Received: 3 December 2021 | Revised: 21 September 2022 | Accepted: 25 September 2022

DOI: 10.1002/for.2915

RESEARCH ARTICLE

WILEY

A dynamic performance evaluation of distress prediction models

Mohammad Mahdi Mousavi[1] | **Jamal Ouenniche**[2] | **Kaoru Tone**[3]

[1]School of Management, University of Bradford, Bradford, UK

[2]Business School, University of Edinburgh, Edinburgh, UK

[3]National Graduate Institute for Policy Studies, Tokyo, Japan

Correspondence
Mohammad Mahdi Mousavi, School of Management, University of Bradford, Bradford, UK.
Email: m.mousavi@bradford.ac.uk

Abstract

So far, the dominant comparative studies of competing distress prediction models (DPMs) have been restricted to the use of static evaluation frameworks and as such overlooked their performance over time. This study fills this gap by proposing a Malmquist Data Envelopment Analysis (DEA)-based multi-period performance evaluation framework for assessing competing static and dynamic statistical DPMs and using it to address a variety of research questions. Our findings suggest that (1) dynamic models developed under duration-dependent frameworks outperform both dynamic models developed under duration-independent frameworks and static models; (2) models fed with financial accounting (FA), market variables (MV), and macroeconomic information (MI) features outperform those fed with either MVMI or FA, regardless of the frameworks under which they are developed; (3) shorter training horizons seem to enhance the aggregate performance of both static and dynamic models.

KEYWORDS

corporate credit risk, distress prediction models, Malmquist productivity index, performance evaluation

1 | INTRODUCTION

An early warning system of corporate failure events (e.g., bankruptcy and financial distress) has such economic benefits for stakeholders (e.g., managers, investors, auditors, and regulators) that many models from the fields of probability and statistics, operational research, and artificial intelligence have been designed to predict them—for a comprehensive classification of distress prediction models, the reader is referred to Abdou and Pointon (2011), Adnan Aziz and Dar (2010), and Demyanyk and Hasan (2010). Among other corporate events, predicting financial distress of a company as the prior phase of bankruptcy has received extensive attention in recent corporate failure studies (e.g., Çelik et al., 2021; Fich & Slezak, 2008; Geng et al., 2015; Hernandez Tinoco & Wilson, 2013; Hwang et al., 2013; Laitinen & Suvas, 2016; Mousavi & Lin, 2020; Wanke et al., 2015; Zhao & Huchzermeier, 2019).

Financial distress refers to the failure of a company to pay its financial obligations in their due time (Beaver, 1966). The financial situation of distressed firms differs from healthy ones (Cleary & Hebb, 2016; Lau, 1987), suggesting that while a firm financial profile weakens, its features shift toward the characteristics of bankrupt firms. In practice, managers could use a distress

prediction model (hereafter, DPM) as an early warning system to take proper preventive actions to avoid bankruptcy.

A strand of the literature on corporate credit risk and failure prediction has focused on comparing the relative predictive performance of DPMs. The findings of these comparative studies suggest that the relative performance of some DPMs is grounded in either the type of classification algorithm or distinctive features of its design (e.g., Bauer & Agarwal, 2014; Fedorova et al., 2013; Li & Faff, 2019; Mousavi et al., 2015; Wu et al., 2010), or the type of implementation decisions of these algorithms such as selected features (e.g., Back, 2005; Bhimani et al., 2013; Hernandez Tinoco & Wilson, 2013; Tian & Yu, 2017), feature selection technique (e.g., Tsai, 2009; Unler & Murat, 2010), sampling technique (e.g., Gilbert et al., 1990; Neves & Vieira, 2006; Zhou, 2013), and performance evaluation framework (Mousavi et al., 2019; e.g., Mousavi & Ouenniche, 2018).

These comparative studies typically evaluate the predictive performance of DPMs using several performance criteria (e.g., correctness of categorical prediction, discriminatory power, information content, and calibration accuracy) and their measures—for a comprehensive list of evaluation criteria and their measures, the reader is referred to Mousavi et al. (2015).

