IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 25, 2021, accepted June 22, 2021, date of publication June 29, 2021, date of current version July 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3093461

Comparing the Performance of Deep Learning Methods to Predict Companies' Financial Failure

H. ALJAWAZNEH[1], A. M. MORA[2], P. GARCÍA-SÁNCHEZ[1], AND P. A. CASTILLO-VALDIVIESO[1]

[1]Department of Computer Architecture and Technology, ETSIIT-CITIC, University of Granada, 18071 Granada, Spain
[2]Department of Signal Theory, Telematics and Communications, ETSIIT-CITIC, University of Granada, 18071 Granada, Spain

Corresponding author: H. Aljawazneh (huthaifa.jawazneh@gmail.com)

This work was supported in part by the Ministerio de Ciencia, Innovación y Universidades under Project RTI2018-102002-A-I00, in part by the Ministerio de Economía y Competitividad under Project TIN2017-85727-C4-2-P and Project PID2020-115570GB-C22, in part by the Fondo Europeo de Desarrollo Regional (FEDER) and Junta de Andalucía under Project B-TIC-402-UGR18, and in part by the Junta de Andalucía under Project P18-RT-4830.

ABSTRACT One of the most crucial problems in the field of business is financial forecasting. Many companies are interested in forecasting their incoming financial status in order to adapt to the current financial and business environment to avoid bankruptcy. In this work, due to the effectiveness of Deep Learning methods with respect to classification tasks, we compare the performance of three well-known Deep Learning methods (Long-Short Term Memory, Deep Belief Network and Multilayer Perceptron model of 6 layers) with three bagging ensemble classifiers (Random Forest, Support Vector Machine and K-Nearest Neighbor) and two boosting ensemble classifiers (Adaptive Boosting and Extreme Gradient Boosting) in companies' financial failure prediction. Because of the inherent nature of the problem addressed, three extremely imbalanced datasets of Spanish, Taiwanese and Polish companies' data have been considered in this study. Thus, five oversampling balancing techniques, two hybrid balancing techniques (oversampling-undersampling) and one clustering-based balancing technique have been applied to avoid data inconsistency problem. Considering the real financial data complexity level and type, the results show that the Multilayer Perceptron model of 6 layers, in conjunction with *SMOTE-ENN* balancing method, yielded the best performance according to the *accuracy*, *recall* and *type II error* metrics. In addition, Long-Short Term Memory and ensemble methods obtained also very good results, outperforming several classifiers used in previous studies with the same datasets.

INDEX TERMS Economic forecasting, classification algorithms, machine learning, deep learning, data balancing.

I. INTRODUCTION

The problem of bankruptcy prediction has attracted the attention of researchers since the Crash of 1929 [1]. The effects of bankruptcy on a company are of great significance, as they affect a large number of stakeholders, including workers, creditors and suppliers, and eventually, even entire countries. Machine Learning (ML), and more recently Deep Learning (DL) [2], have gained the interest of researchers in the financial area. More organizations are interested in collecting this important analytical information. However, the data related to the companies' financial status are inherently imbalanced, since the bankruptcy is relatively uncommon in real life [3]. Several studies have been focused on addressing the lack of patterns of minority classes, such as bankrupt companies in our problem, because it is dramatically affecting the classifiers, causing a decrease in their reliability and performance. The reason is that those methods tend to build a model to predict the majority class. Thus, many balancing techniques have been proposed in order to solve this problem, using their own criteria to balance the data. We have considered the most appropriate and relevant and applied them to financial data.

Moreover, this work aims to advance the research line on bankruptcy prediction that started in [4], in which we compared the performance of several 'classic' classifiers, namely: Random Forest (RF), Naïve Bayes, and J48 to predict

The associate editor coordinating the review of this manuscript and approving it for publication was Joanna Kołodziej.

Spanish companies' financial status. As previously said, the dataset used in that study was extremely imbalanced, so, three balancing techniques (random undersampling, random oversampling and hybrid undersampling-oversampling techniques) were used to avoid the inconsistency problem. Later, in [5], in order to improve the performance of J48, KNN and MLP (Multilayer Perceptron) classifiers, these classifiers were combined with simple deterministic Delay Line Reservoir (DLR) [6] status space. Thus, DLR improved the performance of the classifiers regarding predicting companies' financial status compared to using normal ensemble voting or standalone classifiers.

Also, in this research line, several oversampling techniques were analyzed in order to solve the inconsistency problem. In [7], C4.5 decision tree was used to predict the financial status of Spanish companies. The results showed how the SMOTE-ENN (Synthetic Minority Oversampling TEchnique with Edited Nearest Neighbor) balancing technique obtained superior results according to the metrics used for the performance evaluation.

This paper aims to go a step further, using more advanced classification methods to improve on previous results. To this aim, we have considered DL techniques [8]. DL is a sub-field of ML, whose methods are achieving outstanding success compared to classical ML algorithms in many applications, especially with big data. We have chosen these advanced classification methods since their proven high performance was also reached when dealing with financial data [9]–[12].

Thus, in this study several DL methods, i.e., Deep Belief Network (DBN) [13], Multilayer Perceptron model of 6 Layers (MLP-6L) [14] and Long-Short Term Memory (LSTM) [15], have been considered to predict companies' financial failure. In addition, we have also applied three *bagging* [16] based ensemble methods, i.e, Random Forest (RF) [17], Support Vector Machine (SVM) [18] and K-Nearest Neighbor (KNN) [19], as well as two *boosting* based ensemble methods, i.e, Adaptive Boosting (AdaBoost) [20] and eXtreme Gradient Boosting (XGBoost) [21], given their good achievement in several classification problems shown in the literature. It is noteworthy that RF was the best method in the context of financial data in our previous paper [4].

Therefore, each of the selected DL classifiers belongs to a different type of neural network, aiming to 'cover' the search space in different ways. Thus, MLP-6L is a feed-forward neural network, LSTM is a recurrent neural network, and DBN is a greedily learning stochastic neural network comprised of directed and undirected layers. On the other hand, in order to improve the performance of SVM and KNN, we consider both of them as ensemble models using the bagging technique, whereas RF is an ensemble of decision trees based on bagging. In addition, to making the comparison scope wider, we propose using AdaBoost and XGBoost which are boosting methods. This selection aims to compare different types of approaches that can lead to the identification of 'the best scheme'.

Accordingly, the performance of these methods when working with three complicated financial datasets, consisting of real data with a *highly imbalanced ratio will be tested.* Thus, real Spanish, Taiwanese and Polish companies' datasets have been considered in order to evaluate the efficiency of the methods to predict the companies' financial failure. As previously mentioned, bankruptcy is rare in the real data, so the datasets that we have used in this study are extremely imbalanced. The major differences between these datasets are the complexity level and the data type's diversity. The Spanish companies' dataset is a combination of nominal and numerical attributes values, and contains financial and non-financial data. On the other hand, the Taiwanese and Polish companies' datasets are more complicated according to the number of attributes and records. The Taiwanese dataset contains the largest number of attributes, while the Polish dataset contains the largest number of samples. Both of them contain only numerical financial attributes. These major differences could affect the behavior of the classifiers and the results. In addition, it might lead to making more accurate decisions about the most appropriate classifier to predict companies' financial failure.

To handle the data inconsistent distribution problem, eight advanced balancing techniques from the literature have been applied in the preprocessing stage, namely, SMOTE (Synthetic Minority Oversampling TEchnique) [22], BL-SMOTE (Borderline SMOTE) [23], SMOTE-ENN [24], K-means SMOTE [25], SMOTE-NC (SMOTE Nominal-Continuous) [22], SMOTE-Tomek (SMOTE with Tomek links) [24], SVM-SMOTE (Support Vector Machine with SMOTE) [26] and ADASYN (ADaptive SYNthetic sampling approach) [27]. These resampling techniques significantly enhance the behavior of the classifiers, i.e., they dramatically decrease the classifiers' minority class misclassification. Thus, they had been utilized by several researchers to solve the data inconsistency problem [7], [9]–[11], [28].

Generally, data balancing can be done using one of the following methods: Oversampling, Hybrid Oversampling-Undersampling and Clustering-based techniques. Thus, the techniques selected for this study cover these three data preparation procedures in order to study their influence on the classifiers' performance. The aim is to get to a final decision about the best 'DL/Data balancing' combination to address this financial problem.

Finally, the performance of the proposed methods cannot be evaluated by just considering the usual *accuracy* measure, since in extremely imbalanced data, this is not a reliable value, i.e., the minority class could always be misclassified and the accuracy will be very high. Due to this fact, besides *accuracy*, we have considered *recall*, *specificity*, *precision*, *type I error* and *type II error* as metrics to evaluate the performance of all classifiers. Recall and type II error represent each model's bankruptcy hit and misclassification rates. Specificity and type I error is the solvency hit and misclassification rates. The precision metric shows the performance of each

model with respect to predicting the correct status for each company.

In other words, as aforementioned, due to the importance of the prior knowledge about companies' financial failure to stakeholders, creditors and suppliers, we discuss the use of DL methods as robust tools that could yield very high performance in predicting companies' financial failure compared to standard ML methods. The major problem with bankruptcy real datasets is the inconsistent distribution, which badly affects the reliability of classifiers and strongly raises the need for using balancing techniques. Thus, DL methods with advanced balancing techniques could show outstanding results regarding bankruptcy prediction and outperforming many methods addressed in the literature. On the other hand, the performance of solvent companies' prediction is also an important issue when data balancing is applied, since both classes have a comparable amount of samples (i.e., it is almost similarly hard to predict each class). Moreover this forecasting contributes to improving the overall performance of the classifiers, and helps to make a reliable judging about the companies' financial status. Given this, DL algorithms could reach very high performance in the prediction of solvent companies as well.

Thus, the main contributions of our work are summarized as follows:

- We conduct a complete analysis on the performance of a wide amount of classification techniques working with three real datasets, one of them only available to the authors (the Spanish companies' data).
- The study included different DL methods as solid alternatives outperforming several ensemble classification methods utilized to predict companies' financial failure in the state of the art. This is an advance over our previous works, in which no DL methods were applied.
- We present a novel comparison between three different DL methods, i.e., DBN, LSTM and MLP-6L, and five ensemble classifiers, i.e., RF, SVM, KNN, AdaBoost and XGBoost, in predicting companies' financial failure.
- Given the scarcity of financially failed companies in the real world, the real companies' datasets are extremely imbalanced. Thus, we discuss the impact of several advanced balancing techniques on the DL and ensemble classifiers' behaviors, concluding the most highly recommended technique to improve the reliability of all classifiers' predictions in this situation.

The rest of the paper is organized as follows: first, we describe related works that use ML (including DL) methods to predict companies' financial status using balanced and imbalanced data. After this, the datasets used are described in Section III. Then, the classification algorithms compared are described in detail in Section IV, whereas the data balancing techniques are introduced in Section V. Section VI describes the experimental setup and considered metrics. The experiments' procedures and obtained results are presented and analyzed in Section VII. Best approaches are compared with previous algorithms reported in the state of the art in Section VIII. Finally, Section IX concludes and summarizes the findings and provides directions for future work.

II. RELATED WORKS

Financial failure prediction is a critical matter that occupies the efforts of many researchers, since an inaccurate decision about the companies' financial status could cause costly financial losses. Mostly, the prediction of companies' financial status could be done using statistical techniques such as Linear Discriminant Analysis (LDA), Multi-Discriminant Analysis (MDA) and Logistic Regression (LR or Logit); or by ML algorithms [29]. In the sixties, Altman [30] used MDA to predict companies' financial status using their financial statements. Later, Ohlson [31] adopted Logit to predict companies' financial failure. Brozyna *et al.* [32] used LDA and LR to predict the financial status of Polish and Slovak companies. Jones and Hensher [33] proposed a mixed Logit model, and compared it with a standard Logit model in predicting companies financial distress, proving that the mixed Logit model yields better results than the standard one. More recently, several researchers have compared the statistical techniques with ML techniques on forecasting companies' financial failure. For instance, Pompe and Feelders [34] compared the performance of LDA with classification trees and neural networks in this problem, and proved that neural networks outperform the rest of methods. Min and Lee [35] compared SVM, MDA, Logit and three-layer fully connected back-propagation neural networks regarding bankruptcy prediction, with SVM obtaining the best results. However, in recent studies, ML algorithms showed better performance than the statistical models concerning bankruptcy prediction [35]. For this reason, many researchers have considered it as a classification problem, and have applied standard ML classification or regression methods for prediction [4], [10], [36]–[38].

In addition, some researchers combined several ML algorithms in order to improve the efficiency of the companies' financial failure prediction. Fedorova *et al.* [12] applied several combinations of RBF (Radial Basis Function) network and MLP in order to predict Russian companies' bankruptcy, applied to a balanced dataset (2906 samples) from all the available data. Iturriaga and Sanz [39] combined MLP and SOM (Self-Organized Maps) in order to predict US banks' financial failure up to three years before it occurs. Another balanced dataset of 754 samples was used by Lanbouri and Achchab [40], who proposed a hybrid model (DBN and SVM) to predict French companies' financial distress from a balanced dataset of 966 samples. However, the authors of these works used just one relatively small dataset to evaluate the performance of the proposed combinations of algorithms.

Datasets considered for bankruptcy prediction datasets are not usually balanced, as only a small percentage of companies go bankrupt in real life. Due to this reason, it is necessary to rely on data balancing techniques. SMOTE and its

variants have been applied in several works. For instance, Kim *et al.* [41] used it in combination with their Geometric Mean based Boosting (GMBoost) algorithm and obtained very good results. Islam *et al.* [38] also applied SMOTE to resample an extremely imbalanced dataset in the preprocessing stage, showing an improvement in the performance of the 13 classification and regression algorithms compared. SMOTE was also used in [28], in a combination with different classical classification methods. The author discusses the ideal circumstances to use several balancing techniques, and also the advantages of considering different datasets (Japanese and American companies in this case) to understand how the methods behave in this kind of problems. Because of this, we have also considered more than one dataset in order to improve our insights.

SMOTE variants have been compared as well. Le *et al.* [10] discussed the impact of using different balancing techniques, including the SMOTE variants, on Korean companies' bankruptcy prediction performance. Four classification models were applied to predict the financial status, namely; RF, Decision tree, MLP and SVM. The dataset treated was extremely imbalanced, so, five balancing techniques were tested: the SMOTE variants (SMOTE, BL-SMOTE, SMOTE-ENN, SMOTE-Tomek) and ADASYN. Furthermore, the classification models were applied to the data before and after balancing. RF outperformed the other models in both cases, but using RF with SMOTE-ENN obtained the best results. Consequently, as RF was the superior classifier in that study, we have also adopted RF to predict companies' financial failure, besides other DL and ensemble methods, as previously stated.

On the other hand, according to the positive impact of the data balancing methods shown in the literature, SMOTE, BL-SMOTE, SMOTE-ENN, SMOTE-Tomek and ADASYN have been utilized in our study in order to address the data balancing issue.

One of the motivations of our study is to use DL algorithms as very powerful tools to predict companies' financial failure. Indeed, there are not many studies that apply DL methods to predict companies' financial failure using companies' real data.

Jang *et al.* [9] compared LSTM, Feed-forward neural network and SVM regarding predicting Business Failure relying on listed US construction contractors. The same authors [11] also proposed a model based on LSTM to predict the business failure probability from one to three years using accounting, construction market and macroeconomic variables. Moreover, the SMOTE-Tomek balancing technique was used as a data preprocessing stage in both works, obtaining better results than using only the accounting variables. Therefore, after a second successful application of LSTM and SMOTE-Tomek in the literature, we decided to use these methods in our work and compare their performance to other DL methods and balancing techniques.

Following a different approach, some researchers used financial data as a graphical representation. Yeh *et al.* [42], predicted companies financial status using DBN, the return of stock markets for solvent and bankrupt companies were presented as binary images and then were utilized in order to train the models. They proved that DBN outperforms SVM classical classification method. Also, Hosaka [43] proposed a method based on Convolutional Neural Networks (CNN) to predict bankruptcy using Japanese stock market data represented as a grayscale image. Moreover, the proposed method obtained the optimum results compared to classical and other DL classification methods.

With respect to the datasets considered here, the Spanish companies' dataset has been used as a test-bed in previous works. In the first work [44], it was used to train and compare different neural network architectures. Later, Alfaro-Cid *et al.* [45] combined MLP with genetic programming to predict financial book losses. This dataset has gained attention recently. Jawazneh *et al.* [4] compared the performance of three classical classifiers (RF, Naïve Bayes and J48), in conjunction with three simple balancing techniques. RF outperformed the rest of the classifiers, obtaining the best accuracy, sensitivity and specificity.

This dataset has also been used to compare different combinations of balancing techniques and classification methods. For example, [5] used DLR and MLP, in combination with SMOTE, while [7] used only C4.5 as a classification method, but compared 11 balancing techniques, where SMOTE-ENN obtained the best results. Recently, [46] discovered that the combination of SMOTE and AdaBoost guaranteed promising results compared with basic and ensemble classifiers, as well as using five different Feature Selection approaches. In that study, AdaBoost showed higher performance than the rest of the classifiers applied to the imbalanced dataset without using data resampling techniques, but it did not provide a considerable alternative to use the balancing techniques. Finally, [36] compared the performance of combining three cost-sensitive methods, with several ensemble classifiers. The combination of RF with cost-sensitive classification methods outperformed the rest. Our work presents a further step in this state of the art by utilizing advanced DL methods and balancing techniques to be compared using this dataset.

However, as aforementioned, dealing with only one dataset may not be enough to reach firm and reliable conclusions. For this reason, we also have used the dataset described in [47]. Data from Polish companies that went bankrupt between 2007 and 2013, and from companies that continued to operate between 2000 and 2012, were used to create an extremely imbalanced dataset. In the same work, the Polish companies' dataset was used to compare several classifiers' performance with a novel approach that applies EXtreme Gradient Boosting (EXGB) for learning an ensemble of decision trees, obtaining significant results with respect to the referenced methods they applied, such as J48, RF, SVM and AdaBoost.

