteristics of ensemble techniques that are influenced by the base classifier. In turn, a rigorous study of such methods would provide assistance in designing a model of corporate failure based on ensemble ELM. Furthermore, optimal performance of prediction models developed based on ensemble ELM models can be employed as a baseline prediction model for future research.

The rest of the paper is organized as follows. Section 2 presents the research methodology. Sections 3 and 4 describe the experimental design and results, respectively. Finally, in Sect. 5, the conclusions are summarized.

2 Research Methodology

In this section, we present the method employed in this study. In particular, we describe the extreme learning machine classifier as well as the ensemble modeling techniques.

2.1 Extreme Learning Machine

The Extreme Learning Machine (ELM) classifier was proposed by Huang et al. (2006a). The ELM represents a fast way of creating a Single Layer Hidden Feed-Forward Neural Network (SLFN) by the random initialization of the internal bias and weights. The hidden layer does not need to be iteratively tuned; it bypasses the time-consuming calibration setup performed by artificial intelligence algorithms. As a result, ELM is an extremely fast learning speed while being a simple method. The ELM algorithm can be described as follows:

Consider a set of N observations with features $\boldsymbol{x}_i \in \mathbb{R}^N$ and the corresponding output labels $\boldsymbol{Y} \in \{-1, 1\}^{Nxc}$. A SLFN with m neurons in the hidden layer is written by the following sum:

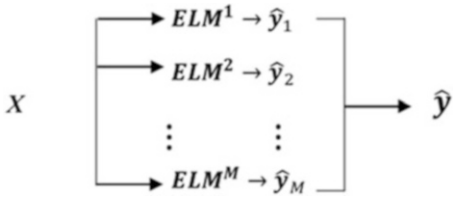

Fig. 1 Architecture of the multiple classifier

$$\Sigma_{j=1}^{m} \beta_j \, \phi\left(w_j \, \boldsymbol{x}_i + b_j\right) = \boldsymbol{Y}_{ik}, i = 1, \ldots, N \; k = 1, \ldots, c, \tag{1}$$

where β_j are the output weights, ϕ is the activation function, w_j are the input weights and b_j represents the biases. The Eq. (1) can be expressed in the form of a matrix as $\boldsymbol{H\beta} = \boldsymbol{Y}$, where

$$\boldsymbol{H} = \begin{pmatrix} \phi(w_1 \, x_1 + b_1) & \cdots & \phi(w_m \, x_1 + b_m) \\ \vdots & \ddots & \vdots \\ \phi(w_1 \, x_N + b_1) & \cdots & \phi(w_m \, x_N + b_m) \end{pmatrix}. \tag{2}$$

$$\boldsymbol{\beta} = (\beta_1 \;\; \cdots \;\; \beta_m)^c \quad \boldsymbol{Y} = (\, Y_1 \;\; \cdots \;\; Y_N \,)^c.$$

Then, the output weights $\boldsymbol{\beta}$ can be calculated by the Ordinary Least Squares method using the Moore-Penrose pseudo inverse of $\boldsymbol{H}$ (Rao & Mitra, 1971):

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{Y}. \tag{3}$$

2.2 Ensemble Techniques

2.2.1 Multiple Classifiers Technique

The multiple classifier technique relies on the simple idea that the combination of multiple classifiers leads to higher classification prediction and efficiency than the single classifier. This approach is equivalent to the wisdom of crowds: the combined opinion of diverse and independent experts usually outperforms the opinion of single individuals. According to Kitter et al. (1998), the multiple classifier technique achieves higher efficiency when learners generalize in different ways, i.e., the diversity of the ensemble is generated. As ELM is based on the random initialization of internal bias and weights, each learner will be different; there is diversity in the ensemble. Therefore, the forecast of several ELMs will be combined using majority voting to produce the final decision rule. Figure 1 shows the general architecture of the multiple classifier.

The classifiers $\boldsymbol{C^1(X)}, \ldots, \boldsymbol{C^M(X)}$ are built based on the data set $\{(\boldsymbol{x}_1, \boldsymbol{y_1}), (\boldsymbol{x}_2, \boldsymbol{y_2}), \ldots, (\boldsymbol{x}_n, \boldsymbol{y_n})\}$. Each classifier provides an output $\widehat{\boldsymbol{y}}_M$ that will be combined into the final output $\widehat{\boldsymbol{y}}$.

2.2.2 Bagging

Bagging (short for bootstrap aggregating) is one of the primal ensemble techniques (Breiman, 1996). Its popularity lies in the fact that it is intuitive and simple to implement, with notably good performance. Bagging generates the essential diversity to create the ensemble process that manipulates the training set. In this regard, the training set samples are randomly resampled in order to generate several different bags of samples. Thus, each bag represents a set of training samples. Finally, the base classifier is applied to each bag, and the output classification is made by a majority vote of all the base classifier results.