The prevailing comparative studies are criticized for using arbitrary performance measures and evaluation criteria (Balcaen & Ooghe, 2006) and for being mono-criterion (Bauer & Agarwal, 2014; Mousavi et al., 2015) because, in each round of evaluation, a single measure (e.g., Type I error) of a single criterion (e.g., the accuracy of categorical prediction) is used to evaluate the performance of DPMs; therefore, the rankings of models under different criteria are typically different and as such one cannot make an informed decision about which model performs better under multiple criteria—this problem of inconsistent rankings for different criteria is clearly illustrated in the results of comparative studies by, for example, Theodossiou (1991), Bandyopadhyay (2006), and Hernandez Tinoco and Wilson (2013), which show mixed rankings, as well as in our conceptual illustration of the inconsistency of rankings' problem in Table 1. To deal with this shortcoming, Mousavi et al. (2015) and Mousavi and Ouenniche (2018) proposed to apply multi-criteria evaluation frameworks (i.e., super-efficiency and context-dependent Data Envelopment Analysis, DEA, models) to assess the relative performance of prediction models under multiple criteria and their measures to obtain a single multi-criteria ranking of models and thus avoid dealing with the problem of several inconsistent rankings.

Zavgren (1983) argued that most traditional failure prediction models are based on the underlying assumption that the nexus between the model's dependent variable (i.e., the probability of failure) and its features (e.g., accounting and market information) is constant over time. Empirical studies, however, indicate that this constancy is extremely arguable (Charitou et al., 2004; Du Jardin & Séverin, 2012) and that model's performance is sensitive to fluctuations in macroeconomic circumstances (Mensah, 1984; Platt et al., 1994). For example, the logit model of Ohlson (1980) performed better in the 1980s, while the discrete-time logit prediction model of Shumway (2001) outperformed other models in the 2000s. The changes in patterns of accounting and market-based information over time suggest that prediction models should be re-engineered regularly to encompass the most recent patterns of information (Grice & Ingram, 2001). Consequently, the static performance of prediction models could be far different from their dynamic performance.

In this study, we contend with an overlooked feature of comparative studies of prediction models that lies in the use of what we refer to as a *static performance evaluation framework* or a *static out-of-sample analysis framework* to compare the performance of models in that, first, historical data is split into a fit period dataset and a test

TABLE 1 Illustration of the inconsistency of rankings' problem

	Criteria (C) and their measures (M)				
	C1		C2		
Models	Ranking position under M1	Ranking position under M2	Ranking position under M1	Ranking position under M2	Ranking position under M3
DPM1	1	2	1	2	1
DPM2	2	3	3	1	2
DPM3	3	1	4	2	5
DPM4	4	5	2	4	3
DPM5	5	4	5	5	4