Moreover, we also consider the dataset faced in [48]. The authors studied the impact of combining the Financial Ratios (FRs) and Corporate Governance Indicators (CGIs) on

the classifiers' performance in predicting Taiwanese companies' financial status. The problem of the inconsistent data distribution was solved by selecting a balanced subset using stratified sampling method. The selected subset contained 239 records for bankrupt companies and another 239 records for solvent companies. Thus, five well-known classifiers were compared, i.e., SVM, KNN, CART, MLP and Naïve Bayes. Then, combining the FRs and the CGIs improved the performance of the classifiers, Stepwise Discriminant Analysis (SDA) Feature Selection method with SVM obtained the best results.

We have considered three datasets in our study in order to evaluate the performance of the classifiers more accurately. The main differences between these datasets are the complexity level and the type of data, given that the Spanish dataset is the simplest and the Polish dataset is the most complicated. In addition, we can compare the performance of our methods with those reported in [47] and [48], in which the same Polish and Taiwanese companies' datasets were considered.

As far as we know, there are no attempts in the literature to compare DL methods (DBN, MLP and LSTM) with ensemble methods (RF, SVM, KNN, AdaBoost and XGBoost) to solve this type of classification problem, even with extremely imbalanced datasets. Thus, in this paper, three different types of balancing techniques used in previous works, have been applied: oversampling (SMOTE, Borderline SMOTE, SMOTE-NC, SVM-SMOTE and ADAYSN), combination of undersampling-oversampling techniques (SMOTE-ENN and SMOTE-Tomek), and clustering-based balancing (K-means SMOTE).

III. DATASETS CONSIDERED

As stated, in this study the problem of predicting companies' financial failure using Spanish, Taiwanese and Polish companies' data has been addressed as a classification problem. The Spanish companies' dataset was obtained from *Infotel* database, a company devoted to gathering information in several domains about companies in Spain. A combination of financial and non-financial real data from 471 companies in Spain during six years (1998 to 2003) has been used. In addition, the Spanish companies dataset used in this work includes particular domain attributes utilized in order to find out whether a company is bankrupt or solvent. It includes 2859 instances, where each one of them consists of 39 independent categorical and numerical variables. In this work, 33 variables have been adopted after eliminating irrelevant attributes (such as internal codes). Thus, 27 of the attributes are numeric, whereas the remaining attributes are categorical. Every instance describes a company for one year, and it contains *Bankruptcy* attribute to mention the financial status of that company. Table 1 shows the independent variables used from the Spanish companies' dataset after eliminating the unnecessary variables.

The Spanish companies' dataset is extremely imbalanced, 2797 records (98%) represent the majority class (solvent companies) and the remaining 62 records (2%) represent the minority class (bankrupt ones). Thus, this situation creates a challenge for the classifiers to work properly; since they always tend to take the easiest way which is predicting the majority class.

However, limiting this study to a single dataset may not be sufficient to understand the differences in the techniques compared. That's why we have also considered Taiwanese and Polish companies datasets to predict bankruptcy, given their differences with the Spanish companies' dataset, particularly, the complexity level and data types. The Taiwanese dataset is more complex than the Spanish one, and the Polish dataset is the most complex of them all. The Taiwanese companies' dataset was collected from the Taiwan Economic Journal over 10 years (1999 to 2009) and it contains 6819 records in total: 6599 records of solvent companies (97%), and the rest corresponding to bankrupt companies (220 records). Besides, it is comprised of 95 financial attributes. However, the companies in this dataset were selected according to two conditions, i.e., the information of each company should be available three years before the decision about its financial status, and the size of each company should match with quite enough companies to compare. On the other hand, the decision about each company's financial status is based mainly on Taiwan's stock exchange business regulations. Further information about this dataset is available in [48].

With respect to the Polish companies' dataset, it contains real data collected from Emerging Markets Information Service (EMIS), but focused on enterprises of that country. EMIS is a database comprised of information about emerging markets around the world. It is important to note that the bankrupt Polish companies' data was collected from 2007 to 2013, and the solvent companies' data refer to 2007 to 2012.

This dataset is also extremely imbalanced; the total amount of samples is 10,000, of which 203 are bankrupt (2.03%) and 9797 samples belong to solvent companies (97.97%). In addition, it has 64 numerical financial attributes, and there are not categorical values. More information about this dataset is available in [47]. It can be downloaded from the *Kaggle* ML community web.[1]

Thus, in the paper we consider these three datasets in order to compare the performance of three different DL methods and five different ensemble approaches, facing the financial failure prediction. The proposed methods are described in detail in Section IV.

Moreover, in order to deal with the extreme imbalanced situation in the datasets, several data balancing techniques have been applied in a preprocessing step, described in Section V.

IV. CLASSIFICATION ALGORITHMS COMPARED

In this section, three different types of advanced DL algorithms, three bagging ensemble and two boosting ensemble methods, which have proven their effectiveness concerning classification tasks, are presented.

[1] https://www.kaggle.com/c/companies-bankruptcy-forecast

TABLE 1. Spanish companies' dataset: financial and non-financial independent variables.

Financial Variables	Description	Type
Debt Structure	Long-Term Liabilities / Current Liabilities	Real
Debt Cost	Interest Cost / Total Liabilities	Real
Debt Paying Ability	Operating Cash Flow / Total Liabilities	Real
Debt Ratio	Total Assets / Total Liabilities	Real
Working Capital	Working Capital / Total Assets	Real
Warranty	Financial Warrant	Real
Operating Income Margin	Operating Income / Net Sales	Real
Return on Operating Assets	Operating Income / Average Operating Assets	Real
Return on Equity	Net Income / Average Total Equity	Real
Return on Assets	Net Income / Average Total Assets	Real
Stock Turnover	Cost of Sales / Average Inventory	Real
Asset Turnover	Net Sales / Average Total Assets	Real
Receivable Turnover	Net Sales / Average Receivables	Real
Asset Rotation	Asset allocation decisions	Real
Financial Solvency	Current Assets / Current Liabilities	Real
Acid Test	(Cash Equivalent + Marketable Securities + Net receivables) / Current Liabilities	Real
Non-financial Variables	**Description**	**Type**
Year	Corresponding to the sample	Integer
Size	Small|Medium|Large	Categorical
Number of employees		Integer
Age of the company		Integer
Type of company	Public Company|Limited Liability Company|Others	Categorical
Linked to a group	If the company is part of a group holding	Binary
Number of partners		Integer
Province code	Code of the location where the company is set	Categorical
Number of changes of location		Integer
Delay	If the company has submitted its annual accounts on time	Binary
Historic number of Judicial incidences	Since the company was created	Integer
Number of judicial incidences	Last year	Integer
Historic amount of money spent on judicial incidences	Since the company was created	Real
Amount of money spent on judicial incidences	Last year	Real
Historic number of serious incidences	Such as strikes, accidents...	Integer
Audited	If the company has been audited	Binary
Auditor's opinion	Favourable|Exceptions|Unfavourable	Categorical

As previously explained, the DL algorithms used (DBN, MLP-6L and LSTM) have been selected as representatives of different types of neural networks, while the ensemble methods (RF, SVM, KNN, AdaBoost and XGBoost) have shown high performance in classification problems in the literature; the RF achieved the best performance in our previous work.

A. DEEP BELIEF NETWORK (DBN)

DBN is a stochastic DL method proposed by Hinton *et al.* [13] in 2006, consisting of several-stacked Restricted Boltzmann Machines (RBM). RBM is an energy-based generative model composed of two layers, visible units and hidden units, where all units are fully bidirectional connected with symmetric weights between layers. As shown in Figure 1, a DBN consists of several stacked RBMs, where the hidden layer of the lower RBM represents the visible layer of the upper RBM, the links between the top two layers are undirected and the links between the remaining layers are directed. In addition, DBN trains greedily; each RBM trains unsupervised on a time. Therefore, the results of each RBM represent the input of the higher RBM, and the final results are fine-tuned with supervised learning.

Every hidden layer is modeled as h^i a binary random vector with elements h^i_j. Thus, Equations 1 and 2 parameterize the whole DBN model and each hidden layer probabilities, respectively,

$$P(V, h^1, \ldots h^\ell) = P(V|h^1)P(h^1|h^2).P(h^\ell - 1|h^\ell) \quad (1)$$

$$P(h^i|h^{i+1}) = \prod_{j=1}^{n_i} P(h^i_j|h^{i+1}) \quad (2)$$

Also, with h^i_j as a stochastic units and 1 as binary activation, Equation 3 represents the probability of each stochastic hidden unit.

$$P(h^i_j = 1|h^{i+1}) = sigm\left(b^i_j + \sum_{k=1}^{n_{i+1}} W^i_{jk} h^{i+1}_k\right) \quad (3)$$

$$sigm(x) = 1/(1 + exp(-x)) \quad (4)$$

The normal sigmoid (Equation 4) represents the activation function, b^i_j are the biases, W^i is the weights matrix. Each RBM follows equation 2 and equation 3 respectively in order; upward from bottom to top [49].

B. LONG-SHORT TERM MEMORY (LSTM)

LSTM is a specific type of Recurrent Neural Networks (RNN) that was proposed by Hochreiter and Schmidhuber in [15]. The essential unit of LSTM is a cell that replaces the hidden layer neurons of the RNN, and each cell is configured mainly by three gates: input gate, output gate and forget gate as shown in Figure 2.

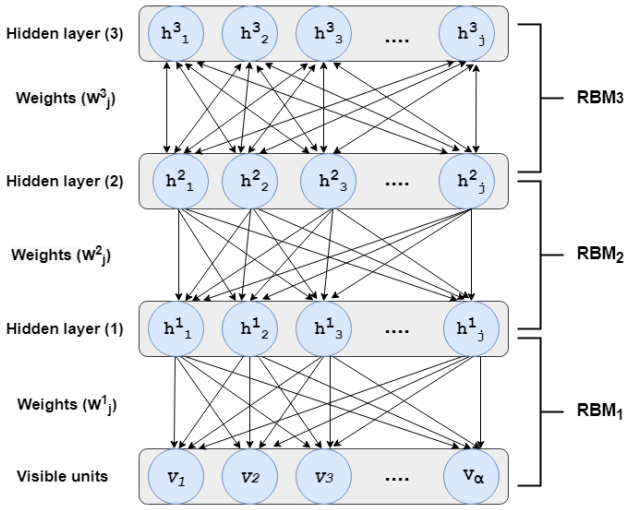

FIGURE 1. The structure of three hidden layers DBN (three RBMs).

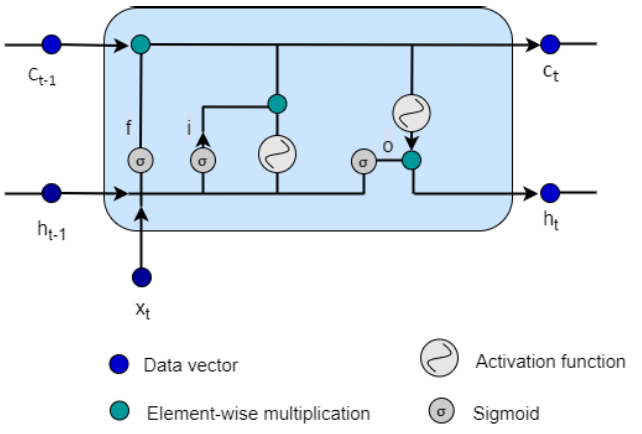

FIGURE 2. The structure of an LSTM memory cell [50].

LSTM architecture gives it the possibility to make a decision whether to forget or update the last hidden status with new information.

The following six equations describe the information processing steps of LSTM [50]:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (5)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (6)$$
$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (7)$$
$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \quad (8)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t + b_o) \quad (9)$$
$$h_t = o_t \odot tanh(c_t) \quad (10)$$

where t represents the time unit, f_t is the forget gate, i_t is the input gate, o_t is the output gate, W_* and U_* are the weight matrices, b_* are bias vectors, x_t is an input vector. Also, c_t represents the memory status vector, and h_t is the hidden status vector output obtained from c_t. in addition, σ is the sigmoid function, $\tilde{c}$ is the input modulation, h_t is the output and $\odot$ is a point-wise multiplication.

Therefore, in the first step of the LSTM process (Equation 5), the sigmoid function identifies the information that will be discarded according to its value. This step represents the forget gate procedure. The second step (Equations 6, 7 and 8) consists on placing the decision to update the information from the input according the sigmoid and *tanh* functions values. In other words, sigmoid makes the decision to update the information or discard the update, *tanh* obtain the value of weights, then the new cell state set by multiplying the values of the sigmoid and the *tanh*. In the final step (Equations 9 and 10) the output will be obtained by relying on the filtered version of the cell state.

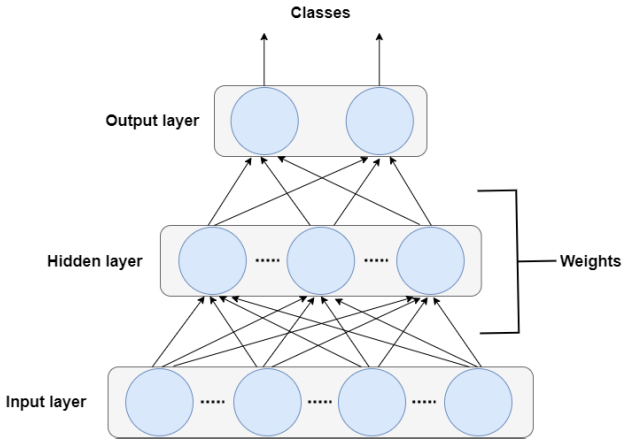

FIGURE 3. The structure of general three layers MLP.

C. MULTILAYER PERCEPTRON WITH 6 LAYERS (MLP-6L)

MLP is a feed-forward neural network, usually applied on supervised learning tasks, based on back-propagation learning [14]. It consists of a neural network with input, output and one or more parallel hidden layers. The architecture of the MLP is described as a fully interconnected network, as shown in Figure 3. However, increasing the number of the hidden layers transforms the MLP from classical learning method into a DL method [51]. In this study, a MLP model with four hidden layers has been used, so, there are six layers in total and thus, we refer to it as MLP-6L.

Each processing unit on each layer is connected with the whole units in the following layer by weighted connections [14], [52]. Also, the input values represent the information fed forward into the network. The processing of the information in the hidden units depends on the input information and the weight value of each input-hidden unit connection. Accordingly, the information obtained by the output units depends on the values of the hidden units and the weight value of each hidden-output units connection [52]. The MLP training is developed gradually in several stages: in each stage, the output units obtained results are compared with the real data allocated in the training data, then an error signal is

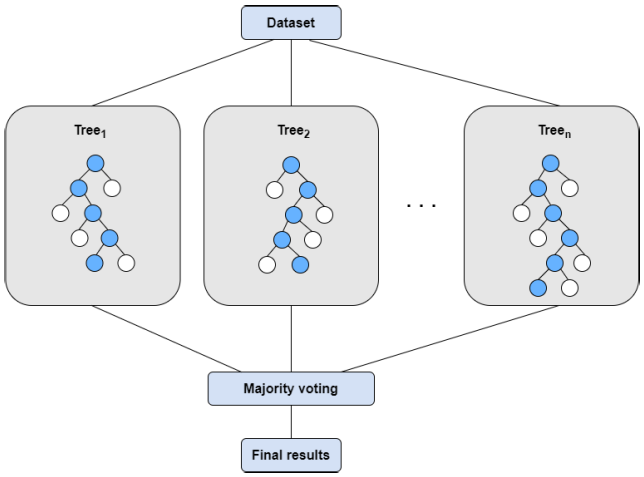

FIGURE 4. The structure of RF.

used to enhance the MLP expectation in the further stages in order to achieve almost identical values gradually [52]. The activation function that obtains the MLP output(h_j) is described in Equation 11 [14],

$$h_j = f\left(\sum_i x_i w_{ij} + b\right), \tag{11}$$

where $f()$ is the activation function, x_i is the activation of ith hidden layer unit, w_{ij} is the weight of the connection joining the jth neuron in a layer with the ith neuron in the previous layer, and b is the bias for the neuron.

Also, Equation 12 describes the error function that can be reduced by enhancing layers interconnection

$$E = \frac{1}{2}\sum_n \sum_k (t_k^n - h_k^n)^2, \tag{12}$$

where t_k^n is the calculated output, h_k^n is the actual output value, n is the number of sample and k is the number of output units.

D. RANDOM FOREST (RF)

Random forest (RF) is an ensemble classification method developed by Breiman [17] in 2001. It is based on the creation of different decision trees from different subsets of the original dataset. Usually, the *bootstrapping* method is used to create these datasets, while the different trees are created using C4.5 (a benchmark decision tree algorithm that stands mainly on entropy and gained values). The RF final classification results are the majority voting of these subtrees.

Figure 4 describes the architecture of RF model.

E. SUPPORT VECTOR MACHINE (SVM)

SVM is one of the most popular supervised ML algorithms proposed mainly for binary classification and regression problems by Cortes and Vapnik in [18]. Basically, it finds the proper separating hyperplane that maximizes the margin in the features space between the two classes. On the other hand, it is not mandatory that the data is linearly separable, thus, to avoid some complex calculations, kernel functions (such as Linear, Polynomial, Sigmoid and Gaussian Radial Basis function (RBF)) are used as a hyperparameter aiming to allocate the separating hyperplanes [53].

In addition, to improve the classification performance, we consider the SVM in this study as an ensemble model applying *bagging* [16]. This uses the *bootstrapping* method in order to create several subsets from the original dataset, and implements the model several times independently and aggregating the ensemble model's final results using majority voting.

F. K-NEAREST NEIGHBOR (KNN)

KNN is another widely used non-parametric ML algorithm proposed by Cover and Hart in [19]. It decides about the class label of each sample according to the similarity with its closest neighbors. Several distance algorithms (Such as Euclidean, Mahalanobis, Minkowski and Hamming) could be used to measure the similarity between samples [54].