Bagging technique generates an improvement in generalization performance due to the reduction in variance while maintaining steady or only slightly increasing the bias, in particular, when it is applied to weak classifiers (Grandvalet, 2004). The bagging algorithm can be expressed as follows:

Given a data set $\{(\boldsymbol{x}_1, \boldsymbol{y_1}), (\boldsymbol{x}_2, \boldsymbol{y_2}), \ldots, (\boldsymbol{x}_n, \boldsymbol{y_n})\}$.

1. Repeat for $i = 1, 2, \ldots, I$.
   (a) Build a bootstrap sample $\left\{\left(\boldsymbol{x}_1^*, \boldsymbol{y}_1^*\right), \left(\boldsymbol{x}_2^*, \boldsymbol{y}_2^*\right), \ldots, \left(\boldsymbol{x}_n^*, \boldsymbol{y}_n^*\right)\right\}$ by randomly selecting n times with replacement from the data $\{(\boldsymbol{x}_1, \boldsymbol{y_1}), (\boldsymbol{x}_2, \boldsymbol{y_2}), \ldots, (\boldsymbol{x}_n, \boldsymbol{y_n})\}$.
   (b) Fitting the bootstrapped classifier $\boldsymbol{C}_i$ on corresponding bootstrap sample.
2. Calculate the output of the final classifier:

$$\boldsymbol{C}(\boldsymbol{x}) = I^{-1} \sum_i^I C_i(x). \tag{4}$$

2.2.3 Boosting

Unlike the bagging technique, the boosting technique combines inaccurate and relatively weak rules to produce highly accurate predictions. That is, it progressively gives more weight to observations that have been misclassified by previously generated classifiers in order to generate new classifiers and then combines the classifiers of different iterations with weighted voting to make final predictions. Since numerous algorithms for boosting have been proposed, we use the Adaboost algorithm (Freund & Schapire, 1996) which is one of the most popular boosting techniques applied to pattern recognition (Verikas et al., 2010). The Adaboost algorithm can be described as follows:

Given a data set $\{(\boldsymbol{x}_1, \boldsymbol{y_1}), (\boldsymbol{x}_2, \boldsymbol{y_2}), \ldots, (\boldsymbol{x}_n, \boldsymbol{y_n})\}$.

1. Initialize the weight vector of the training set:

$$W_1(i) = 1/N \text{ for } i = 1, \ldots, N. \tag{5}$$

2. For $t = 1, \ldots, T$,

   (a) Train the weak classifier C_t on the weighted training samples.
   (b) Calculate the sum of weighted errors of C_t:

$$\varepsilon_t = \sum_{i=1}^{N} W_i^t, Y_i \neq C_t\ (X_i). \tag{6}$$

(c) Choose

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right). \tag{7}$$

(d) Update the weights:

$$W_i^{t+1} = \frac{W_i^t \exp\left(-\alpha_t Y_i C_t\ (X_i)\right)}{Z_t}, \tag{8}$$

where Z_t is a normalization factor.

3. Output:

$$f(x) = sign\left(\sum_{t=1}^{T} \alpha_t C_t\ (x)\right). \tag{9}$$

2.2.4 Random Subspace

The random subspace (Ho, 1998) bases its ensemble process on the modification of the feature space. That is, it creates different bags of training samples by randomly selecting features drawn for the initial feature set that characterizes each sample. The training sample $\boldsymbol{X}_i (i = 1, \ldots, n)$ in the training set $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ is a p-dimensional vector $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ip})$, where p represents the feature components. Within the random subspace, the k-dimensional subspace is randomly selected from the original p-dimensional feature space, $k < p$. The new learning samples $\boldsymbol{X}^b = \left(\boldsymbol{X}_1^b, \boldsymbol{X}_2^b, \ldots, \boldsymbol{X}_n^b\right)$ in a k-dimensional subspace $\boldsymbol{X}_i^b = \left(\boldsymbol{x}_{i1}^b, \boldsymbol{x}_{i2}^b, \ldots, \boldsymbol{x}_{in}^b\right)$,

where $\boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{b}}(j = 1, \ldots, r)$, are built and then, the classifiers in the random subspace $\boldsymbol{X}^{b}$ are combined using majority voting to create the final decision rule. Thus, the random subspace can be organized as follows:

1. Repeat b times, with $b = 1, 2, \ldots, B$
   (a) Randomly select a k-dimensional subspace $\boldsymbol{X}^{b}$ among the initial p-dimensional feature space $\boldsymbol{X}$.
   (b) Design a classifier $\boldsymbol{C}^{\boldsymbol{b}}(\boldsymbol{x})$ using the sample $\boldsymbol{X}^{b}$.
2. Combine the forecast of $\boldsymbol{C}^{\boldsymbol{b}}(\boldsymbol{x})$ classifiers using majority voting to a final decision rule.

$$\mathrm{Prev}(x) = \underset{y \in \{-1;1\}}{\operatorname{argmax}} \sum_{b=1}^{B} \delta_{sgn\left(\boldsymbol{C}^{\boldsymbol{b}}(\boldsymbol{x})\right),\boldsymbol{y}}. \tag{10}$$

3 Experimental Design

3.1 Data

Our empirical study uses non-listed French firms taken from the Diane database created by Bureau Van Dijk. The French companies must submit annual reports to the French Commercial Court under French law provide accounting and income statements to the Bureau Van Dijk authority. We drew firms from all sectors of activity (excluding financial companies) for the years 2016–2018, allowing us to examine the model's capacity to create good prediction rules in a real-world scenario.

The Diane database provides the information on whether firms have failed or remain healthy; in the case of failure, it also provides the date. A firm is considered to be failed if it proceeded to be liquidated or reorganized, and non-failed firms were those that continued their activity for at least a year after the period studied. We decided to be conservative in the selection of non-failed firm in order to avoid the inclusion of healthy companies that may suddenly fail and ensure a reliable sample that does not fail. Moreover, firms that presented missing values in their financial statement, as well as outliers, were excluded to ensure the prediction model stability. Consequently, the collected dataset is composed of 3000 failed and 3000 non-failed firms.[1]

[1] Corporate failure is a rare phenomenon in the real world, so failed firms are clearly outnumbered by non-failed ones. That is why the sample selection process becomes a significant paradigm. If one design a model based on the actual population, the dataset must be imbalanced. However, this procedure has a main drawback: it is likely to lead to significant degradation of the prediction performance due to low percentage of failed firm in the entire sample (López et al., 2013; Shajalal et al., 2021). Therefore, we collect a stratified sample with same observations of failed and non-failed based on matched pair technique (Ciampi, 2015), in which failed firms are matched with non-failed firms according to industry sector, size, and firm age.

To minimize the bias effect and sample variability that might influence the model prediction performance, we carried out a tenfold cross-validation method in which the dataset is split into ten distinct training and test set in order to learn and evaluate the model prediction. This procedure was repeated ten times to ensure the reliability of our results. Therefore, the final prediction performance is calculated as the average of 100 testing results.

3.2 Variables

Financial dimensions characterize the main explanatory factors for corporate failure. Therefore, the balance sheets and income statements of the collected firms were used to calculate 30 financial ratios to use as explanatory variables. This representation layer is important because it guarantees that the variables, we have used actually represent all aspects of the phenomenon.

The initial set of financial ratios that we compute includes at least four indicators representing six categories: liquidity, solvency, profitability, financial structure, turnover, and activity. These variables are presented in Table 1.

However, using all financial ratios may result in very high-dimensional feature space, which may reduce model predictive capability. Therefore, a variable selection process has been performed in order to choose a subset of the most relevant financial ratios. Following the study by Kainulainen et al. (2011), a feed-forward variable selection process was performed to retain the necessary information for prediction.

3.3 Evaluation Metrics

The evaluation criteria of our experiments are adopted from standard measures established in the field of prediction (Shahriare et al., 2021). These measures include average accuracy, type error I, and type error II. The formula of these measures provided below can be explained with respect to the confusion matrix shown in Table 2.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{11}$$

$$\text{Type} - \text{I error} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{12}$$

$$\text{Type} - \text{II error} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{13}$$

In addition to these evaluation metrics, we also used the area under the receiver operating characteristic curve (AUC) to estimate the model performance. This is a

Table 1 Initial set of variables

Profitability		Liquidity	
X1	Profit before Tax/Shareholders' Funds	X16	Cash/total assets
X2	Net income/shareholders' funds	X17	Current assets/current liabilities
X3	EBITDA/Total assets	X18	Current assets/total debts
X4	EBIT/Total assets	X19	Quick assets/Total assets
X5	Net income/Total assets	X20	(Cash +Marketable securities)/Total sales
Financial structure		Turnover	
X6	Shareholder's funds/Total assets	X21	Inventory/Total sales
X7	Total debt/shareholders' funds	X22	Net operating working /Total sales
X8	Total debt/Total assets	X23	Accounts receivable/Total sales
X9	Net operating working/Total assets	X24	Accounts payable/Total sales
X10	Long term debt/Total assets	X25	Current assets/Total sales
Solvency		Activity	
X11	Financial expenses/Total sales	X26	Cash flow/total sales
X12	Labor expenses/Total sales	X27	Total sales/total assets
X13	Financial debts/equity	X28	Value added/total sales
X14	Financial expenses/EBITDA	X29	Net income/value added
X15	Financial expenses/net income	X30	EBITDA/Total sales