period dataset, then out-of-sample testing is conducted to assess the predictive performance of a DPM. In this paper, we shall refer to a *dynamic performance evaluation framework* or a *dynamic out-of-sample analysis framework* as a framework, which implements a static out-of-sample analysis framework in a dynamic fashion; that is, using a rolling horizon technique. To be more specific, the static out-of-sample analysis framework described above is run several times, each time with a different fit period length, where fit periods are overlapping to comply with a rolling horizon technique such as the fixed-origin rolling horizon technique or the variable-origin rolling horizon technique. The fixed origin rolling horizon technique uses the same historical time period, referred to as the fixed origin, as the starting date of the fit period but increases the length of the fit period by one period at a time in each run; thus, the time horizon is being rolled. By contrast, the variable-origin rolling horizon technique keeps the length of the fit period constant but drops the first period and adds the next period of the historical time horizon from the fit period in each run; thus, again the time horizon is being rolled. Given the above definitions of static and dynamic performance evaluation frameworks, a conceptual comparative analysis of their designs suggests that the static framework ignores the time dynamics in the data, which could bias the performance outlook of the predictive models, on one hand, and neglects the potential need of re-engineering the models as time goes by, on the other hand, whereas the dynamic framework takes account of time dynamics and the corresponding changes in patterns within the data, on one hand, and allows for the re-engineering of the models (i.e., computing revised estimates of the parameters of the models) from one period to the next, on the other hand. Furthermore, the static framework produces a single performance outlook, whereas the dynamic framework produces as many performance outlooks as the number of runs or time horizons considered in the analysis. In this paper, we address the above-mentioned issues with the use of static performance evaluation frameworks by using the variable-origin rolling horizon technique to implement the dynamic out-of-sample analysis framework. In addition, we use the Malmquist-DEA as a dynamic multi-criteria framework for assessing the multi-period performance of DPMs to address the issues with the use of mono-criterion performance evaluation frameworks mentioned in previous paragraphs. In sum, in our proposal, we overcome both the drawbacks of the static out-of-sample analysis framework and the drawbacks of the static multi-criteria performance evaluation frameworks of prediction models. Note that, by design, the proposed Malmquist-DEA framework for the performance evaluation of prediction models under multiple criteria and in a dynamic fashion provides multi-criteria rankings of competing DPMs over multiple periods. In practice, such a framework will not only allow one to highlight the models' performance dynamics over time but also reveal sources of inefficiencies in these models. As far as we know, no previous study has investigated the performance of DPMs using a dynamic multi-criteria assessment framework. However, these types of multi-criteria evaluation frameworks have been used in assessing other types of units in a variety of application areas such as renewable energy (e.g., Zeng et al., 2020; Zeng et al., 2019) kiwifruit production (e.g., Mohammadi et al., 2011), electricity generation (Alizadeh et al., 2020), insurance companies (e.g., Beiragh et al., 2020), supply chains (e.g., Shafiei Kaleibari et al., 2016), and energy efficiency (Houshyar et al., 2010; Mohammadi et al., 2011; Zhou et al., 2008).

Furthermore, this study uses the proposed dynamic multi-criteria evaluation framework to address the following *research questions*:

a. What is the effect of the modeling framework design on the performance of models?
b. What is the effect of the type of information with which models are fed on their performance?
c. What is the effect of the length of the training sample on the models' performance? and
d. Which DPMs perform better in predicting distress over the years with a high distress rate (HDR)?

The rest of the paper is structured as follows. In Section 2, we review other comparative studies of failure prediction models. In Section 3, we describe the proposed dynamic evaluation framework under multiple criteria, namely, Malmquist DEA, and how to operationalize it for our application. In Section 4, we provide details on our experimental design including data, sample selection, and prediction models. In Section 5, we discuss the empirical findings and provide answers to our research questions. Finally, Section 6 concludes the paper.

2 | LITERATURE REVIEW

In this section, we provide a brief survey and classification of the literature on bankruptcy and financial distress prediction along with references to the most cited models and tools and, where appropriate, we critically assess the literature and highlight the gaps.

Recent failure prediction studies have employed non-parametric classifiers form the area of expert system and artificial intelligence such as case-based reasoning (Li &

Sun, 2011; Sartori et al., 2016), neural networks (Hosaka, 2019; Odom & Sharda, 1990; Tsai, 2009; Tsai & Wu, 2008), recursively partitioned decision trees (e.g., Frydman et al., 1985), rough set theory (Ahn et al., 2000; Yeh et al., 2012), genetic programming (e.g., Alfaro-Cid et al., 2007; Etemadi et al., 2009; McKee & Lensberg, 2002), and classifiers from the area of operations research such as data envelopment analysis (e.g., Sun et al., 2014; Shetty et al., 2012; Yeh et al., 2010; Ouenniche & Tone, 2017) and multi-criteria decision-making (e.g., Corazza et al., 2016; Doumpos et al., 2002; Michalis Doumpos & Figueira, 2019; Ouenniche et al., 2017; Ouenniche et al., 2019; Ouenniche et al., 2018).