Also, to avoid the confusion during the class assignment, the value of K is set to an odd number. In this study, we propose KNN as an ensemble model using the *bagging* method.

G. ADAPTIVE BOOST (AdaBoost)

AdaBoost is an iterative ensemble ML algorithm that applies its base classifiers in a sequence based on *boosting*, a technique that combines a set of 'weak' learners applied sequentially for developing a 'strong' learner [20]. In other words, AdaBoost applies the base classifier (normally a *Decision Tree*) several times iteratively. In the first iteration of the model, the weights are set equally to all samples, and in the remaining iterations, the weights increase for the misclassified samples and decrease for the correctly classified samples in the previous iterations in order to improve the overall performance of the model. The final results are the combination of the ensemble classifiers predictions using weighted majority voting [55].

H. EXTREME GRADIENT BOOSTING (XGBoost)

XGBoost is a relatively new ensemble *boosting* ML method proposed by Chen and Guestrin in [21]. The procedure of XGBoost is based mainly on Gradient Boosting but with further steps in order to improve the performance of predictions by controlling the overfitting using regularization. The base learner used in XGBoost is Classification And Regression Trees (CART). The final result is the sum of the CARTs' scores.

V. BALANCING TECHNIQUES APPLIED

In order to improve the performance and the reliability of the classifiers, it is essential to balance the datasets used for training, by either oversampling, undersampling or a mixture of both. Oversampling is the procedure of increasing the amount of the minority class instances, for instance replicating some of the existing samples (called *random oversampling*) or generating synthetic ones. On the other hand,

undersampling is the procedure to decrease the amount of the majority class instances, usually by removing random ones or following any other criterion to decide which ones to eliminate [56].

In this section, three different types of balancing techniques are presented, i.e., Oversampling (SMOTE, BL-SMOTE, SVM-SMOTE, ADASYN and SMOTE-NC), Hybrid Undersampling-Oversampling (SMOTE-ENN and SMOTE-TOMEK) and a clustering-based balancing technique (K-means SMOTE). These techniques have been chosen specifically due to the variety of their procedures in order to analyze their effects on the performance of the studied DL methods and RF algorithm.

1) SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

SMOTE is a more advanced technique than random oversampling. Instead, it generates new instances depending on the nearest neighbors of each instance in the minority class, by a combination of their features [22]. Many modifications have been made since its proposal.

2) BORDERLINE SMOTE (BL-SMOTE)

It is another oversampling technique based on the standard SMOTE, but targeting only the borderline minority class instances, that is, the ones closer to the line separating the data of two classes [23].

3) SMOTE WITH EDITED NEAREST NEIGHBOR (SMOTE-ENN)

This is a balancing method combining the oversampling using SMOTE and the undersampling using ENN. The ENN step calculates the nearest k neighbors of each instance, and if most neighbors are of a different class, it eliminates the instance [24].

4) K-MEANS SMOTE

It is a combination of clustering and oversampling, based on generating minority class samples, with SMOTE, only in safe and crucial areas. These areas are the clusters, calculated with K-Means, with a high ratio of minority observations [25].

5) SMOTE NOMINAL-CONTINUOUS (SMOTE-NC)

It is a variation designed in order to handle datasets consisting of nominal and continuous data. The categories of a newly generated sample are decided by picking the most frequent category of the nearest neighbors, instead of creating a synthetic continuous value for that feature [22].

6) SMOTE WITH TOMEK LINKS (SMOTE-TOMEK)

In this method, Tomek links are used to clean the overlap between classes, by detecting noise or border instances, and therefore establishing well-defined clusters in the dataset. Oversampling is performed with SMOTE as the name suggests [24].

7) SUPPORT VECTOR MACHINE WITH SMOTE (SVM-SMOTE)

In this case, SVM is applied in order to approximate the decision boundary and borderline. Then, for the instances far away from the borderline, an extrapolation technique is used to generate minority class instances. On the other hand, for the instances closer to the borderline an interpolation technique similar to SMOTE is used to generate the minority instances [26].

8) ADAPTIVE SYNTHETIC SAMPLING APPROACH (ADASYN)

ADASYN uses a weighted distribution for different samples of the minority class according to their level of learning difficulty. That is, more synthetic data is generated for the samples of minority classes that are more difficult to learn [27].

VI. EXPERIMENTAL SETUP

As stated before, we aim to predict companies' financial failure considering it as a classification problem. Accordingly, due to the high attainment of DL algorithms regarding classification, we are comparing the performance of three different types of DL methods, i.e., DBN, MLP-6L and LSTM, and five well-known (and very effective) ensemble classification methods, i.e., RF, SVM, KNN, AdaBoost and XGBoost; looking for the best classifier with respect to the problem addressed.

The methods are compared by considering a set of evaluation metrics rather than just computing the usual 'accuracy' measure, due to the high imbalance existing in the datasets. These are described in Subsection VI-A.

In this regard, and previously to the application of the algorithms, it is mandatory to address the existing data inconsistency problem, in order to improve the classification performance. To this end, several advanced data resampling techniques have been applied to balance the datasets considered. They aim to enhance the classification performance by increasing the minority class instances and decreasing the majority class instances (depending on each balancing technique procedure).

Table 2 presents the majority and the minority classes' distribution in the three datasets before and after the data balancing step.

As shown in the table, the oversampling and the clustering-based techniques, i.e., SMOTE, BL-SMOTE, ADASYN, SMOTE-NC and K-Means SMOTE, generated balanced datasets that contain almost the same percentage for both classes (50% of samples of the majority class and 50% of minority class instances). ADASYN generated more minority class instances compared to the other oversampling methods, depending on the minority data distribution density. Therefore, it generates more synthetic samples in the case of low data distribution density. On the other hand, SVM-SMOTE generated less balanced datasets (32% of the minority class and 68% of the majority class instances), so, this will model a more realistic situation for the problem (with less 'artificial'

TABLE 2. The distribution of the majority and minority classes in the Spanish, the Taiwanese and the Polish companies' datasets before and after resampling using each balancing technique. Note that SMOTE-NC does not apply on the Taiwanese nor on the Polish datasets, since they have no categorical variables.

		Spanish companies' data		*Taiwanese companies' data*		*Polish companies' data*	
	Balancing techniques	Bankrupt	Solvent	Bankrupt	Solvent	Bankrupt	Solvent
	Original dataset	62	2797	220	6599	203	9797
Oversampling	SMOTE	2797	2797	6599	6599	9797	9797
	BL-SMOTE	2797	2797	6599	6599	9797	9797
	SVM-SMOTE	1346	2797	2308	6599	4708	9797
	ADASYN	2806	2797	6523	6599	9860	9797
	SMOTE-NC	2797	2797	-	-	-	-
Oversampling-undersampling	SMOTE-Tomek	2765	2765	6565	6565	9794	9794
	SMOTE-ENN	2651	2494	6260	5433	9757	8546
Clustering based Oversampling	K-means SMOTE	2797	2797	6599	6599	9797	9797

information added), this could have a negative effect on the results obtained by the classifiers.

The hybrid oversampling-undersampling techniques, i.e., SMOTE-ENN and SMOTE-Tomek, first oversample the minority class to achieve a very close amount to the majority class instances, then the majority class instances are undersampled depending on their cleaning procedure: ENN and Tomek Links. Thus, they generated datasets with variant balancing proportions, even turning the ratio to have more samples of the minority class than those of the majority class.

Specifically, the number of Tomek links in the Spanish companies' dataset (after SMOTE oversampling) was 32, so, after removing those links, the generated balanced dataset contains 2765 instances for each class. In the Taiwanese dataset the number of links was 34, thus, the generated dataset after balancing contains 6565 records for each class, whereas in the Polish dataset the number of links was just three, so, the data distribution among classes remains almost the same. On the other hand, SMOTE-ENN conducts a deeper data cleaning than SMOTE-Tomek. So, after oversampling, it removes the instances belonging to different classes when compared with at least two of their neighbors, that is why there is a considerable difference in the final data distribution.

Moreover, SMOTE-NC is devoted to balancing datasets including nominal and continuous attributes, thus, given the data types in the Spanish, Taiwanese and Polish companies' datasets, SMOTE-NC has only been used to balance the Spanish one, since it contains a mix of categorical and numerical variables, whereas all the variables in the other datasets are numerical.

A. EVALUATION METRICS

Several metrics can be used to measure the performance of any classifier, computed by combining the results obtained in the confusion matrix (see Table 3) [57]. Four categories are composing this matrix:

1) *True Positives (TP):* amount of samples correctly classified as bankrupt.
2) *False Positives (FP):* amount of samples incorrectly classified as bankrupt.
3) *True Negatives (TN):* amount of samples correctly classified as solvent.

TABLE 3. The confusion matrix in an output that describes the performance of each classifier, showing the distribution of correctly classified and misclassified patterns. True Positives (TP), True Negatives (TN), False Positive (FP) and False Negatives (FN) are shown.

		Actual	
		Bankrupt	Solvent
Classified	Bankrupt	TP	FN
	Solvent	FP	TN

4) *False Negatives (FN):* amount of samples incorrectly classified as solvent.

Thus, since the binary classification accuracy results are not reliable while the data considered is extremely imbalanced (the classifiers always tend to predict the majority class and ignore the minority class), several metrics have been computed to make a better judgment about each classifier's performance and reliability. These metrics are:

- Accuracy [58]: Performance of the classifier in terms of assigning the correct class to each instance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

- Recall (Sensitivity) [59]: Performance of the classifier regarding assigning each sample/company to the 'bankrupt' class (prediction) while it is actually bankrupt (real status).

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

- Specificity [59]: Represents the performance of the classifier assigning every company to the 'solvent' class (prediction) while it is actually solvent (real status).

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

- Precision [59]: Performance of the classifier regarding the ratio of assigning the correct class to each sample.

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

- Type I error [58]: Also known as False Positive Rate (FPR). It represents the failure of the classifier to assign

bankrupt companies to the 'bankrupt' class (wrong prediction), while its actual class is 'bankrupt' (real status).

$$Type\ I\ error = \frac{FP}{TN + FP} = 1 - Specificity \quad (17)$$

- Type II error [58]: Also known as False Negative Rate (FNR), represents the failure of the classifier in assigning solvent companies to 'solvent' class (wrong prediction), while its actual class is 'solvent' (real status).

$$Type\ II\ error = \frac{FN}{TP + FN} = 1 - Recall \quad (18)$$

Moreover, while the aim of this study is to predict the companies' failure, *recall* and *Type II error* are the most important metrics; they evaluate the performance of the classifiers regarding classifying and misclassifying the bankrupt companies. On the other hand, wholly focusing on the prediction of bankruptcy does not describe the real effectiveness of the classifiers, thus, the remaining metrics are important also to evaluate their overall performance.

B. IMPLEMENTATION DETAILS

In order to implement the DL methods in this study, we used *tensorflow* [51], which is a software library introduced by Google mainly for DL processing. Also, *tensorflow* has been used as a back end for *Keras* [51], which is an open-source library framework written in Python adopted to implement some DL methods. On the other hand, RF, ensemble SVM, ensemble KNN and AdaBoost have been applied using *scikit-learn* module [60]; it is a package in *Python* programming language comprising implementations of several supervised and unsupervised ML algorithms. Furthermore, XGBoost algorithm has been implemented using *xgboost* [2] package.

To achieve the highest performance and obtain more reliable results from each classifier used in this study, 10-fold cross-validation has been considered in the datasets. That is, the data are split randomly into 10 partitions, nine partitions are used to train the classifier in each iteration, and the remaining partition is used to test the obtained model. Thus, 10 iterations are performed, using in each of them a different test partition [61].

C. MAIN HYPERPARAMETERS CONSIDERED

Generally, the performance of any DL or ensemble method is tightly linked to the appropriate selection of its hyperparameters values, since some of them have a big effect on the method learning behavior. Selecting the ideal values of the hyperparameters highly depend on the problem to solve, so, they are normally set after an exhaustive experimentation process. Thus, some of the hyperparameters considered in the methods are:

1) *Learning Rate:* it is the most important DL method hyperparameter used to adapt the model to the problem. It is in the range [0.0, 1.0]. Selecting an appropriate value to obtain the optimum results is critical: a large learning rate leads the model to suboptimal solutions, whereas a too small learning rate extremely increases the probability of model stagnation.
2) *Epoch:* is a complete learning cycle of the training data in the model.
3) *Batch Size:* normally used in artificial neural networks (NNs), is the number of instances that pass through the NN in each epoch, and used to train the model before updating its internal configuration (the weights).
4) *Sigmoid Activation Function:* is a popular function associated with each neuron in the neural networks, adopted to transform its input data between 0.0 and 1.0 (model output).
5) *Dropout:* is a regulation technique able to improve the accuracy (reducing the overfitting problem) by 'removing' some visible or hidden neurons [62].
6) *Softmax Activation Function:* it is a function that transforms the outputs in a vector of probabilities, the sum of these probabilities must be up to one. In addition, it improves the accuracy of the classifiers by increasing the probability of the selected class at the expense of the others. Equation 19 presents the *softmax* formula [63]

$$f(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)} \quad (19)$$

where x_i is an input vector, x_j is an output vector and *exp* is a standard exponential function.
7) *Categorical Cross-Entropy Loss Function:* it is a function that reduces the wrong predictions by calculating the difference between the classes probabilities generated by the softmax activation function and the desired output probability (0,1), this allows the model to minimize its deficiencies by adjusting the weights.
8) *Adam Optimizer:* is a fast gradient descent optimization algorithm [64] applied in deep NNs. It also guarantees more accurate results by updating the model weights and learning rate during the training.
9) *Rectified Linear Unit (ReLU) Activation Function:* it is a fast nearly linear function associated normally with each hidden layer or with the output layer for certain types of applications, it is used to solve the vanishing gradient problem during the backpropagation learning by returning 0 if the input value is less than 0 and the same value if it is 0 or more, equation 20 shows the formula of ReLU [63].

$$f(x) = max(0, x) = \begin{cases} x_i, & if\ x_i \geq 0 \\ 0, & if\ x_i < 0 \end{cases} \quad (20)$$

10) *Kernel Function:* It is a mathematical operator used to transform the data into a required form in order to avoid some complex calculations in the classifiers, and to improve their performance. Several kernel functions could be used to enhance the performance of SVM such as Linear, Polynomial, Sigmoid and Radial Based

[2] https://xgboost.readthedocs.io/en/latest/

Function (RBF) depending on the data distribution [53]. In our work, we have used RBF described by the following formula:

$$K(x, x_i) = exp(-\gamma ||x - x_i||^2), \quad \gamma > 0 \qquad (21)$$

where x and x_i are input samples, and γ is a kernel function parameter.

11) *Number of Estimators (N_estimators):* defines the number of base estimators constructing the ensemble model.
12) *Number of Samples (Max_Samples):* is the number of samples to select from the dataset to train in each estimator, it could be set as a float number in the range of 0.0 to 1.0 (representing a percentage amount of samples), or an integer value in the range of 1 to the total number of samples in the dataset (representing the exact number of samples to fit in each ensemble's estimator).
13) *Number of Features (Max_Features):* is the number of features to select from the dataset to train in each estimator, also it could be set as a float number in the range of 0.0 to 1.0 (representing a percentage amount of features), or an integer value in the range of 1 to the total number of features in the dataset.
14) *Nearest Neighbors Parameter (K):* it is a parameter associated with the KNN method, representing the number of nearest neighbors to select for each sample in the dataset.
15) *BallTree Algorithm:* is an algorithm used to search for the nearest neighbors; it organizes the data according to distance metric values defined using two points [65].
16) *Minkowski Distance:* is a metric that measures the similarity (distance) between two points in the data according to the following formula:

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^{\frac{1}{p}} \right)^p, \qquad (22)$$

where x_i and y_i are data points, and p is an integer. The value of p could make Minkowski metric equal to other metrics, if $p = 1$ (City block distance), $p = 2$ (Euclidean distance) and $p = \infty$ (Chebyshev distance) [66].
17) *Entropy Criterion:* is an impurity measure used to evaluate the quality of the data splitting during building the decision trees, less entropy (impurity) more information gain (pure data).

It is important to note that not all these parameters are considered in every DL and ensemble method applied in our study. Tables 4 and 5 enumerate the usage and considered value for each hyperparameter in the DL and in the Ensemble methods respectively. In the following Section, more details about the selection of each hyperparameter values are given.

VII. RESULTS

As aforementioned, in binary classification problems, the performance and the reliability of any classifier are dramatically affected by the data balancing ratio. Furthermore, in case the data is extremely imbalanced, the classifier tends always to predict the majority class and somehow 'ignore' the minority class. This behavior gains significant results with respect to the accuracy, but it yields poor performance regarding the prediction of the minority class. In other words, the model obtains high accuracy at the expense of the reliability [4], [10].

Therefore, due to the sensitivity of the DL methods to the data distribution inconsistency, eight advanced balancing techniques have been previously applied in this study.

Thus, different datasets have been generated from the Spanish, Taiwanese and Polish datasets - one per each balancing method -. Therefore, every classification algorithm has been tested considering each of the produced datasets.

The following sections present and discuss the results obtained, analyzing specifically the behavior of each DL and ensemble method regarding bankruptcy and solvency misclassification, as well as the effects of each balancing technique on their outcomes.

A. DBN + BALANCING TECHNIQUES

DBN with 10-folds cross-validation was applied to each dataset generated using the balancing techniques. Thus, after several trials of DBN application on the three datasets, we found that the most appropriate *RBM* and fine-tuning learning rates must be set respectively to 0.005 and 0.1 (See Table 4), also the optimum value of batch size is 40. The activation function for DBN obtaining the best results in our work is *sigmoid*, while *dropout* was set to 0.2.