EBIT, earnings before interest and taxes; EBITDA, earning before interest, taxes, depreciation, and amortization

Table 2 Confusion matrix for the prediction of corporate failure

		Actually	
		Failed	Healthy
Prediction	Failed	*True positive (TP)*	*False positive (FP)*
	Healthy	*False negative (FN)*	*True negative (TN)*

graphical plot used to represent the model performance while changing the cutoff value. In this case, the proportion of true positive and false positive are plotted on the *x*-axis and *y*-axis of the curve. AUC has become a widely used evaluation metric in corporate failure prediction because it is insensitive to the matrix of misclassification cost[2] to assess the discrimination ability of a model. In summary, two classifiers can be easily compared according to differences in the ROC curve performance. A classifier should get as close to the top left corner as possible, where its value will be close to 1.

With the data set mentioned above, a cross-validation loop (tenfold cross-validation repeated ten times) was performed to estimate the average evaluation measures. To compare the classifier performance, Demšar (2006) recommends a

[2]The misclassification of a failed firm (predict that a firm is healthy when it fails) represent a loss in capital, while the misclassification of a healthy firm (predict that a firm is failed when it survives) represents only a loss of commercial bargain. That is why, misclassified a failed firm is considered to be more costly.

Wilcoxon signed ranks non-parametric test because it only assumes limited commensurability and can be applied to prediction accuracy, misclassification errors or any other evaluation metric. It is expressed as follows:

Given R^+ be the sum of ranks when the second classifier outperforms the first one, R^- be the sum of ranks for the opposite and the ranks of $d_i = 0$ are split evenly among the sums:

$$R^+ = \sum_{d_i > 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i), \tag{14}$$

$$R^- = \sum_{d_i < 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i). \tag{15}$$

Let T be the smaller of the sums, $T = \min(R^+, R^-)$, the normal approximation can be used and the following statistic is used to calculate the z-statistics with a corresponding p-value:

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}. \tag{16}$$

However, Garcia and Herrera (2008) caution that several repeated pairwise comparison tests between algorithms conducted by us may lead to loss of control over family-wise errors.

4 Results

Experimental analysis is designed to compare the prediction ability of different ensemble methods based on extreme learning machine classifier. Table 3 indicates the evaluation metrics achieved to assess the performance of the methods. Furthermore, this table is complemented by Table 4, which highlights whether the differences between the methods are statistically significant.[3]

We first analyze the overall performance of the methods. Boosting ELM and Bagging ELM achieve the best mean accuracy values, 82.2% and 82.6%, respectively, while Random subspace ELM attains mean accuracy value of 81.7% and that of 81.4% is achieved with Multiple ELM. All ensemble methods are more accurate than the single ELM (80.4% of the mean accuracy). Thus, it confirms that ensemble ELM methods produce greater predictive power compared to a single ELM

[3] Appendix 1 shows the results on the database using ELM and ELM-ensemble methods. Figures 2 and 3 indicates the testing results with different number of hidden nodes and the average classification error of the ELM-ensemble methods as a function of the number of ensemble members.

Table 3 Performance of different ELM-based ensemble methods

	Accuracy	Type-I error	Type-II error	AUC
ELM	80.4%	21.7%	17.5%	0.821
Multiple ELM	81.4%	20.3%	16.7%	0.834
Bagging ELM	82.6%	18.2%	16.5%	0.849
Boosting ELM	82.2%	18.8%	16.8%	0.842
Random subspace ELM	81.7%	20.0%	16.6%	0.836

Table 4 Significance levels of a test of differences by method and evaluation metric

	Multiple ELM	Bagging ELM	Boosting ELM	Random subspace ELM
	Accuracy			
ELM	0.0866*	0.0001***	0.0012***	0.0338**
Multiple ELM		0.0463**	0.0971*	0.3372
Bagging ELM			0.2908	0.985*
Boosting ELM				0.2883
	Type-I error			
ELM	0.0976*	0.0001***	0.0001***	0.0652*
Multiple ELM		0.0179**	0.0751*	0.7871
Bagging ELM			0.5584	0.0386**
Boosting ELM				0.182
	Type-II error			
ELM	0.4275	0.0987*	0.4752	0.1255
Multiple ELM		0.7213	0.6531	0.6466
Bagging ELM			0.7889	0.6777
Boosting ELM				0.5133
	AUC			
ELM	0.0610*	0.0001***	0.0001***	0.0462**
Multiple ELM		0.0133**	0.1170	0.8674
Bagging ELM			0.2891	0.0811*
Boosting ELM				0.3746