The results obtained by DBN applied to the balanced datasets are shown in Table 6. These are approximately divergent according to the metrics used in this study to evaluate the performance of all DL and ensemble methods. As it can be seen, DBN shows higher performance with the simplest data (i.e., Spanish dataset) than the other dataset used in this study; more complex data leads to less *accuracy*, *recall* and *precision* metrics values. The results on the Spanish dataset are better overall than the results on the Taiwanese and Polish data, given that the first dataset is 'the simplest' one. For the Polish companies' dataset (the most complex), DBN moderately misses predicting solvent and bankrupt companies as stated in *recall* and *specificity* metrics. Looking at the Spanish companies' dataset, the use of SMOTE-ENN yields the best results in combination with DBN regarding *accuracy*, *recall* and *type II error*, also getting this approach the second-best *precision* value. On the other hand, SVM-SMOTE obtained the complementary best results, i.e., the best *specificity*, and *type I error*. Moreover, ADASYN obtains the best *precision*, also gets the second-best in the remaining metrics. In other words, SMOTE-ENN shows the optimum performance in the classification of bankrupt companies, whereas SVM-SMOTE is the best in classifying solvent ones in this dataset.

With respect to the Taiwanese companies' dataset, again the combination of SMOTE-ENN and DBN yields the lowest bankruptcy misprediction. It obtains the best *recall* and

TABLE 4. Hyperparameters considered for each DL method, '✓' or a value indicates the use of the hyperparameter, and '–' indicates no use.

DL methods	Learning rate	Adam optimizer	Batch size	Epochs	Sigmoid	Softmax	ReLU	Categorical cross-entropy	Dropout
DBN	0.005, 0.1	-	40	150	✓	-	-	-	0.2
LSTM	-	✓	15	200	-	✓	-	✓	-
MLP-6L	-	✓	15	200	-	✓	✓	✓	0.2

TABLE 5. Hyperparameters considered for each ensemble method, '✓' or a value indicates the use of the hyperparameter, and '–' indicates no use.

Ensemble methods	Kernel Function	N_estimators	Max_Samples	Max_Features	K	BallTree algorithm	Minkowski distance	Entropy
RF	-	100	1.0	1.0	-	-	-	✓
SVM	RBF	40	1.0	1.0	-	-	-	-
KNN	-	40	1.0	1.0	5	✓	✓	-
AdaBoost	-	50	1.0	1.0	-	-	-	✓
XGBoost	-	50	1.0	1.0	-	-	-	✓

TABLE 6. The evaluation metrics values yielded by DBN applied to balanced datasets generated by each balancing technique. The first and second-best results for each metric are marked in boldface, the best value also has gray background.

Spanish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9174	0.9640	0.8707	0.8819	0.1293	0.0360
BL-SMOTE	0.9175	0.9462	0.8888	0.8953	0.1112	0.0538
SVM-SMOTE	0.9095	0.9094	**0.9096**	0.8516	**0.0904**	0.0906
ADASYN	**0.9409**	**0.9783**	**0.9011**	**0.9136**	**0.0989**	**0.0217**
SMOTE-NC	0.9232	0.9615	0.8849	0.8934	0.1151	0.0385
SMOTE-Tomek	0.9180	0.9677	0.8683	0.8806	0.1317	0.0323
SMOTE-ENN	**0.9414**	**0.9854**	0.8946	**0.9092**	0.1054	**0.0146**
K-means SMOTE	0.9080	0.9687	0.8472	0.8638	0.1528	0.0313
Taiwanese companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.8538	0.8861	0.8216	0.8339	0.1784	0.1139
BL-SMOTE	0.8728	**0.9204**	0.8251	0.8414	0.1749	**0.0796**
SVM-SMOTE	0.8714	0.8673	**0.8729**	0.7062	**0.1271**	0.1327
ADASYN	0.8543	0.8532	**0.8555**	**0.8578**	**0.1445**	0.1468
SMOTE-Tomek	**0.8821**	0.9173	0.8468	**0.8589**	0.1532	0.0827
SMOTE-ENN	**0.8809**	**0.9465**	0.8153	0.8367	0.1847	**0.0535**
K-means SMOTE	0.8485	0.8742	0.8227	0.8314	0.1773	0.1258
Polish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.7557	**0.8398**	0.6716	0.7191	0.3284	**0.1602**
BL-SMOTE	0.7638	0.7909	0.7368	**0.7536**	0.2632	0.2091
SVM-SMOTE	**0.8075**	0.6654	**0.8758**	0.7290	**0.1242**	0.3346
ADASYN	0.7353	0.7732	0.6973	0.7217	0.3027	0.2268
SMOTE-Tomek	0.7434	0.8230	0.6637	0.7103	0.3363	0.1770
SMOTE-ENN	**0.8151**	**0.8531**	**0.7718**	**0.8115**	**0.2282**	**0.1469**
K-means SMOTE	0.7463	0.7868	0.7057	0.7287	0.2943	0.2132

type II error, while BL-SMOTE is the second-best technique according to the same two metrics. On the other hand, SVM-SMOTE yields the lowest solvency misprediction as stated in *specificity* and *type I error*, being ADASYN the second-best technique. SMOTE-Tomek can be also remarked, as it obtains the best *accuracy* and *precision*, but very close to SMOTE-ENN and ADASYN methods, respectively.

Regarding the Polish companies' dataset, SMOTE-ENN leads DBN to obtain the optimum *accuracy*, *recall*, *precision* and *type II error*, and the second-best *specificity* and *type I error*. Thus, this is the best approach overall. However, again SVM-SMOTE obtains the best results according to the *specificity* and *type I error*.

Generally, the performance of any ML algorithm depends mainly on the data fitting. Over-fitting yields excellent training results but poor validation ones, whereas under-fitting obtains poor training and validation results [67]. Good-fitting yields relatively close training and validation values. So, discovering if the DBN model learns adequately is quite tricky [68]; since the model's input is just training data. Figure VII-A illustrates the training loss in the DBN fine-tuning stage. It specifically shows the loss of the superior combination of DBN with balancing techniques with respect to bankruptcy prediction on the three datasets (i.e., DBN + SMOTE-ENN).

The loss of the DBN on the Spanish and Polish data starts from 0.6, and decreases until epoch 100, then they are stabilized at around 0.13 and 0.38, respectively. The loss on the Polish data starts at around 0.5 and stabilizes after epoch 40 at around 0.22 of loss value. From these curves we can conclude that the model is not over-fitted, given that the model they

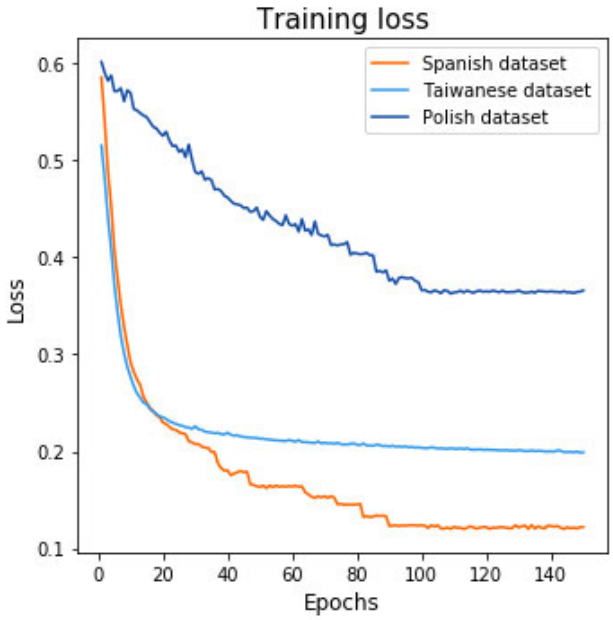

FIGURE 5. Training loss curves in the DBN fine-tuning stage obtained by the best combination of DBN with a balancing technique (DBN + SMOTE-ENN) in the three datasets.

show is not training perfectly (loss values are not very low). On the other hand, while the input of DBN model is just training data (no validation curves) it is hard to determine if the model was under-fitted. To solve this issue, we have extracted 20 subsets of different sizes from each balanced dataset generated in this study. For each subset, DBN was trained and tested using 10-fold cross-validation to trace the model's performance with respect to the training and validation data sizes. Thus, Figure 6 illustrates the DBN model's training and validation curves obtained using the best data balancing technique for predicting bankrupt companies (i.e., SMOTE-ENN). As it is shown in the figure, each training curve is relatively close to the validation curve (both belong to the same dataset), which shows good-fitting of the Spanish, Taiwanese and Polish data to the DBN model.

Thus, the firm conclusion extracted from the DBN results is that utilizing SMOTE-ENN to balance the datasets helps the DL method to reach the lowest bankruptcy misclassification, which is the most relevant prediction in the problem we are solving.

B. LSTM + BALANCING TECHNIQUES

In this experiment, we applied 4 stacked layers of LSTM on the datasets generated by the balancing techniques in order to improve the accuracy of the model outcomes. Moreover, we have considered the *softmax* activation function in the output layer, so we previously transformed the class attribute into two attributes, since using this function gets extremely better results than considering *sigmoid* one. In addition, considering that the class attribute transformed into two attributes, *categorical cross − entropy* applied as a loss function. Moreover, after intensive experimentation on LSTM model with the three datasets, setting the batch size to 15 yielded the best results.

Besides, *Adam optimizer* was used to reduce the loss and to obtain more accurate outcomes by updating the model weights and learning rate during the training phase.

Looking at the obtained results presented in Table 7, it can be seen the substantial improvement compared to the previous experiment with regard to the Polish companies' dataset. Thus, applying SMOTE-ENN previously to LSTM, leads to outperforming the rest of the balancing techniques with the hardest dataset. This combination obtains the best values for all the metrics, so, it is the most adequate for classifying both bankrupt and solvent companies in this dataset. SMOTE-Tomek gets very close results for *accuracy*, *recall*, and *type II error* metrics (related to bankruptcy prediction), whereas BL-SMOTE is the second-best regarding *specificity*, *precision*, and *type I error* (focused on healthy companies prediction).

In the case of the Spanish companies' dataset, the roles are inverted, and SMOTE-Tomek together with LSTM yielded the best outcomes according to all the metrics. This time K-Means SMOTE is the second-best also in all the metrics, however, SMOTE-ENN gets very close results to these approaches, even reaching a perfect value for *recall* and *type II error*, as it is also obtained by the two aforementioned methods.

Regarding the Taiwanese companies' dataset, the combination of ADASYN with LSTM shows perfect performance regarding predicting the bankrupt companies according to *recall* and *type II error*, whereas SMOTE obtains the second-best results in the same case. On the other hand, the combination of K-means SMOTE with LSTM yields the best results in the metrics related to the prediction of the solvent companies, while BL-SMOTE is the second-best. SMOTE-ENN obtains the best *accuracy* and *precision* values.

Moreover, differently than in the previous experiment, LSTM inputs are training and validation data. Figure 7 illustrates the curves of LSTM training and validation accuracy/loss in each epoch, obtained from the best combinations in bankruptcy prediction in the three datasets. The best fusions of LSTM are with SMOTE-Tomek, ADASYN and SMOTE-ENN. Thus, the training shows higher accuracy and lower loss than the validation in all curves, but very close values showing good-fitting of the LSTM model to the Spanish, Taiwanese and Polish data.

It is important to note that LSTM results are much better than DBN ones, obtaining great indicators for all the metrics in almost all the cases (all the balancing techniques) and in the three datasets. Thus, for every balanced dataset, LSTM tends to predict the correct status of the companies accurately whether they are bankrupt or solvent, with values mostly above 0.98 and errors close to 0 for almost all the approaches.

C. MLP-6L + BALANCING TECHNIQUES

Following the process of previous experiments, a deep learning MLP approach with 6 layers, called MLP-6L, has been

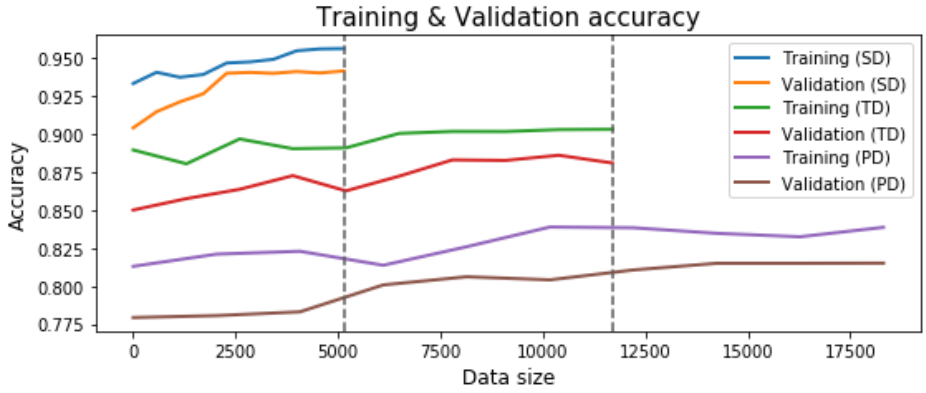

FIGURE 6. Training and validation accuracy curves obtained by the superior combination of DBN with a balancing technique (i.e., DBN + SMOTE-ENN). The dashed vertical lines indicate the size of the Spanish dataset (SD), Taiwanese dataset (TD) and Polish dataset (PD).

TABLE 7. Evaluation metrics values yielded by LSTM applied to the different datasets generated after applying each balancing technique. First and second-best results for each metric are marked in boldface, the best values also have a gray background.

Spanish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9929	0.9996	0.9861	0.9863	0.0139	0.0004
BL-SMOTE	0.9925	0.9982	0.9868	0.9869	0.0132	0.0018
SVM-SMOTE	0.9923	0.9964	0.9882	0.9884	0.0118	0.0036
ADASYN	0.9939	**1.0**	0.9878	0.9881	0.0122	**0.0**
SMOTE-NC	0.9887	0.9925	0.9850	0.9852	0.0150	0.0075
SMOTE-Tomek	**0.9955**	**1.0**	**0.9908**	**0.9914**	**0.0092**	**0.0**
SMOTE-ENN	0.9942	**1.0**	0.9880	0.9889	0.0120	**0.0**
K-means SMOTE	**0.9948**	**1.0**	**0.9896**	**0.9898**	**0.0104**	**0.0**
Taiwanese companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9891	**0.9998**	0.9783	0.9788	0.0217	**0.0002**
BL-SMOTE	0.9884	0.9955	**0.9814**	0.9816	**0.0186**	0.0045
SVM-SMOTE	0.9788	0.9809	0.9780	0.9399	0.0220	0.0191
ADASYN	**0.9897**	**1.0**	0.9795	0.9797	0.0205	**0.0**
SMOTE-Tomek	0.9879	0.9995	0.9762	0.9768	0.0238	0.0005
SMOTE-ENN	**0.9905**	0.9997	0.9799	**0.9829**	0.0201	0.0003
K-means SMOTE	0.9789	0.9758	**0.9821**	**0.9820**	**0.0179**	0.0242
Polish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9896	**0.9985**	0.9808	0.9812	0.0192	**0.0015**
BL-SMOTE	0.9891	0.9921	**0.9860**	**0.9861**	**0.0140**	0.0079
SVM-SMOTE	0.9851	0.9847	0.9700	0.9700	0.0300	0.0153
ADASYN	0.9898	0.9969	0.9827	0.9830	0.0173	0.0031
SMOTE-Tomek	**0.9904**	**0.9985**	0.9823	0.9826	0.0177	**0.0015**
SMOTE-ENN	**0.9931**	**0.9988**	**0.9866**	**0.9884**	**0.0134**	**0.0012**
K-means SMOTE	0.9902	0.9955	0.9849	0.9851	0.0151	0.0045

applied to the eight datasets built after applying each one of the balancing methods. In this case, after exhaustive experimentation on the MLP-6L model, setting the batch size to 15, and using *Adam optimizer* yielded the best results. Again, in this experiment, we have transformed the class attribute into two attributes in order to use *softmax* in the output layer, as it was compared with *sigmoid* in preliminary tests yielding much better outputs. However, in the hidden layers, we have considered *Rectified Linear Unit (ReLU)* as the activation function. In addition, because the class attribute was transformed into two attributes, *categorical cross − entropy* was used as loss function.

As it can be seen in Table 8, the results obtained by the MLP-6L model applied on all the generated datasets are excellent. Being precise, SMOTE-ENN obtains the best outcomes for almost all the metrics in all the datasets, just getting the second-best for solvent-related metrics (*specificity* and *type I error*) in the Taiwanese and Polish companies' datasets.

In summary, SMOTE-ENN leads the MLP-6L model to obtain the highest *accuracy* and lowest bankrupt companies misclassification rate according to *recall* and *type II error* metrics. Also, the combination of MLP-6L with SMOTE-ENN obtains the second-best solvent companies classification and misclassification rates as stated in

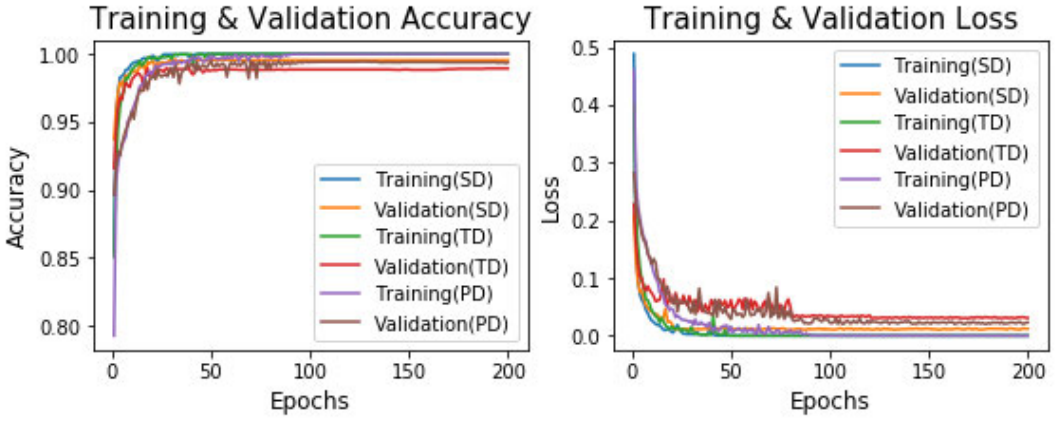

FIGURE 7. Training and validation accuracy/loss curves were obtained by the superior combination of LSTM with a balancing technique in the three datasets; LSTM + SMOTE-Tomek in the Spanish dataset (SD), LSTM + ADASYN in the Taiwanese dataset (TD) and LSTM + SMOTE-ENN in the Polish dataset (PD).

TABLE 8. Evaluation metrics values yielded by MLP-6L applied to balanced datasets generated by each balancing technique. First and second-best results for each metric are marked in boldface, the best value also has gray background.

Spanish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	**0.9980**	**1.0**	**0.9961**	**0.9961**	**0.0039**	**0.0**
BL-SMOTE	0.9959	0.9975	0.9943	0.9943	0.0057	0.0025
SVM-SMOTE	0.9952	0.9963	0.9946	0.9890	0.0054	0.0037
ADASYN	0.9975	**1.0**	0.9950	0.9951	0.0050	**0.0**
SMOTE-NC	0.9936	0.9939	0.9932	0.9933	0.0068	0.0061
SMOTE-Tomek	0.9973	**1.0**	0.9944	0.9948	0.0056	**0.0**
SMOTE-ENN	**0.9986**	**1.0**	**0.9972**	**0.9974**	**0.0028**	**0.0**
K-means SMOTE	0.9964	0.9989	0.9939	0.9940	0.0061	0.0011
Taiwanese companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9910	**1.0**	0.9820	0.9823	0.0180	**0.0**
BL-SMOTE	0.9900	0.9939	0.9861	0.9862	0.0139	0.0061
SVM-SMOTE	0.9839	0.9753	0.9870	0.9633	0.0130	0.0247
ADASYN	0.9905	**1.0**	0.9813	0.9813	0.0187	**0.0**
SMOTE-Tomek	**0.9912**	0.9998	0.9826	0.9830	0.0174	0.0002
SMOTE-ENN	**0.9963**	**1.0**	**0.9921**	**0.9931**	**0.0079**	**0.00**
K-means SMOTE	0.9832	0.9736	**0.9927**	**0.9926**	**0.0073**	0.0264
Polish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9945	0.9991	0.9899	0.9900	0.0101	0.0009
BL-SMOTE	0.9936	0.9950	0.9922	**0.9923**	0.0078	0.0050
SVM-SMOTE	0.9921	0.9911	**0.9932**	0.9859	**0.0068**	0.0089
ADASYN	**0.9953**	**0.9996**	0.9909	0.9911	0.0091	**0.0004**
SMOTE-Tomek	0.9947	0.9995	0.9899	0.9900	0.0101	0.0005
SMOTE-ENN	**0.9967**	**0.9999**	**0.9930**	**0.9939**	**0.0070**	**0.0001**
K-means SMOTE	0.9915	0.9994	0.9826	0.9850	0.0174	0.0006

specificity and *type I error*. Again, it can be seen the complexity of each dataset, since four approaches are able to obtain the highest scores (1.0) for some metrics and the lowest errors (0) with the Spanish dataset, and three approaches with Taiwanese dataset, whereas they are close to the maximum and minimum values, but did not reach them with the Polish data.

Furthermore, Figure 8 illustrates the curves of training and validation accuracy/loss obtained from the best MLP-6L combinations in predicting the companies failure (MLP-6L + SMOTE-ENN) in the three datasets. As it is shown in the figure, the validation accuracy and loss are better than the training ones. On the other hand, the validation loss shows relatively better performance than in the previous experiment. Also in this experiment, the training and validation accuracy/loss curves are very close, which shows good-fitting of the three datasets to the MLP-6L model.

D. RF + BALANCING TECHNIQUES

This section presents the results on the application of RF, which is a very effective classifier based on ensembles. As previously told, this is not a DL approach, but it has been extensively used in many difficult classification problems in the literature and obtained excellent results. Thus, after an

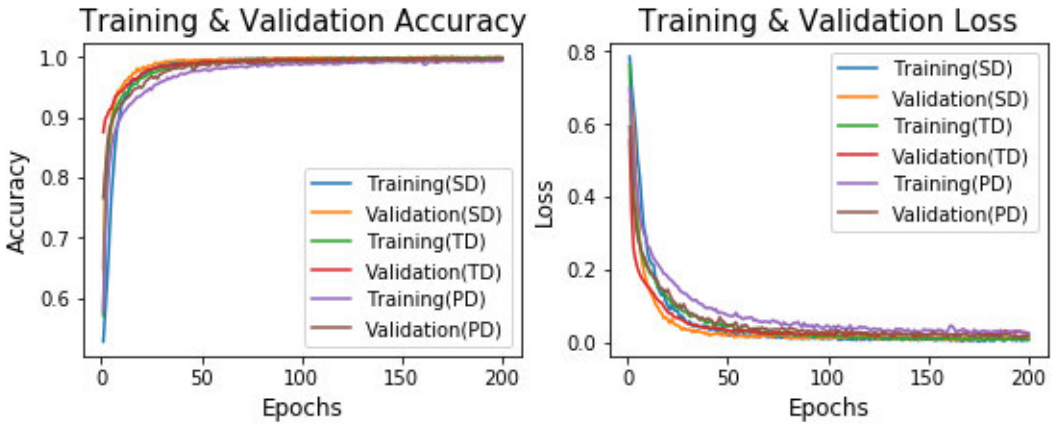

FIGURE 8. Training and validation accuracy/loss curves obtained by the superior combination of MLP-6L with balancing techniques concerning bankruptcy prediction (MLP-6L + SMOTE-ENN) in the Spanish dataset (SD), Taiwanese dataset (TD) and Polish dataset (PD).

exhaustive experimentation process with the three datasets, we found that the most appropriate number of ensemble model estimators (*N_Estimators*) is 100; a larger number causes an increment in the computation time and obtains the same results, whereas a smaller number of estimators does not obtain results as good as 100. In the other hand, to fit an entire *bootstrap* in each estimator, *Max_features* and *Max_Samples* are set to 1.0. Besides, the trees splitting criterion is *entropy* while it yields better results than the other criteria in this work. Here we have combined this method with the same eight balancing techniques as the DL approaches and tested it on the same three datasets.

Table 9 shows the obtained results for RF in conjunction with the data balancing methods. Thus, the combination of the classifier with SMOTE-ENN outperforms the rest of the approaches concerning the prediction of the bankrupt companies in all the datasets, getting the best *recall* and *type II error*, and also it obtains the best *accuracy* and the second-best values for the remaining metrics in the Spanish and Taiwanese datasets. On the other hand, the combination of the classifier and K-means SMOTE yields the best results in predicting the solvent companies in all the datasets, obtaining the best *specificity*, *precision* and *type I error*.

Furthermore, to trace the performance of the RF model with respect to the training and validation data sizes, the same procedure addressed in experiment (VII-A) has been conducted in this experiment as well. In other words, also in this experiment, from each balanced dataset generated in this study, 20 subsets of different sizes were extracted, and then each subset was adopted to train and test RF using 10-fold cross-validation. Figure 9 shows the training and validation accuracy curves obtained by the best approaches regarding bankruptcy prediction in the three datasets (RF + SMOTE-ENN). The best three combinations' training accuracy is equal to 1.0, whereas each validation accuracy curve is increasing being close to the training curves and showing that the three datasets good-fitted to RF.

E. ENSEMBLE SVM + BALANCING TECHNIQUES

In this experiment, we have applied SVM to the generated datasets after balancing. Thus, in order to improve the performance of this algorithm, we propose it as an ensemble model using the bagging technique. The SVM kernel function that we have used is *RBF* because it yielded better performance compared to the other well-known kernel functions (*Linear* and *Sigmoid*) in our problem. Moreover, the most appropriate number of estimators (*N_Estimators*) is 40, whereas the *Max_features* and *Max_Samples* are set as in the previous experiment.

As it can be seen in Table 10, the combination of the ensemble SVM and SMOTE-ENN yields the best performance regarding predicting and mispredicting bankrupt companies in the three datasets according to *recall* and *type II error* values, whereas applying K-means SMOTE obtains the second-best values for the same metrics in the Spanish data, and the application of BL-SMOTE to the Taiwanese and Polish companies' datasets yields the second-best values as well. On the other hand, K-means SMOTE shows the best performance with respect to predicting and mispredicting the solvent companies in all datasets. It obtains the best values of *specificity*, *precision* and *type I error*. In addition, K-means SMOTE obtains the highest *accuracy* values in the Spanish and Polish companies' datasets and the second-best for the Taiwanese one, whereas ADASYN obtains the best *accuracy* value on the Taiwanese data. In addition, also in this experiment, the same procedure addressed in the previous experiment used to trace the performance of the ensemble SVM model.

Figure 10 illustrates the training and validation curves obtained by the best approaches in predicting bankrupt companies in the three datasets (i.e., SVM + SMOTE-ENN). Thus, different from the previous experiments, the training and validation values increase along with each other simultaneously; the validation curves are very close to the training ones for each approach, which makes us conclude

TABLE 9. Evaluation metrics values yielded by RF applied to balanced datasets generated by each balancing technique. First and second-best results for each metric are marked in boldface, the best value also has gray background.

Spanish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9968	0.9979	0.9957	0.9957	0.0043	0.0021
BL-SMOTE	0.9964	0.9961	**0.9968**	0.9968	**0.0032**	0.0039
SVM-SMOTE	0.9947	0.9930	0.9957	0.9924	0.0043	0.0070
ADASYN	**0.9971**	**0.9989**	0.9952	0.9955	0.0048	**0.0011**
SMOTE-NC	0.9955	0.9982	0.9928	0.9929	0.0072	0.0018
SMOTE-Tomek	0.9967	0.9986	0.9949	0.9950	0.0051	0.0014
SMOTE-ENN	**0.9979**	**0.9989**	**0.9968**	**0.9970**	**0.0032**	**0.0011**
K-means SMOTE	0.9921	0.9850	**0.9993**	**0.9993**	**0.0007**	0.0150
Taiwanese companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9796	0.9950	0.9642	0.9653	0.0358	0.0050
BL-SMOTE	0.9841	0.9918	0.9764	0.9768	0.0236	0.0082
SVM-SMOTE	0.9745	0.9584	0.9801	0.9443	0.0199	0.0416
ADASYN	0.9796	0.9966	0.9627	0.9636	0.0373	0.0034
SMOTE-Tomek	0.9807	**0.9968**	0.9647	0.9658	0.0353	**0.0032**
SMOTE-ENN	**0.9906**	**0.9970**	**0.9834**	**0.9857**	**0.0166**	**0.0030**
K-means SMOTE	**0.9844**	0.9724	**0.9964**	**0.9963**	**0.0036**	0.0276
Polish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9932	0.9984	0.9880	0.9881	0.0120	0.0016
BL-SMOTE	**0.9935**	0.9949	0.9920	**0.9921**	0.0080	0.0051
SVM-SMOTE	0.9919	0.9858	**0.9948**	0.9891	**0.0052**	0.0142
ADASYN	0.9930	0.9986	0.9874	0.9877	0.0126	0.0014
SMOTE-Tomek	**0.9933**	**0.9988**	0.9878	0.9879	0.0122	**0.0012**
SMOTE-ENN	0.9926	**0.9992**	0.9852	0.9871	0.0148	**0.0008**
K-means SMOTE	0.9892	0.9797	**0.9988**	**0.9988**	**0.0012**	0.0203

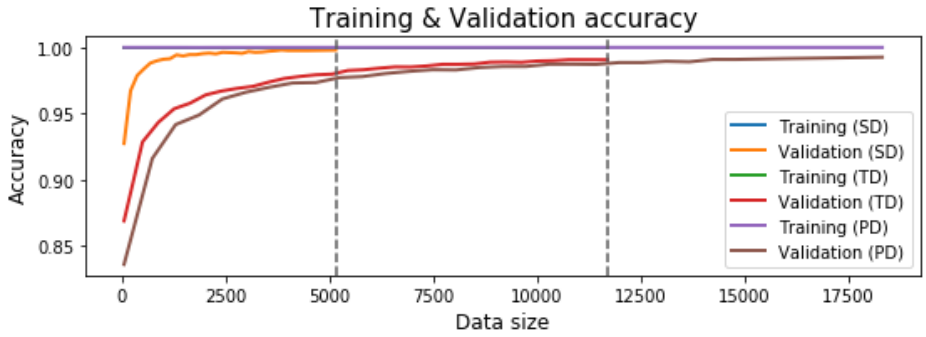

FIGURE 9. Training and validation accuracy curves obtained by the superior combination of RF with a balancing technique (RF + SMOTE-ENN). The dashed vertical lines indicate the size of the Spanish dataset (SD), Taiwanese dataset (TD) and Polish dataset (PD).

that the three datasets are good-fitted by the ensemble SVM model.

F. ENSEMBLE KNN + BALANCING TECHNIQUES

This subsection presents the results obtained by the application of KNN to the balanced datasets generated by the eight balancing techniques. In this experiment, again we proposed the KNN as an ensemble model using the *bagging* technique as was the case in the previous experiment. Thus, the most proper value of K in the KNN was 5; a lower value increases the impact of the noise on the classifier's results, whereas a higher value increases the computation time and obtained worse results. In addition, *BallTree* algorithm has been used to find the nearest neighbors; it obtains better results than the other searching algorithms in our datasets. The distance metric used is *Minkowski*. Moreover, the *N_Estimators* parameter was set to 40; a larger number causes an increment in the computation time and obtains the same results. Also in this experiment, in order to fit an entire *bootstrap* in each estimator, *Max_features* and *Max_Samples* are set to 1.0.

As it is shown in Table 11, and as in the previous set of experiments, the combination of the ensemble KNN and SMOTE-ENN obtained the best *recall* and *type II error* values for the three datasets, i.e., a good bankrupt classification performance.

Paying attention to each dataset, SMOTE-ENN leads the ensemble KNN to obtain the best values of all evaluation metrics in the Spanish companies' dataset showing that it is the ideal combination in order to predict the bankrupt and solvent companies, while the combination with ADASYN

TABLE 10. Evaluation metrics values yielded by ensemble SVM applied to balanced datasets generated by each balancing technique. The first and second-best results for each metric are marked in boldface, the best value also has a gray background.

Spanish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9694	0.9796	0.9592	0.9602	0.0408	0.0204
BL-SMOTE	0.9726	0.9782	0.9671	0.9675	0.0329	0.0218
SVM-SMOTE	0.9697	0.9630	**0.9735**	0.9539	**0.0265**	0.0370
ADASYN	0.9806	0.9913	0.9691	0.9716	0.0309	0.0087
SMOTE-NC	0.9680	0.9864	0.9496	0.9515	0.0504	0.0136
SMOTE-Tomek	0.9697	0.9791	0.9603	0.9611	0.0397	0.0209
SMOTE-ENN	**0.9825**	**0.9940**	0.9703	**0.9729**	0.0297	**0.0060**
K-means SMOTE	**0.9884**	**0.9800**	**0.9968**	**0.9967**	**0.0032**	**0.0200**
Taiwanese companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9560	0.9848	0.9271	0.9311	0.0729	0.0152
BL-SMOTE	0.9639	**0.9851**	0.9426	0.9450	0.0574	**0.0149**
SVM-SMOTE	0.9498	0.9125	0.9629	0.8958	0.0371	0.0875
ADASYN	**0.9885**	0.9784	**0.9987**	**0.9986**	**0.0013**	0.0216
SMOTE-Tomek	0.9543	0.9837	0.9249	0.9291	0.0751	0.0163
SMOTE-ENN	0.9741	**0.9921**	0.9535	0.9605	0.0465	**0.0079**
K-means SMOTE	**0.9836**	0.9677	**0.9994**	**0.9994**	**0.0006**	0.0323
Polish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.8660	0.9694	0.7626	0.8034	0.2374	0.0306
BL-SMOTE	**0.9136**	**0.9824**	0.8447	**0.8636**	0.1553	**0.0176**
SVM-SMOTE	0.9056	0.9023	**0.9072**	0.8240	**0.0928**	0.0977
ADASYN	0.8629	0.9686	0.7567	0.8004	0.2433	0.0314
SMOTE-Tomek	0.8655	0.9692	0.7618	0.8029	0.2382	0.0308
SMOTE-ENN	0.9018	**0.9846**	0.8074	0.8536	0.1926	**0.0154**
K-means SMOTE	**0.9885**	0.9784	**0.9987**	**0.9986**	**0.0013**	0.0216

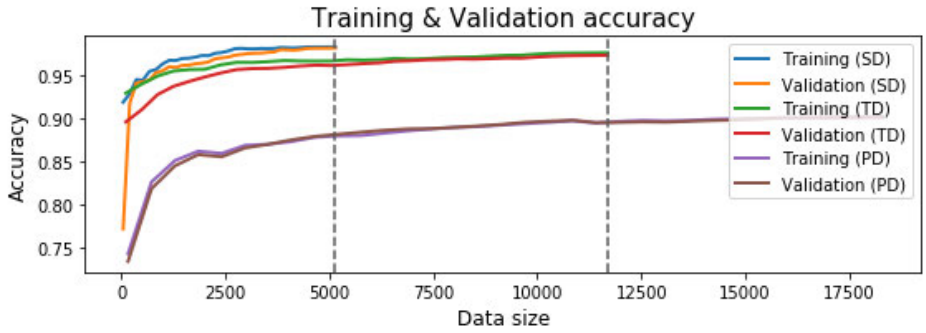

FIGURE 10. Training and validation accuracy curves obtained by the superior combination of ensemble SVM with a balancing technique (ensemble SVM + SMOTE-ENN). The dashed vertical lines indicate the size of the Spanish dataset (SD), Taiwanese dataset (TD) and Polish dataset (PD).

yields very close results proving that it could be an adequate alternative.

Regarding the Taiwanese companies' dataset, SMOTE-ENN obtains the best values of *accuracy*, *recall* and *type II error*, and also the second-best values of the remaining metrics. K-means SMOTE shows a great performance regarding predicting the solvent companies compared to the other balancing techniques, as it obtains the best *specificity*, *precision* and *type I error*, and the second-best *accuracy*.