*Significant at 10% threshold; **Significant at 5% threshold; ***Significant at 1% threshold

classification. The fact that Bagging and Boosting ensembles lead to the best reduction in the generalization error is not entirely surprising, as it is well documented their robustness to overfitting (Xiao et al., 2013; González et al., 2020). In contrast, variation of the parameters of the classifiers, such as Multiple ensemble and Random Subspace, can generate greater diversity (Bi, 2012). Nonetheless, the information perceived by the varying diversity does not generate consistent guidance so that the ensemble classifier can obtain a good generalization. On the whole, the key of Boosting and Bagging is that they build a set of diverse classifiers, while they benefit from the balance between diversity and accuracy, which is an important determinant of the performance of ensemble classifiers.

Secondly, we find no uniform improvement among the ensemble methods. If the misclassification errors are analyzed, Boosting ELM and Bagging ELM, here as

well, lead to lower misclassification error for failed firms, 18.8% and 18.2%, respectively, significant at 1% threshold in comparison with ELM. In contrast, we do not observe any significant differences in misclassification error for non-failed firms across ensemble methods; rather, the mean type-II error ranges from 16.5% with Bagging ELM and Random Subspace ELM to 18.8% with Bagging ELM.

Finally, the Bagging and Boosting ELM-based methods lead to higher AUC values than the other ensemble methods, which is in line with the previous results. In particular, Bagging ELM seems to be the most optimal ensemble method for corporate failure prediction as results are significantly better than those achieved with the other ensemble methods, but with respect to Boosting ELM.

In sum, the better overall prediction of Bagging and Boosting methods over the other ensemble methods, as previously observed, is due to their capacity to better identify failed firms. The superiority of Bagging ELM is based on the creation of a unique training set for each ensemble member because the perturbation generated in the learning set causes a significant change in the prediction constructed. As a model's prediction is order-correct for most of the replicated observation, the bagging-based ELM can be transformed into a nearly optimal predictor, in particular, for failed firms. Furthermore, one of major reasons why boosted ELM better identifies failed firms may be due to the fact that the new classifier generation gives more relevance to misclassified observation, mostly failed firms. That is, the likelihood of instances that have been misclassified by the previously generated classifier increases, and the set of classifiers grows progressively diverse. This trend explains why this method provides higher accuracy for the minority class without jeopardizing the accuracy of the majority class.

4.1 Further Validation

In order to further evaluate the effectiveness of the ensemble extreme learning machine for the corporate failure prediction task, a new data set has been collected. In general, there is no universal accepted definition of corporate failure; bankruptcy, the more severe form of failure, is commonly used. The popularity of bankruptcy as the definition of failure is based on two concepts: on the one hand, it provides an objective criterion to distinguish failed and non-failed firms, and, on the other hand, the moment of failure can be dated when a firm fills in the bankruptcy procedure. Therefore, the bankruptcy notion offers a discrimination criterion for obtaining a well-defined dichotomy, or at least, a representation of corporate failure, that can be applied methodologically. Nonetheless, numerous studies (Sun et al., 2014; Brédart et al., 2021) consider that corporate failure begins when a firm experiences financial distress. That is, when a firm encounters financial difficulties or struggles to fulfill its obligations. Accordingly, we collected a data set considering financial distress as the definition of corporate failure. We consider the criterion provided by Balcaen et al. (2011), who define financial distress as a firm with negative recurring profit after

Table 5 Performance of different prediction methods

	Accuracy	Type-I error	Type-II error	AUC
ELM	78.2%	24.7%	18.9%	0.790
Multiple ELM	79.5%	23.0%	18.0%	0.804
Bagging ELM	81.1%	20.7%	17.1%	0.824
Boosting ELM	80.5%	21.4%	17.6%	0.812
Random subspace ELM	80.0%	22.1%	17.9%	0.808

Table 6 Significance levels of a test of differences by method and evaluation metric