For the Polish companies' dataset, again the combination of ensemble KNN with SMOTE-ENN yields the best results regarding predicting the bankrupt companies, and shows that it is the second-best combination to predict the solvent companies. On the other hand, K-means SMOTE this time also shows a big improvement in the prediction of solvent companies.

Furthermore, as explained in the previous experiment, Figure 11 illustrates the training and validation curves obtained from the best ensemble KNN and balancing techniques in predicting bankrupt companies in the three datasets (KNN + SMOTE-ENN). Thus, in this experiment, the training and validation values also increase along with each other simultaneously for each approach, thus, the Spanish, Taiwanese and Polish data are good-fitted to the ensemble KNN.

G. AdaBoost + BALANCING TECHNIQUES

AdaBoost is an ensemble method based on the *boosting* technique that showed better performance than bagging and

TABLE 11. Evaluation metrics values yielded by ensemble KNN applied to balanced datasets generated by each balancing technique. The first and second-best results for each metric are marked in boldface, the best value also has a gray background.

Spanish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9753	**1.0**	0.9507	0.9531	0.0493	**0.0**
BL-SMOTE	0.9793	0.9986	0.9600	0.9617	0.0400	0.0014
SVM-SMOTE	0.9704	0.9955	0.9564	0.9277	0.0436	0.0045
ADASYN	**0.9833**	**1.0**	**0.9655**	**0.9687**	**0.0345**	**0.0**
SMOTE-NC	0.9703	0.9950	0.9456	0.9484	0.0544	0.0050
SMOTE-Tomek	0.9762	**1.0**	0.9523	0.9547	0.0477	**0.0**
SMOTE-ENN	**0.9835**	**1.0**	**0.9659**	**0.9691**	**0.0341**	**0.0**
K-means SMOTE	0.9739	0.9996	0.9482	0.9508	0.0518	0.0004
Taiwanese companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9478	**1.0**	0.8956	0.9056	0.1044	**0.0**
BL-SMOTE	0.9565	0.9965	0.9165	0.9227	0.0835	0.0035
SVM-SMOTE	0.9512	0.9783	0.9417	0.8547	0.0583	0.0217
ADASYN	0.9467	0.9998	0.8941	0.9032	0.1059	0.0002
SMOTE-Tomek	0.9481	**1.0**	0.8963	0.9060	0.1037	**0.0**
SMOTE-ENN	**0.9843**	**1.0**	**0.9665**	**0.9714**	**0.0335**	**0.0**
K-means SMOTE	**0.9833**	0.9709	**0.9958**	**0.9957**	**0.0042**	0.0291
Polish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9455	**0.9997**	0.8914	0.9020	0.1086	**0.0003**
BL-SMOTE	0.9649	0.9959	0.9340	0.9379	0.0660	0.0041
SVM-SMOTE	0.9566	0.9904	0.9403	0.8886	0.0597	0.0096
ADASYN	0.9448	**0.9997**	0.8895	0.9011	0.1105	**0.0003**
SMOTE-Tomek	0.9453	0.9992	0.8915	0.9022	0.1085	0.0008
SMOTE-ENN	**0.9788**	**0.9999**	**0.9546**	**0.9618**	**0.0454**	**0.0001**
K-means SMOTE	**0.9895**	0.9798	**0.9993**	**0.9993**	**0.0007**	0.0202

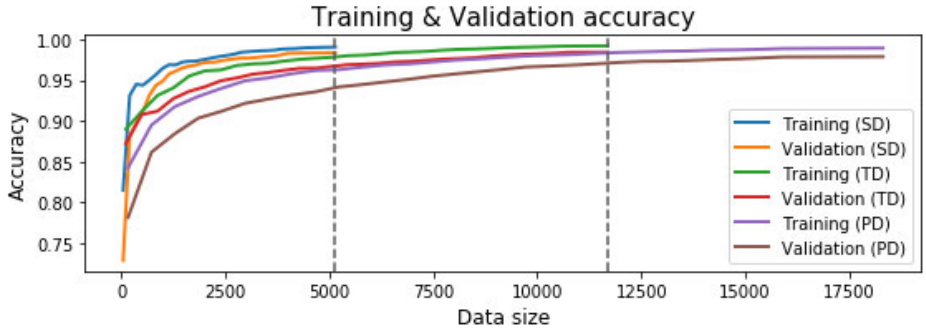

FIGURE 11. Training and testing accuracy curves obtained by the superior combination of ensemble KNN with a balancing technique (ensemble KNN + SMOTE-ENN) in the three datasets. The dashed vertical lines indicate the size of the Spanish dataset (SD), Taiwanese dataset (TD) and Polish dataset (PD).

standard classifiers in the classification using imbalanced datasets. In a previous study considered the Spanish companies' dataset [46], applying AdaBoost on the dataset without resampling or Feature Selection obtained the highest *recall* value (0.6) compared to the other classifiers. That value is still relatively low, but it significantly improved later after applying the balancing techniques. In this subsection, we present the results obtained by AdaBoost method applied to the generated datasets after balancing. Thus, in this experiment, the most appropriate number of estimators (*N_Estimators*) is 50, whereas *Max_features* and *Max_Samples* are set to 1.0 in order to fit the entire data in each estimator. In addition, *Entropy* has been used to measure the quality of data splitting during building the decision trees.

In Table 12, it can be seen the significant impact of K-means SMOTE on the AdaBoost regarding predicting the bankrupt and solvent companies in the Taiwanese and Polish datasets. It shows a considerable improvement in the metrics values compared to the rest of the balancing techniques. In the Taiwanese dataset, the second-best results in predicting bankrupt companies have been obtained by using SMOTE-ENN, and BL-SMOTE in the Polish companies' dataset.

Regarding the Spanish companies' dataset, ADASYN leads AdaBoost to the minimum bankruptcy misclassification; it obtains the best *recall* and *type II error*, and the second-best *accuracy*. On the other hand, in the Spanish dataset also, K-means SMOTE shows the optimum performance with respect to predicting the solvent companies; it

TABLE 12. Evaluation metrics values yielded by AdaBoost applied to balanced datasets generated by each balancing technique. The first and second-best results for each metric are marked in boldface, the best value also has a gray background.

Spanish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9873	0.9946	0.9800	0.9803	0.0200	0.0054
BL-SMOTE	0.9864	0.9954	0.9775	0.9779	0.0225	0.0046
SVM-SMOTE	0.9828	0.9879	0.9800	0.9651	0.0200	0.0121
ADASYN	**0.9911**	**0.9970**	0.9847	0.9859	0.0153	**0.0030**
SMOTE-NC	0.9828	0.9946	0.9710	0.9718	0.0290	0.0054
SMOTE-Tomek	0.9886	0.9957	0.9816	0.9819	0.0184	0.0043
SMOTE-ENN	**0.9916**	**0.9962**	**0.9867**	**0.9877**	**0.0133**	**0.0038**
K-means SMOTE	0.9900	0.9875	**0.9925**	**0.9925**	**0.0075**	0.0125
Taiwanese companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9343	0.9454	0.9232	0.9250	0.0768	0.0546
BL-SMOTE	0.9545	0.9645	0.9445	0.9457	0.0555	0.0355
SVM-SMOTE	0.9449	0.9016	**0.9600**	0.8878	**0.0400**	0.0984
ADASYN	0.9417	0.9528	0.9307	0.9316	0.0693	0.0472
SMOTE-Tomek	0.9440	0.9538	0.9342	0.9355	0.0658	0.0462
SMOTE-ENN	**0.9600**	**0.9697**	0.9491	**0.9559**	0.0509	**0.0303**
K-means SMOTE	**0.9817**	**0.9750**	**0.9883**	**0.9882**	**0.0117**	**0.0250**
Polish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.8976	0.9288	0.8664	0.8743	0.1336	0.0712
BL-SMOTE	**0.9359**	**0.9650**	0.9068	**0.9120**	0.0932	**0.0350**
SVM-SMOTE	0.9327	0.9157	**0.9409**	0.8818	**0.0591**	0.0843
ADASYN	0.8972	0.9275	0.8667	0.8751	0.1333	0.0725
SMOTE-Tomek	0.8933	0.9232	0.8633	0.8711	0.1367	0.0768
SMOTE-ENN	0.9203	0.9513	0.8849	0.9040	0.1151	0.0487
K-means SMOTE	**0.9895**	**0.9822**	**0.9968**	**0.9968**	**0.0032**	**0.0178**

obtains the best *specificity*, *precision* and *type I error* values. Furthermore, SMOTE-ENN produces close values to the best results, it obtains the best *accuracy* and the second-best values for the rest of the metrics. Figure 12 shows the training and validation curves obtained by the best approaches concerning bankruptcy prediction in all datasets, which are AdaBoost in conjunction with ADASYN for the Spanish data, and AdaBoost in conjunction with K-means SMOTE for the Taiwanese and Polish data. The figure shows that AdaBoost fits small training data better than larger one; there is an inverse correlation between the training accuracy and the data size. On the other hand, the validation accuracy is relatively the same for all data sizes, specifically, with the Taiwanese and Polish data. Also in this experiment, the training curves are close to the validation ones, showing that the datasets are well-fitted by AdaBoost.

From this experiment, we can conclude that using K-means SMOTE as a preprocessing stage for the AdaBoost method incredibly improves its performance regarding predicting both classes, i.e., bankruptcy and solvency, in the numeric datasets (Taiwanese and Polish datasets).

H. XGBoost + BALANCING TECHNIQUES

This subsection presents the last experiment, carried out to find the most appropriate combination of classification algorithm XGBoost which is another *boosting*-based ensemble method, with the balancing approaches. XGBoost has also shown a higher performance in the imbalanced data classification compared to the other standard and ensemble methods [69], but as in the case of AdaBoost, using the balancing techniques significantly improves the classification performance. Thus, as in the previous experiment, the optimum number of estimators (*N_Estimators*) is 50, and also the entire data was fitted in each estimator by setting *Max_features* and *Max_Samples* to 1.0.

Table 13 shows the results achieved by the XGBoost method with each balancing technique. SMOTE-ENN surpassed the rest of the techniques in predicting the bankrupt companies (on the three datasets), while K-means SMOTE is the optimum to predict the solvent companies in all datasets also. Paying attention to the Spanish companies' dataset, as before, the conjunction of SMOTE-ENN with XGBoost gets the best *accuracy*, *recall* and *type II error*, whereas SMOTE-NC and SMOTE-Tomek obtain the second-best *recall* and *type II error*. In addition, again K-means SMOTE yields the best *specificity*, *precision* and *type I error*, while BL-SMOTE achieves the second-best values for the same metrics. Also, in the Taiwanese companies' dataset, the conjunction with SMOTE-ENN leads the XGBoost model to make the best bankrupt companies prediction in comparison with the other balancing techniques. Thus, it gets the best *accuracy*, *recall* and *type II error*, and the second-best values for the remaining metrics.

In addition, the outperforming balancing techniques applied to the Polish companies' dataset are the same used with Spanish and Taiwanese ones (SMOTE-ENN and K-means SMOTE), the second-best *recall* and *type II error* are obtained by SMOTE, while SVM-SMOTE yields the second-best *specificity* and *type I error*. In other words, SMOTE-ENN could be considered as the more adequate

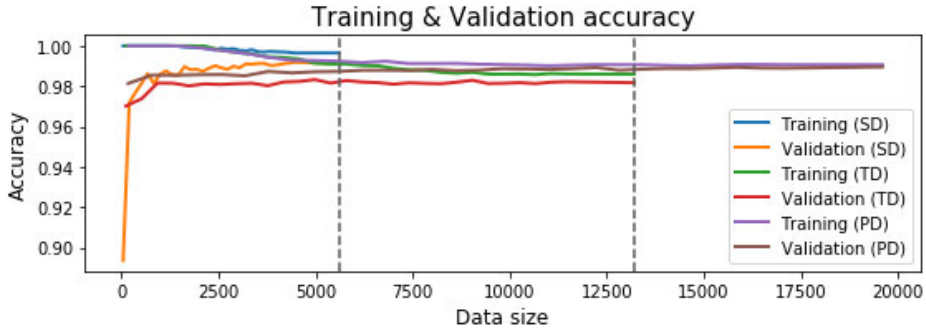

FIGURE 12. Training and validation accuracy curves obtained by the superior combination of AdaBoost with a balancing technique; AdaBoost + ADASYN in the Spanish dataset (SD), and AdaBoost + K-means SMOTE in the Taiwanese dataset (TD) and Polish dataset (PD). The dashed vertical lines indicate the size of the SD, TD and PD.

TABLE 13. Evaluation metrics values yielded by XGBoost applied to balanced datasets generated by each balancing technique. The first and second-best results for each metric are marked in boldface, the best value also has a gray background.

Spanish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	**0.9950**	0.9989	0.9911	0.9912	0.0089	0.0011
BL-SMOTE	0.9948	0.9968	**0.9929**	**0.9929**	**0.0071**	0.0032
SVM-SMOTE	0.9931	0.9955	0.9918	0.9855	0.0082	0.0045
ADASYN	0.9941	0.9982	0.9900	0.9901	0.0100	0.0018
SMOTE-NC	0.9941	**0.9993**	0.9889	0.9891	0.0111	**0.0007**
SMOTE-Tomek	0.9944	**0.9993**	0.9895	0.9897	0.0105	**0.0007**
SMOTE-ENN	**0.9955**	**0.9996**	0.9912	0.9918	0.0088	**0.0004**
K-means SMOTE	0.9916	0.9864	**0.9968**	**0.9968**	**0.0032**	0.0136
Taiwanese companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	0.9870	0.9979	0.9762	0.9768	0.0238	0.0021
BL-SMOTE	0.9858	0.9927	0.9788	0.9791	0.0212	0.0073
SVM-SMOTE	0.9800	0.9692	0.9838	0.9546	0.0162	0.0308
ADASYN	**0.9872**	**0.9974**	0.9771	0.9774	0.0229	**0.0026**
SMOTE-Tomek	0.9867	**0.9974**	0.9761	0.9766	0.0239	**0.0026**
SMOTE-ENN	**0.9923**	**0.9991**	**0.9846**	**0.9867**	**0.0154**	**0.0009**
K-means SMOTE	0.9845	0.9741	**0.9948**	**0.9947**	**0.0052**	0.0259
Polish companies' dataset						
Balancing technique	**Accuracy**	**Recall**	**Specificity**	**Precision**	**Type I error**	**Type II error**
SMOTE	**0.9942**	**0.9996**	0.9889	0.9890	0.0111	**0.0004**
BL-SMOTE	0.9941	0.9956	0.9925	**0.9926**	0.0075	0.0044
SVM-SMOTE	0.9932	0.9887	**0.9954**	0.9905	**0.0046**	0.0113
ADASYN	0.9941	0.9994	0.9889	0.9891	0.0111	0.0006
SMOTE-Tomek	**0.9943**	0.9993	0.9894	0.9895	0.0106	0.0007
SMOTE-ENN	0.9936	**0.9998**	0.9864	0.9883	0.0136	**0.0002**
K-means SMOTE	0.9918	0.9846	**0.9990**	**0.9990**	**0.0010**	0.0154

preprocessing stage before applying XGBoost aiming to predict the bankruptcy situation, whereas K-means SMOTE is the best to predict the solvent companies.

Furthermore, Figure 13 illustrates the training and validation accuracy curves obtained by the superiors XGBoost and balancing techniques combinations concerning bankruptcy prediction in all datasets (XGBoost + SMOTE-ENN). Thus, the same as experiment (VII-D) curves, all approaches training curves are equal to 1.0, and the validation curves are less, but increasing being close to the training ones. Also in this experiment, the three datasets good-fitted to the XGBoost model.

I. SUMMARY OF RESULTS

This section aims to clarify the findings of the experiments conducted.

Firstly, we present a visual representation of all metrics values obtained by each combination of classifier and balancing technique addressed in this study, and applied to the Spanish, Taiwanese and Polish companies' datasets in the Figures 14, 15 and 16, respectively, where it can be seen the high values reached by almost all the methods in most of the metrics (very low errors also). Just DBN shows a lower performance overall, with a noticeable variation of the metrics values obtained.

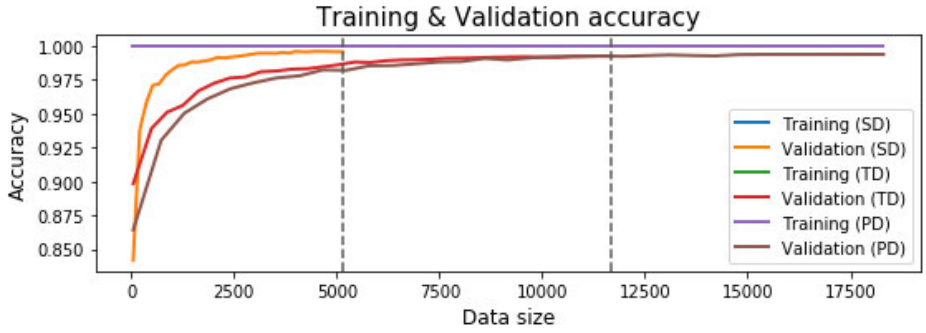

FIGURE 13. Training and validation accuracy curves obtained by the superior combination of XGBoost with a balancing technique; XGBoost + SMOTE-ENN in the three datasets. The dashed vertical lines indicate the size of the Spanish dataset (SD), Taiwanese dataset (TD) and Polish dataset (PD).

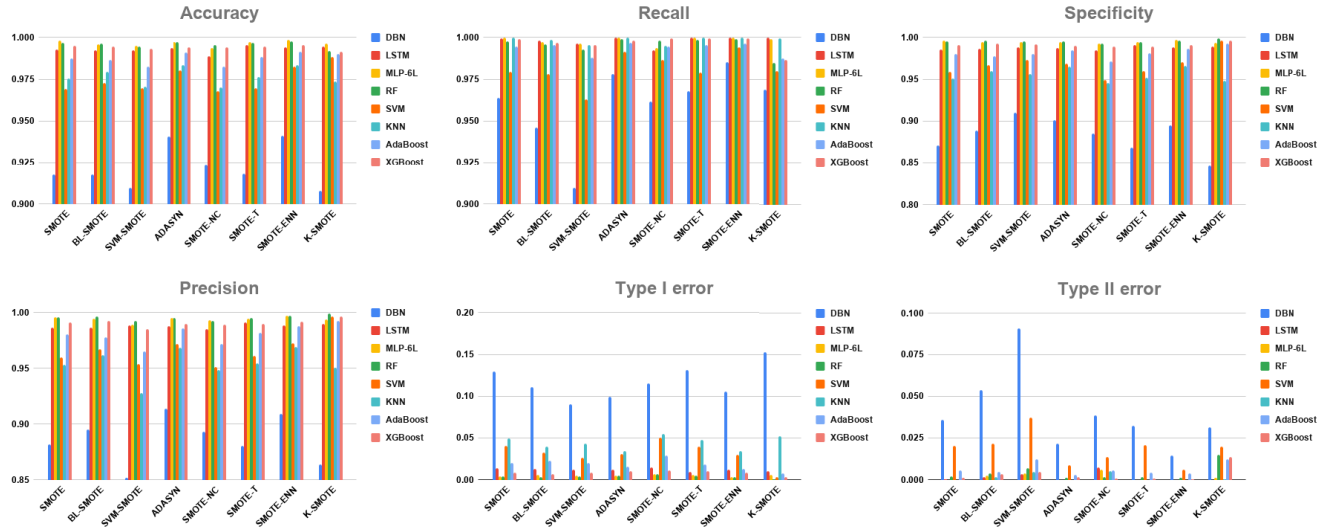

FIGURE 14. Accuracy, recall, specificity, precision, type I and type II error obtained by all the methods on the Spanish companies' dataset.

The reason is DBN was designed mainly to process a different type of data than those considered in this study, i.e., images; since it was focused on image and pattern classification [42], [70], [71].

MLP-6L, LSTM, RF and XGBoost in conjunction with every balancing technique stand out with high efficiency regarding predicting companies' financial status. In other words, the evaluation metrics values obtained by these algorithms applied on the Spanish companies' dataset are excellent: *accuracy* and *recall* metric values exceeded 0.99, while *type I error* is around 0.01, and *type II error* is very close to 0.0 most of times. Thus, these methods yield very low bankruptcy and solvency misclassification.

Moreover, ensemble SVM, ensemble KNN and AdaBoost show good performance regarding predicting the bankrupt companies in the Spanish dataset, ensemble SVM achieved *accuracy* and *recall* metric values exceeded 0.96. Ensemble KNN shows significant performance in predicting the bankrupt companies, it obtains *recall* around 1.0 for all the balancing techniques, but not high *specificity* as much as the other classifiers. Also, AdaBoost proved that it could be a competitor in predicting bankruptcy with *accuracy* and *recall* exceeded 0.98.

In order to select the best approach to address the bankruptcy prediction problem, now we present a comparison of all the best combinations between classifiers + data balancing techniques, according to the obtained results for each of them previously described in Sections VII-A to VII-H. These are summarized in Table 14, which shows the evaluation metrics values obtained by the best approaches.

As it can be seen in the table, SMOTE-ENN is the best data balancing method applied to the Spanish companies' dataset, and just in two cases other algorithms reach better results: SMOTE-Tomek and ADASYN. The combination of MLP-6L with SMOTE-ENN shows the highest performance in predicting both the bankrupt and solvent companies, but the combination of RF and SMOTE-ENN obtains very close results.

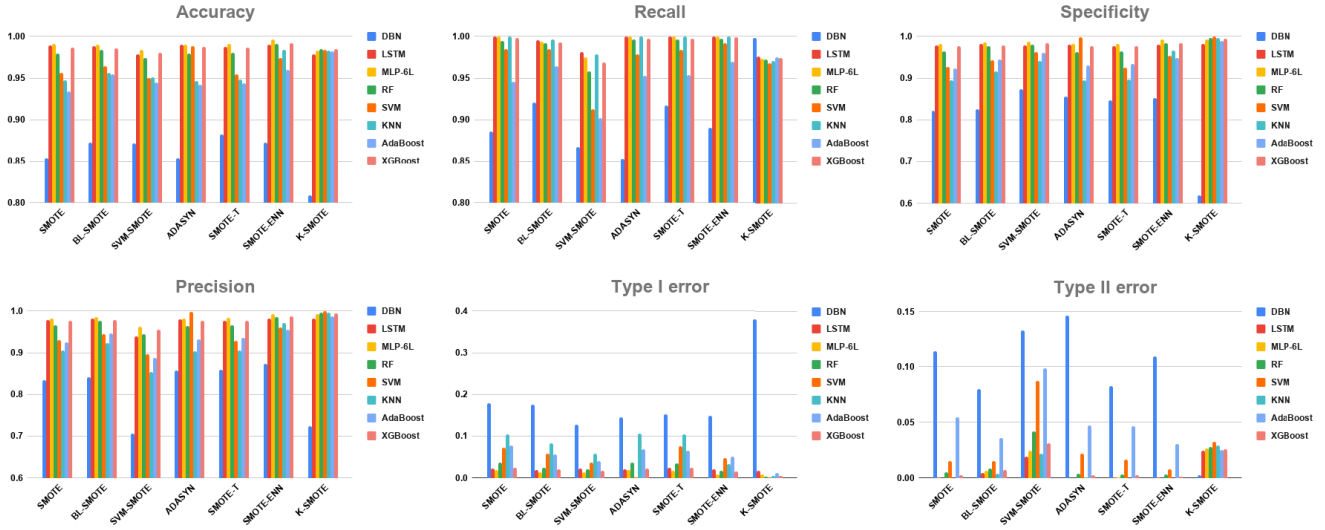

FIGURE 15. Accuracy, recall, specificity, precision, type I and type II error obtained by all the methods on the Taiwanese companies' dataset.

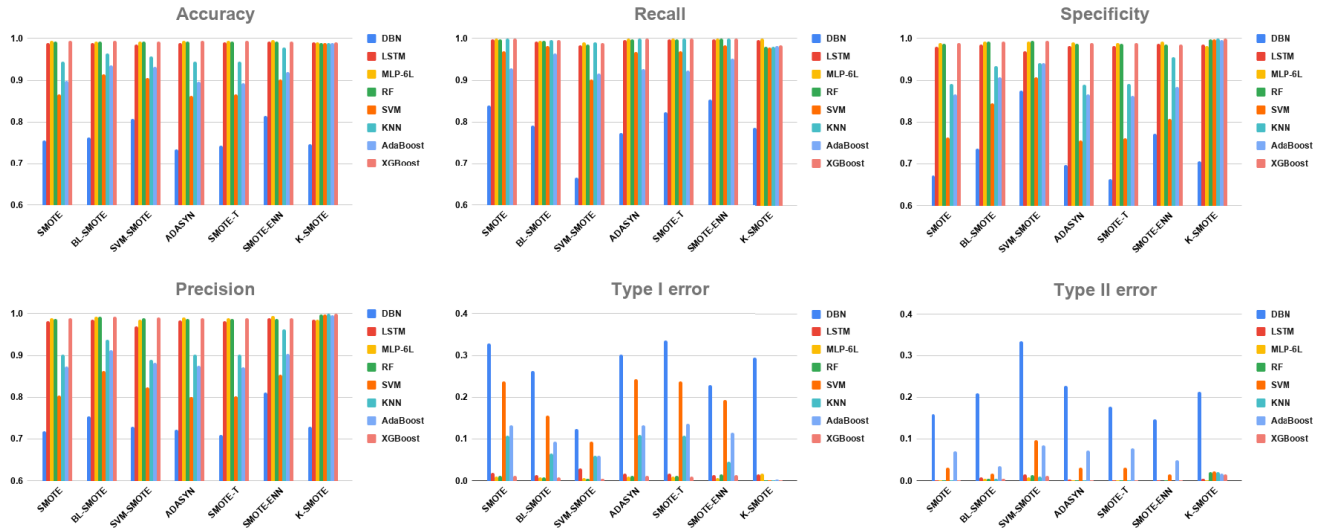

FIGURE 16. Accuracy, recall, specificity, precision, type I and type II error obtained by all the methods on the Polish companies' dataset.

With respect to the Taiwanese companies' dataset, due to the increment of its complexity compared to the Spanish one, it is expected that the extra load on the classifier could affect its performance whether decrease bankruptcy, solvency prediction or both. Thus, the increase of the complexity nearly imperceptibly affected the performance of MLP-6L and XGBoost concerning the estimation of bankrupt companies, and shows a noticeable affection in all metrics of the remaining classifiers. Furthermore, also in this dataset, the combination of MLP-6L with SMOTE-ENN outperforms the other combinations in predicting the bankrupt and solvent companies with the best values of all metrics. The combinations of LSTM + ADASYN and ensemble KNN + SMOTE-ENN obtain the best results concerning predict the bankrupt companies, both combinations got the same values of *recall* and *type II error* obtained by the superior combination (MLP-6L + SMOTE-ENN).

Regarding the Polish companies' dataset, it is the most complex one in this study, with this data also, MLP-6L with a data preprocessing stage using SMOTE-ENN is the best approach to classify the bankrupt companies reaching the best *accuracy*, *recall* and *type II error*, whereas the combination of ensemble SVM with K-means SMOTE shows the optimum performance in predicting the solvent companies.

However, from this summary, we can conclude that SMOTE-ENN shows significant performance with most of the classifiers no matter the data complexity. The combination of MLP-6L with SMOTE-ENN obtains the best bankrupt

TABLE 14. Results of the outperforming classifiers + balancing techniques combinations according to the considered metrics. Best results for each metric are marked in boldface and gray background. The second-best value is just in boldface.

Dataset	Combinations	Accuracy	Recall	Specificity	Precision	Type I error	Type II error
Spanish companies' dataset	**DBN + SMOTE-ENN**	0.9414	0.9854	0.8946	0.9092	0.1054	0.0146
	LSTM + SMOTE-Tomek	0.9955	**1.0**	0.9908	0.9914	0.0092	**0.0**
	MLP-6L + SMOTE-ENN	**0.9986**	**1.0**	**0.9972**	**0.9974**	**0.0028**	**0.0**
	RF + SMOTE-ENN	**0.9979**	0.9989	**0.9968**	**0.9970**	**0.0032**	0.0011
	SVM + SMOTE-ENN	0.9825	0.9940	0.9703	0.9729	0.0297	0.0060
	KNN +SMOTE-ENN	0.9835	**1.0**	0.9659	0.9691	0.0341	**0.0**
	AdaBoost + ADASYN	0.9911	0.9970	0.9847	0.9859	0.0153	0.0030
	XGBoost + SMOTE-ENN	0.9955	0.9996	0.9912	0.9918	0.0088	0.0004
Taiwanese companies' dataset	**DBN + SMOTE-ENN**	0.8809	0.9465	0.8153	0.8367	0.1847	0.0535
	LSTM + ADASYN	0.9897	**1.0**	0.9795	0.9797	0.0205	**0.0**
	MLP + SMOTE-ENN	**0.9963**	**1.0**	**0.9921**	**0.9931**	**0.0079**	**0.0**
	RF + SMOTE-ENN	0.9906	0.9970	0.9834	0.9857	0.0166	0.0030
	SVM + SMOTE-ENN	0.9741	0.9921	0.9535	0.9605	0.0465	0.0079
	KNN + SMOTE-ENN	0.9843	**1.0**	0.9665	0.9714	0.0335	**0.0**
	AdaBoost + K-means SMOTE	0.9817	0.9750	**0.9883**	**0.9882**	**0.0117**	0.0250
	XGBoost + SMOTE-ENN	**0.9923**	0.9991	0.9846	0.9867	0.0154	0.0009
Polish companies' dataset	**DBN +SMOTE-ENN**	0.8151	0.8531	0.7718	0.8115	0.2282	0.1469
	LSTM + SMOTE-ENN	0.9931	0.9988	0.9866	0.9884	0.0134	0.0012
	MLP-6L + SMOTE-ENN	**0.9967**	**0.9999**	0.9930	0.9939	0.007	**0.0001**
	RF +SMOTE-ENN	0.9926	0.9992	0.9852	0.9871	0.0148	0.0008
	SVM +K-means SMOTE	0.9885	0.9784	**0.9987**	**0.9986**	**0.0013**	0.0216
	KNN +SMOTE-ENN	0.9788	**0.9999**	0.9546	0.9618	0.0454	**0.0001**
	AdaBoost + K-means SMOTE	0.9895	0.9822	**0.9968**	**0.9968**	**0.0032**	0.0178
	XGBoost + SMOTE-ENN	**0.9936**	0.9998	0.9864	0.9883	0.0136	0.0002

companies classification in all the datasets according to the *accuracy*, *recall* and *type II error*. Also, yields the best solvent companies prediction in the Spanish and Polish companies, whereas ensemble SVM with K-means SMOTE were the best in the Polish one.

Furthermore, the aforementioned summary is based mainly on the outperforming combinations of classifiers and balancing techniques in predicting bankruptcy (the aim of this study). The real superior combinations in predicting the solvent companies are K-means SMOTE with RF, ensemble SVM and ensemble KNN in the Spanish, Taiwanese and Polish datasets, respectively.

VIII. COMPARISON WITH THE STATE OF THE ART

Once we have tested the DL and ensemble methods together with the data balancing techniques, we aim now to compare the most effective ones with previous algorithms/results of the state of the art working with the same datasets considered in this study.

First, focusing on the Spanish companies' dataset, Table 15 presents a comparison between the best approach selected in this work (MLP-6L + SMOTE-ENN) and the best algorithms or combinations found in five previous works, namely: [4], [5], [7], [36], [46]. In [4], the superior classifier was RF applied to the Spanish dataset balanced using a technique based on dividing the dataset into subsets, which were balanced using a simple oversampling approach. In [7], several balancing techniques were utilized in order to solve the data inconstancy problem as a preprocessing stage before applying C4.5 classifier. SMOTE-ENN balancing technique outperformed all the rest of the balancing techniques. In addition, in [5], combining simple DLR status space with MLP obtained the best results compared with other classifiers. Just SMOTE was applied to solve the data inconsistency problem. Whereas, in [46], the combination of SMOTE and AdaBoost ensemble methods utilizing Reduced Error Pruning Tree (REPT), yielded the best results compared with other basic and ensemble classifiers. It also outperformed the results of using this combination with five different Feature Selection approaches. Finally, in [36], the combination of RF with a Cost-Sensitive Classification (CSC) method outperformed many other ensemble and cost-sensitive methods.

Common metrics computed in all the studies are considered in Table 15. As it can be seen in the table, the outputs obtained by MLP-6L + SMOTE-ENN clearly outperform the rest of the results obtained by previous approaches, regarding the bankruptcy prediction and misprediction, according to *accuracy*, *recall* and *type II error* metrics. Our tested approach outperformed even the most promising approaches proposed in our most recent paper in this scope [36].

The second comparison will be focused on the Taiwanese dataset, which is more complicated than the Spanish one. Here we compare the performance of the superior combination of classifier and balancing techniques applied to the Taiwanese dataset in this study with the best approach addressed in [48]. In that work, the authors proved that combining the Financial Ratios (FRs) and Corporate Governance Indicators (CGIs) improves the classifiers' performance in predicting Taiwanese companies' financial status. Moreover, five Feature Selection approaches were compared for reducing data dimensionality after this combination. Thus, SVM with Stepwise Discriminant Analysis (SDA) Feature Selection method to the combination of FRs and CGIs (FC) obtained the best results. Table 16 shows the results of

TABLE 15. Best accuracy, recall and type II error metrics values obtained in this and previous studies on the Spanish companies' dataset. Best values are marked in gray background and boldface.

Metric	MLP-6L+ SMOTE-ENN	RF [4]	C4.5+ SMOTE-ENN [7]	DLR-MLP+ SMOTE [5]	REPT-AdaBoost+SMOTE [46]	RF + CSC [36]
Accuracy	**0.9986**	0.9129	0.8761	0.985	0.983	0.9117
Recall	**1.0**	0.6018	0.8763	0.677	0.55	0.912
Type II error	**0.0**	0.3982	0.1237	0.323	0.45	0.088

TABLE 16. Best accuracy, recall, specificity, type I error and type II error metrics values obtained in this and previous study on the Taiwanese companies' dataset. Best values are marked in gray background and boldface.

Combination	Accuracy	Recall	Specificity	Type I error	Type II error
MLP-6L + SMOTE-ENN	**0.9963**	**1.0**	**0.9921**	**0.0079**	**0.0**
SVM + SDA-FC [48]	0.815	0.792	0.837	0.163	0.208

TABLE 17. AUC metric values for all the approaches of this study and a previous one from the state of the art on the Polish companies' dataset. Best values are marked in gray background and boldface.

Balancing Technique	DBN	LSTM	MLP-6L	RF	SVM	KNN	AdaBoost	XGBoost	EXGB [47]
SMOTE	0.7557	0.9897	0.9945	0.9932	0.8660	0.9456	0.8976	0.9942	**0.959**
BL-SMOTE	0.7639	0.9890	0.9936	**0.9934**	0.9136	0.9649	0.9359	0.9941	0.944
SVM-SMOTE	0.7706	0.9773	0.9921	0.9903	0.9047	0.9653	0.9283	0.9920	0.940
ADASYN	0.7352	0.9898	0.9952	0.9930	0.8627	0.9446	0.8971	0.9941	0.941
SMOTE-Tomek	0.7433	0.9904	0.9947	0.9933	0.8655	0.9453	0.8932	**0.9943**	0.955
SMOTE-ENN	**0.8125**	**0.9927**	**0.9965**	0.9922	0.8960	0.9772	0.9181	0.9931	
K-means SMOTE	0.7463	0.9902	0.9910	0.9892	**0.9886**	**0.9895**	**0.9895**	0.9918	

SVM + SDA-FC applied to the Taiwanese dataset, and the best approach used in this study. As it can be seen in the table, MLP-6L + SMOTE-ENN yields much better results than the best approach in the previous study.