	Accuracy			
	Multiple ELM	Bagging ELM	Boosting ELM	Random subspace ELM
ELM	0.0753*	0.0001***	0.0032**	0.0217**
Multiple ELM		0.0265**	0.1333	0.2766
Bagging ELM			0.1267	0.0836*
Boosting ELM				0.3045
	Type-I error			
	Multiple ELM	Bagging ELM	Boosting ELM	Random subspace ELM
ELM	0.0592*	0.0001***	0.0001***	0.0154**
Multiple ELM		0.0144**	0.0869*	0.1936
Bagging ELM			0.1709	0.0935*
Boosting ELM				0.2423
	Type-II error			
	Multiple ELM	Bagging ELM	Boosting ELM	Random subspace ELM
ELM	0.2611	0.0348**	0.0107	0.2414
Multiple ELM		0.2560	0.3987	0.5612
Bagging ELM			0.6214	0.3521
Boosting ELM				0.3951
	AUC			
	Multiple ELM	Bagging ELM	Boosting ELM	Random subspace ELM
ELM	0.0509*	0.0001***	0.0028***	0.0131**
Multiple ELM		0.0106**	0.1635	0.5145
Bagging ELM			0.0958*	0.0439**
Boosting ELM				0.3153

*Significant at 10% threshold; **Significant at 5% threshold; ***Significant at 1% threshold

taxes over two consecutive years. Consequently, the collected dataset is composed of 2500 failed and 2500 non-failed firms.[4]

The results presented in Tables 5 and 6 are consistent with those of the previous ones. Boosting ELM and Bagging ELM achieve the highest accuracy values, in particular, due to their effectiveness in the reducing the type-I error in comparison to

[4]To design the prediction methods, the same procedure used in Sect. 3.2 was followed. Then, they were evaluated based on a 10-cross validation and using the abovementioned evaluation metrics.

the single ELM.[5] Moreover, it is important to mention that the prediction performance of the methods in this data set is inferior to the previous one. Thus, it is more arduous to differentiate failed firms from healthy ones in the initial steps of failure, when firms just experience financial distress. The literature documented that firms have shown a certain resilience for a long time, even though their financial situation resembles to a bankrupt one (Iftikhar et al., 2021). In contrast, firms that seem completely sound may suddenly fail. Therefore, the inability to know whether the echoes of financial distress may result in corporate failure makes it difficult to capture distinguishable factors that might reinforce model accuracy. That is why the performance of models is lower when corporate failure is represented as financial distress than when it is defined as bankruptcy.

5 Conclusion

In this study, we propose to evaluate several ensemble methods applied to corporate failure prediction in order to improve the classification performance of ELM. An ensemble strategy that combines the predictions of individual models is more performance-based than relying on the prediction capacity of a single model. Our results confirm that the Extreme Learning Machine-based ensemble is more accurate and robust than the "individual best" ELM model using two real financial datasets. In particular, the ensemble methods used in this study increase, on average, the classification accuracy estimated for the single ELM by 1.6 and 2.1 percentage points for the bankruptcy data and financial distress data, respectively. An increase in prediction performance of these magnitudes may seem modest, but the readers need to understand that financial institutions and banks can save a huge amount of the limited financial resources with decision technology that can increase the prediction power by 2%.

As Bagging ELM and Boosting ELM give similar results – there is some evidence that the bagging strategy is more effective for the prediction of corporate failure using ELM – it is arduous to make a design recommendation for which method is more optimal. However, we do notice that both methods, which operate by taking a base learner and invoking it multiple times using different training sets, are most effective in the ensemble ELM prediction method. We also notice that bagged ELM is more computationally efficient, as it requires 40–50 ensemble members, while 60–70 members as necessary for the boosting ensemble.

Acknowledgments We sincerely thank Prof. Abedin and Prof. Hajek for their assistance.

[5] The Appendix 2 shows graphically the testing results with different hidden nodes (Fig. 4) and the average classification error of ELM-ensemble methods as a function of ensemble members (Fig. 5).

Appendices

Appendix 1

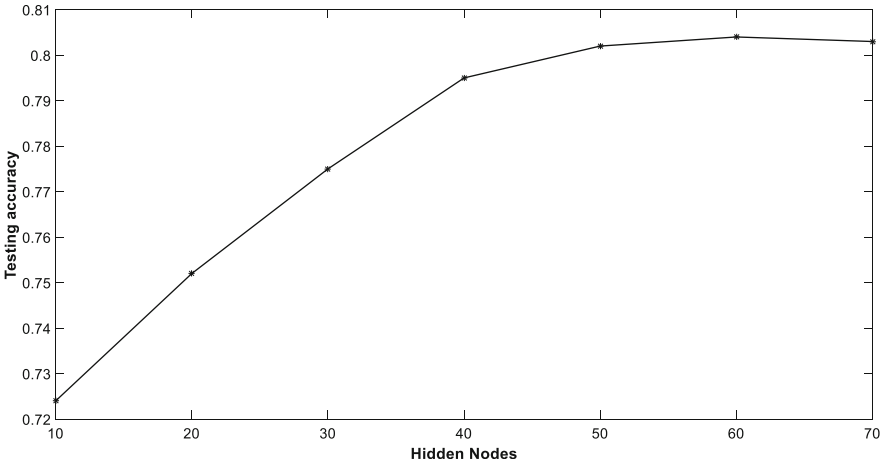

Fig. 2 Testing results for different hidden nodes in ELM for bankruptcy data

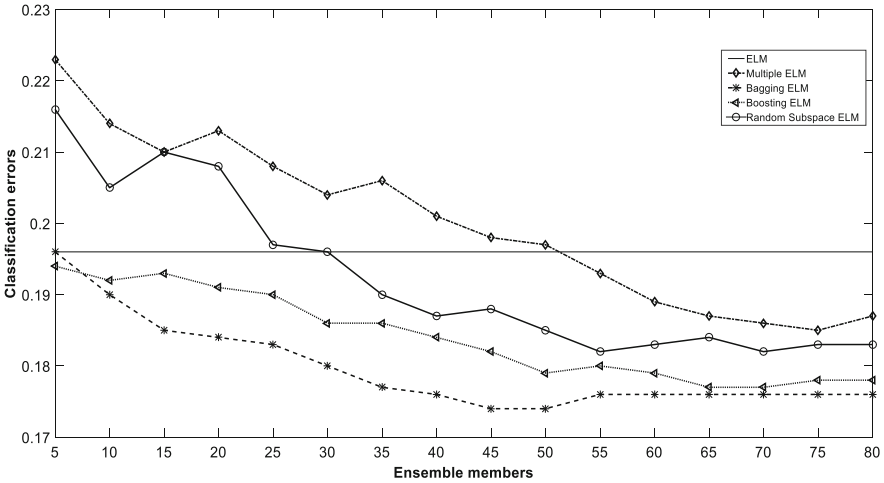

Fig. 3 Average classification errors of the Ensemble ELM methods by ensemble members for bankruptcy data

Appendix 2

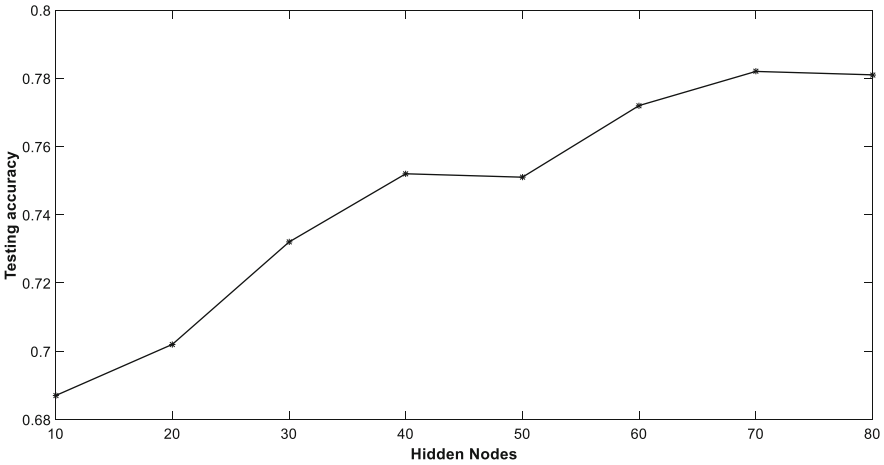

Fig. 4 Testing results for different hidden nodes in ELM for financial distress data

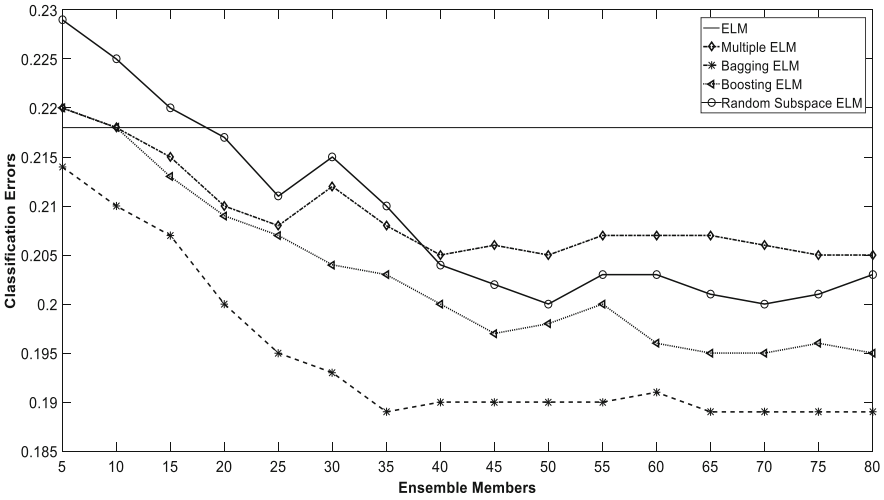

Fig. 5 Average classification errors of the Ensemble ELM methods by ensemble members for financial distress data

References

Abedin, M. Z., Chi, G., Colombage, S., & Moula, F. E. (2018). Credit default prediction by using a support vector machine and a probabilistic neural network. *Journal of Credit Risk, 14*(2), 1–27.