The final comparison is focused on the Polish companies' dataset. We compare the best results obtained by all the approaches tested here with the algorithm proposed in [47], where the authors compared the performance of several classifiers and regression methods with a novel approach that utilizes Extreme Gradient Boosting (EXGB) for learning an ensemble of decision trees in order to predict companies' financial status. Moreover, EXGB is widely used in Kaggle competitions[3] on classification problems. Thus, the results found in [47] outperformed all the referenced methods there, regarding the prediction of companies' financial status. In their study, they divided the dataset into five subsets depending on the years.

Area Under ROC Curve (AUC) is the only metric used to evaluate the classification and regression models' performance. Receiver Operating Characteristics (ROC) curve [72] is a graph showing the performance of classifiers and regression models by means of a series of thresholds. In the case of binary classification, there is only one threshold value. Equation 23 represents the AUC formula:

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{23}$$

Thus, in order to compare the methods, we computed the AUC value for all the DL and RF methods considered in this study together with the data balancing techniques. Table 17 shows the results of this metric for all the approaches and the one in [47].

As it can be seen, LSTM, MLP-6L, RF and XGBoost, combined with all the balancing techniques, and ensemble SVM, ensemble KNN and AdaBoost in conjunction with just K-means SMOTE, get better results than the EXGB algorithm. Thus, again MLP-6L + SMOTE-ENN reaches the best metric performance beating in almost four points the state of the art method.

Accordingly, the firm conclusion extracted from the comparison results is that the superior approach adopted to predict companies' financial failure in this study (i.e., MLP-6L + SMOTE-ENN) outperformed the other approaches addressed in the state of the art with the same Spanish, Taiwanese, and Polish companies' datasets.

IX. CONCLUSION AND FUTURE WORK

This paper has addressed companies' financial status classification problem using three DL algorithms, three bagging ensemble and two boosting ensemble classification methods. As DL algorithms, the Deep Belief Network (DBN), Multi-Layer Perceptron with 6 layers (MLP-6L), and Long-Short Term Memory (LSTM) have been chosen. The bagging ensemble classifiers are Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN); and the boosting ensemble classifiers are ADaptive boost (AdaBoost) and Extreme Gradient Boosting (XGBoost).

A difficult and very imbalanced problem is faced in this work, using three different datasets: A Spanish companies' dataset, which was provided by Infotel company, a Taiwanese companies' dataset, collected from Taiwan economic journal, and a Polish companies' dataset, collected from the

[3]https://www.kaggle.com/competitions

Emerging Markets Information Service (EMIS). In order to cope with this problem, three types of balancing techniques have been utilized, namely: oversampling (SMOTE, Borderline SMOTE, SMOTE-NC, SVM-SMOTE and ADAYSN), oversampling+undersampling (SMOTE-ENN and SMOTE-Tomek); and clustering-based balancing (K-means SMOTE), so eight data resampling methods have been combined with the classifiers.

Several metrics have been considered (in addition to the classical accuracy) to properly measure the performance of each classification method applied to each balanced dataset. After extensive experiments were done, and according to the evaluation metrics, MLP-6L applied to the datasets generated using SMOTE-ENN balancing technique obtained the best results regarding predicting companies' financial failure.

Indeed SMOTE-ENN was the mutual superior balancing technique for all DL methods leading them to reach the lowest bankruptcy misclassification. The best DL approaches have been compared with state of the art methods applied by other authors to the same datasets, outperforming the results previously obtained.

For future work, we will study how to improve the DL methods, by fine-tuning their parameter values, for instance by means of meta-optimization applying Evolutionary Algorithms. We will consider more complex datasets related to this problem as well as to other classification problems. We also aim to develop a specialized new data balancing technique, in order to handle the problem of data distribution inconsistency more efficiently.

REFERENCES

[1] J. L. Bellovary, D. E. Giacomino, and M. D. Akers, "A review of bankruptcy prediction studies: 1930 to present," *J. Financial Educ.*, vol. 33, pp. 1–42, Dec. 2007.

[2] D. L. Minh, A. Sadeghi-Niaraki, H. D. Huy, K. Min, and H. Moon, "Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network," *IEEE Access*, vol. 6, pp. 55392–55404, 2018.

[3] D. Veganzones and E. Séverin, "An investigation of bankruptcy prediction in imbalanced datasets," *Decis. Support Syst.*, vol. 112, pp. 111–124, Aug. 2018.

[4] H. Jawazneh, A. Mora, and P. Castillo, "Predicting the financial status of companies using data balancing and classification methods," in *Proc. Int. Work-Conf. Time Ser. (ITISE)*. Granada, Spain: Godel Impresiones Digitales SL, 2017, pp. 661–673.

[5] A. Rodan, P. A. Castillo, H. Faris, A. M. Mora, and H. Jawazneh, "Forecasting business failure in highly imbalanced distribution based on delay line reservoir," in *Proc. 26th Eur. Symp. Artif. Neural Netw. (ESANN)*, Bruges, Belgium, Apr. 2018, pp. 1–6. [Online]. Available: http://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2018-105.pdf

[6] A. Rodan and P. Tino, "Minimum complexity echo state network," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 131–144, Jan. 2011.

[7] W. Alswiti, H. Faris, H. Aljawazneh, S. Safi, P. Castillo, A. Mora, R. A. Khurma, and H. Alsawalqah, "Empirical evaluation of advanced oversampling methods for improving bankruptcy prediction," in *Proc. Int. Conf. Time Ser. Forecasting (ITISE)*, Granada, Spain, Sep. 2018, pp. 1495–1506.

[8] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[9] Y. Jang, I.-B. Jeong, Y. K. Cho, and Y. Ahn, "Predicting business failure of construction contractors using long short-term memory recurrent neural network," *J. Construct. Eng. Manage.*, vol. 145, no. 11, Nov. 2019, Art. no. 04019067.

[10] T. Le, M. Lee, J. Park, and S. Baik, "Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset," *Symmetry*, vol. 10, no. 4, p. 79, Mar. 2018.

[11] Y. Jang, I. Jeong, and Y. K. Cho, "Business failure prediction of construction contractors using a LSTM RNN with accounting, construction market, and macroeconomic variables," *J. Manage. Eng.*, vol. 36, no. 2, Mar. 2020, Art. no. 04019039.

[12] E. Fedorova, E. Gilenko, and S. Dovzhenko, "Bankruptcy prediction for Russian companies: Application of combined classifiers," *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7285–7293, Dec. 2013.

[13] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[14] A. A. Kasgari, M. Divsalar, M. R. Javid, and S. J. Ebrahimian, "Prediction of bankruptcy Iranian corporations through artificial neural network and probit-based analyses," *Neural Comput. Appl.*, vol. 23, nos. 3–4, pp. 927–936, Sep. 2013.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

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

[19] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[20] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learn. (ICML)*. San Francisco, CA, USA: Morgan Kaufmann, 1996, pp. 148–156.

[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[23] N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving imbalanced dataset classification using oversampling and gradient boosting," in *Proc. 5th Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2019, pp. 217–222.

[24] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[25] F. Last, G. Douzas, and F. Bacao, "Oversampling for imbalanced learning based on k-means and SMOTE," 2017, *arXiv:1711.00837*. [Online]. Available: http://arxiv.org/abs/1711.00837

[26] B. Zhou, Z. Li, S. Zhang, X. Zhang, X. Liu, and Q. Ma, "Analysis of factors affecting hit-and-run and non-hit-and-run in vehicle-bicycle crashes: A non-parametric approach incorporating data imbalance treatment," *Sustainability*, vol. 11, no. 5, p. 1327, Mar. 2019.

[27] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 1322–1328.

[28] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," *Knowl.-Based Syst.*, vol. 41, pp. 16–25, Mar. 2013.

[29] S. S. Devi and Y. Radhika, "A survey on machine learning and statistical techniques in bankruptcy prediction," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 2, pp. 133–139, Apr. 2018.

[30] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *J. Finance*, vol. 23, no. 4, pp. 589–609, Sep. 1968.

[31] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *J. Accounting Res.*, vol. 18, no. 1, pp. 109–131, Apr. 1980.

[32] J. Brozyna, G. Mentel, and T. Pisula, "Statistical methods of the bankruptcy prediction in the logistics sector in Poland and Slovakia," *Transformations Bus. Econ.*, vol. 15, no. 1, pp. 80–96, 2016.

[33] S. Jones and D. A. Hensher, "Predicting firm financial distress: A mixed logit model," *Accounting Rev.*, vol. 79, no. 4, pp. 1011–1038, Oct. 2004.

[34] P. P. M. Pompe and A. J. Feelders, "Using machine learning, neural networks, and statistics to predict corporate bankruptcy," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 12, no. 4, pp. 267–276, Jul. 1997.

[35] J. Min and Y. Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," *Expert Syst. Appl.*, vol. 28, no. 4, pp. 603–614, May 2005.

[36] N. Ghatasheh, H. Faris, R. Abukhurma, P. A. Castillo, N. Al-Madi, A. M. Mora, A. M. Al-Zoubi, and A. Hassanat, "Cost-sensitive ensemble methods for bankruptcy prediction in a highly imbalanced data distribution: A real case from the Spanish market," *Prog. Artif. Intell.*, vol. 9, no. 4, pp. 361–375, Dec. 2020.

[37] K.-S. Shin, T. S. Lee, and H.-J. Kim, "An application of support vector machines in bankruptcy prediction model," *Expert Syst. Appl.*, vol. 28, no. 1, pp. 127–135, Jan. 2005.

[38] S. R. Islam, W. Eberle, S. K. Ghafoor, S. C. Bundy, D. A. Talbert, and A. Siraj, "Investigating bankruptcy prediction models in the presence of extreme class imbalance and multiple stages of economy," 2019, *arXiv:1911.09858*. [Online]. Available: http://arxiv.org/abs/1911.09858

[39] F. J. L. Iturriaga and I. P. Sanz, "Bankruptcy visualization and prediction using neural networks: A study of U.S. Commercial banks," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 2857–2869, Apr. 2015.

[40] Z. Lanbouri and S. Achchab, "A hybrid deep belief network approach for financial distress prediction," in *Proc. 10th Int. Conf. Intell. Syst., Theories Appl. (SITA)*, Oct. 2015, pp. 1–6.

[41] M.-J. Kim, D.-K. Kang, and H. B. Kim, "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1074–1082, Feb. 2015.

[42] S.-H. Yeh, C.-J. Wang, and M.-F. Tsai, "Deep belief networks for predicting corporate defaults," in *Proc. 24th Wireless Opt. Commun. Conf. (WOCC)*, Oct. 2015, pp. 159–163.

[43] T. Hosaka, "Bankruptcy prediction using imaged financial ratios and convolutional neural networks," *Expert Syst. Appl.*, vol. 117, pp. 287–299, Mar. 2019.

[44] A. Vieira, P. A. Castillo, and J. J. Merelo, "Application of HLVQ and G-prop neural networks to the problem of bankruptcy prediction," in *Proc. Int. Work-Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2003, pp. 655–662.

[45] E. Alfaro-Cid, P. Castillo, A. Esparcia, K. Sharman, J. Merelo, A. Prieto, A. M. Mora, and J. L. J. Laredo, "Comparing multiobjective evolutionary ensembles for minimizing type I and II errors for bankruptcy prediction," in *Proc. IEEE Congr. Evol. Comput., IEEE World Congr. Comput. Intell.*, Jul. 2008, pp. 2902–2908.

[46] H. Faris, R. Abukhurma, W. Almanaseer, M. Saadeh, A. M. Mora, P. A. Castillo, and I. Aljarah, "Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market," *Prog. Artif. Intell.*, vol. 9, no. 1, pp. 31–53, Mar. 2020.

[47] M. Zięba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, Oct. 2016.

[48] D. Liang, C.-C. Lu, C.-F. Tsai, and G.-A. Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," *Eur. J. Oper. Res.*, vol. 252, no. 2, pp. 561–572, Jul. 2016.

[49] N. Le Roux and Y. Bengio, "Deep belief networks are compact universal approximators," *Neural Comput.*, vol. 22, no. 8, pp. 2192–2207, Aug. 2010.

[50] L. Zhang, C. Aggarwal, and G.-J. Qi, "Stock price prediction via discovering multi-frequency trading patterns," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 2141–2149.

[51] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," *IEEE Access*, vol. 6, pp. 24411–24432, 2018.

[52] C. Tsai and J. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2639–2649, May 2008.

[53] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," in *Proc. Int. Conf. Adv. Technol. Eng. (ICATE)*, Jan. 2013, pp. 1–9.

[54] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, "An empirical study of distance metrics for k-nearest neighbor algorithm," in *Proc. 2nd Int. Conf. Ind. Appl. Eng.*, 2015, pp. 280–285.

[55] M.-J. Kim and D.-K. Kang, "Ensemble with neural networks for bankruptcy prediction," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3373–3379, Apr. 2010.

[56] P. Kaur and A. Gosain, "Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise," in *Proc. ICT Based Innov.* Singapore: Springer, 2018, pp. 23–30.

[57] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *Proc. Int. Symp. Intell. Comput. Appl.* Berlin, Germany: Springer, 2009, pp. 461–471.

[58] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," *J. Mach. Learn. Technol*, vol. 2, pp. 2229–3981, Jan. 2011.

[59] M. Stojanović, M. Apostolović, D. Stojanović, Z. Milošević, A. Toplaović, V. Mitić-Lakušić, and M. Golubović, "Understanding sensitivity, specificity and predictive values," *Vojnosanitetski Pregled*, vol. 71, no. 11, pp. 1062–1065, 2014.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[61] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015.

[62] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[63] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018, *arXiv:1811.03378*. [Online]. Available: http://arxiv.org/abs/1811.03378

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[65] S. M. Omohundro, *Five Balltree Construction Algorithms*. Berkeley, CA, USA: International Computer Science Institute Berkeley, 1989.

[66] A. Singh, A. Yadav, and A. Rana, "K-means with three different distance metrics," *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, Apr. 2013.

[67] H. Jabbar and R. Z. Khan, "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)," *Comput. Sci., Commun. Instrum. Devices*, pp. 163–172, Dec. 2015.

[68] J. Hearty, *Advanced Machine Learning With Python*. Birmingham, U.K.: Packt, 2016.

[69] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: Adaption of different imbalance ratios," *Expert Syst. Appl.*, vol. 98, pp. 105–117, May 2018.

[70] M. Ahmad, D. Ai, G. Xie, S. F. Qadri, H. Song, Y. Huang, Y. Wang, and J. Yang, "Deep belief network modeling for automatic liver segmentation," *IEEE Access*, vol. 7, pp. 20585–20595, 2019.

[71] D. Bibin, M. S. Nair, and P. Punitha, "Malaria parasite detection from peripheral blood smear images using deep belief networks," *IEEE Access*, vol. 5, pp. 9099–9108, 2017.

[72] J. Muschelli, "ROC and AUC with a binary predictor: A potentially misleading metric," *J. Classification*, vol. 37, no. 3, pp. 1–13, Dec. 2019.

H. ALJAWAZNEH was born in Amman, Jordan, in 1991. He received the bachelor's degree in computer science from The University of Jordan, Amman, in 2014, and the master's degree in computer engineering from Yarmouk University, Irbid, Jordan, in 2016. He is currently pursuing the Ph.D. degree in computer science with the University of Granada, Spain. His research interests include deep learning and financial status forecasting.

A. M. MORA was born in Granada, Spain, in 1977. He received the degree in computer sciences and the Ph.D. degree, researching on multi-objective ant colony optimization for solving a bi-criteria pathfinding problem in a military environment, from the University of Granada, in 2001 and 2009, respectively.

He is currently an Associate Professor with the Department of Signal Theory, Telematics and Communications, University of Granada, where he has worked as a Contracted Researcher and a Substitute Professor for 14 years. He has published more than 25 articles in international journals and more than 100 papers in top-rated international conferences. His research interests include bioinspired algorithms and their applications to data analysis, network security, and video games, among others.

Dr. Mora is a Founder Member of the Spanish Society on Videogames Research (SECiVi). He has being an Active Member of review boards in IEEE journals, such as IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES, and a PC and an Organization Member of IEEE conferences, such as IEEE Conference on Computational Intelligence and Games and IEEE Evolutionary Computation Conference.

P. GARCÍA-SÁNCHEZ was born in Granada, Spain, in 1983. He received the degree in computer science and the Ph.D. degree with a dissertation in the field of service-oriented evolutionary algorithms from the University of Granada, Spain, in 2007 and 2014, respectively.

He has been a Substitute Associate Lecturer with the University of Cádiz, Spain, for four years, and he has also been working as an Associate Lecturer with the Department of Computer Architecture and Technology, University of Granada, since 2021. He has been the Lead Researcher of the PETRA project, granted by the Spanish Traffic Agency. He has published more than 60 papers in international conferences and 18 articles in indexed journals. His research interests include distributed and parallel algorithms, artificial intelligence applied to videogames, and scientometrics.

Dr. García-Sánchez is a member of the Spanish Network of Excellence in Research and Development and Science in Videogames (RiDiVi) and has been a program committee member of more than ten international conferences, being part of the organization team of conferences, such as the IEEE Computational Intelligence in Games (CIG) and the European Conference on the Applications of Evolutionary and bio-inspired Computation (EvoAPPS), among others.

P. A. CASTILLO-VALDIVIESO was born in Granada, Spain, in 1974. He received the B.Sc. degree in computer science and the Ph.D. degree from the University of Granada, Spain, in 1997 and 2000, respectively.

He was a Teaching Assistant with the Department of Computer Science, University of Jaén, Spain. He was also a Visiting Researcher with Napier University, Edinburg, U.K., in July 1998, and Santa Fe Institute, Santa Fe, NM, USA, in September 2000. He has been the leader of several research projects, and directed five Ph.D. students. He is currently an Associate Professor with the Department of Computer Architecture and Technology, University of Granada. His research interests include bio-inspired systems, hybrid systems, and the combination of evolutionary algorithms and neural networks.