Abedin, M. Z., Hassan, M. K., Petr, H., & Uddin, M. M. (2021). Machine learning in finance and accounting. In *The essentials of machine learning in finance and accounting,* Taylor & Francis.

Abedin, M. Z., Chi, G., Hajek, P., & Tong, Z. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747-021-00614-4

Akusok, A., Veganzones, D., Miche, Y., Björk, K. M., Du Jardin, P., Severin, E., & Lendasse, A. (2015). MD-ELM: Originally mislabeled samples detection using OP-ELM model. *Neurocomputing, 159*, 242–250.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609.

Balcaen, S., Manigart, S., & Ooghe, H. (2011). From distress to exit: Determinants of the time to exit. *Journal of Evolutionary Economics, 21*, 407–446.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, 4*, 71–111.

Bi, Y. (2012). The impact of diversity on the accuracy of evidential classifier ensembles. *International Journal of Approximate Reasoning, 53*(4), 584–607.

Brédart, X., Séverin, E., & Veganzones, D. (2021). Human resources and corporate failure prediction modeling: Evidence from Belgium. *Journal of Forecasting, 40*(7), 1325–1341.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises: An empirical analysis of Italian firms. *Journal of Business Research, 68*(5), 1012–1025.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research, 7*, 1–30.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning* (pp. 148–156). IEEE.

Garcia, S., & Herrera, F. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research, 9*(12), 2677–2694.

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion, 64*, 205–237.

Grandvalet, Y. (2004). Bagging equalizes influences. *Machine Learning, 55*(3), 251–270.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006a). Extreme learning machine: Theory and applications. *Neurocomputing, 70*(1), 489–501.

Huang, G. B., Chen, L., & Siew, C. K. (2006b). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks, 17*(4), 879–892.

Iftikhar, A., Purvis, L., & Giannoccaro, I. (2021). A meta-analytical review of antecedents and outcomes of firm resilience. *Journal of Business Research, 135*, 408–425.

Kainulainen, L., Miche, Y., Eirola, E., Yu, Q., Frénay, B., Séverin, E., & Lendasse, A. (2011). Ensembles of local linear models for bankruptcy analysis and prediction. *Case Studies in Business, Industry and Government Statistics, 4*(2), 116–133.

Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373–3379.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226–239.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques: A review. *European Journal of Operational Research, 180*(1), 1–28.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113–141.

Moula, F. E., Chi, G., & Abedin, M. Z. (2017). Credit default prediction modeling: An application of support vector machine. *Risk Management, 19*(2), 158–187.

Ouenniche, J., & Tone, K. (2017). An out-of-sample evaluation framework for DEA with application in bankruptcy prediction. *Annals of Operations Research, 254*(1), 235–250.

Rao, C. R., & Mitra, S. S. K. (1971). *Generalized inverse of matrix and its application* (Wiley Series in Probability and Mathematical Studies). Wiley.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1), 1–39.

Séverin, E., & Veganzones, D. (2021). Can earnings management information improve bankruptcy prediction models? *Annals of Operations Research, 306*(1), 247–272.

Shahriare S, Khair A, Abedin MZ (2021, December 19–21). Performance analysis of machine learning techniques that predict hotel booking cancellations in hospitality industry. In *ICCIT 2020: 23rd International Conference on Computer and Information Technology, Dhaka*.

Shajalal, M., Abedin, M. Z., & Uddin, M. M. (2021). Handling class imbalance data in business domain. In: *The essentials of machine learning in finance and accounting*. Taylor & Francis.

Sun, J., Li, H., Huang, Q. H., & He, K. Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems, 57*, 41–56.

Veganzones, D., & Severin, E. (2020). Corporate failure prediction models in the twenty-first century: A review. *European Business Review, 33*(2), 204–226.

Verikas, A., Kalsyte, Z., Bacauskiene, M., & Gelzinis, A. (2010). Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey. *Soft Computing, 14*(9), 995–1010.

Xiao, T., Zhu, J., & Liu, T. (2013). Bagging and boosting statistical machine translation systems. *Artificial Intelligence, 195*, 496–527.

Yu, Q., Miche, Y., Séverin, E., & Lendasse, A. (2014). Bankruptcy prediction using extreme learning machine and financial expertise. *Neurocomputing, 128*, 296–302.