Statistical classifiers such as discriminant analysis and logit analysis have been very popular in the field of failure and distress prediction and remain very popular as benchmarks in the field. The first generation of statistical models is based on discriminant analysis techniques. Beaver (1966, 1968) is the ground-breaking study that proposed a univariate discriminant analysis to predict bankruptcy. Later, Altman (1968) applied the multivariate discriminant analysis (MDA) to estimate the renowned "Z-score," which is used as a proxy of the financial situation of a company. The MDA technique has been frequently used in later studies (e.g., Altman, 1982; Altman et al., 1977; Blum., 1974; Deakin, 1972; Serrano-Cinca & Gutiérrez-Nieto, 2013).

The majority of subsequent studies have applied the second generation of statistical techniques; that is, the linear probability model (LPM) (Meyer & Pifer, 1970), logit analysis (LA) (Martin, 1977; Ohlson, 1980), and probit analysis (PA) (Zmijewski, 1984). The first and second generations of models could be viewed as empirical models in that they are driven by practical considerations such as an accurate prediction of the risk class or an exact estimate of the probability of belonging to a risk class; in sum, the selection of the explanatory variables is driven by the predictive performance of the models. These models and their application in some previous studies are not without limitations. Some of the assumptions underlying the modeling frameworks may not be reasonably satisfied for some data sets, on one hand, and the earliest studies restricted the type of information to accounting-based one, on the other hand. Also, these models have a static structure and therefore cannot explicitly account for changes over time in the profiles of companies.

The third generation of statistical models is survival analysis (SA) and contingent claims analysis (CCA) models that overcome some of the limitations of the first- and second-generation models. The underlying modeling frameworks of both SA models and CCA models are dynamic by design. To be more specific, SA models are used to estimate time-varying probabilities of failure. Despite the application of SA models in failure prediction dates back to the mid-1980s (e.g., Crapp & Stevenson, 1987; Lane et al., 1986; Luoma & Laitinen, 1991), Shumway (2001) was the pioneering study, which made its use famous by providing an attractive estimation methodology based on an equivalence between multi-period logit models and a discrete-time hazard model. Thereafter, the suggested discrete-time hazard model—also referred to as a discrete-time logit model—was frequently used in later studies (e.g., Bauer & Agarwal, 2014; Chava et al., 2004; Hernandez Tinoco & Wilson, 2013; Mousavi et al., 2015; Wu et al., 2010) to estimate the coefficients of time-varying accounting and market-based covariates of SA models. Unlike the first-generation, the second-generation and SA models, which are empirical, CCA models—also referred to as Black–Scholes–Merton (BSM)-based models—are theoretically grounded. CCA models are grounded in option-pricing theory, as introduced in Black and Scholes (1973) and Merton (1974) whereby the position of an equity holder of a firm is assumed to be the position of long in a call option. Therefore, as suggested by McDonald (2006), the probability of failure could be interpreted as the likelihood that the value of the assets is less than the face value of the liabilities of the firm at maturity; that is, the call option expires worthless. Like any modeling framework, CCA models are not without their limitations. For example, CCA models implicitly assume the same maturities for the liabilities of the firm, which in practice is a limitation (Allen & Saunders, 2004). Also, one might argue that the approximation process of unobservable variables (e.g., expected return, the volatility of return, and market value of assets) is not free of potential measurement errors (Aktug, 2014). Table 2 provides a summary of applied techniques in distress prediction.

Compared to many studies that have focused on proposing new classification techniques, several studies have aimed at investigating the impact of different types of features (i.e., accounting, market, corporate governance, and macroeconomic information) on the performance of models (e.g., Chandrashekar & Sahin, 2014; Lin et al., 2010; Mousavi et al., 2019, 2015; Mousavi & Lin, 2020; Sun et al., 2013; Trujillo-Ponce et al., 2014; Wang et al., 2014). In practice, along with the advancement of prediction techniques, the literature has seen developments in employing new features for developing prediction models. The first and second generation of models, that is, UDA, MDA, LA, and PA, restricted the type of features to accounting-based financial ratios (e.g., Altman, 1968; Beaver, 1968; Begley & Watts, 1997;

TABLE 2 Parametric and non-parametric techniques applied for distress prediction

		Model	Some pioneer studies
Non-parametric	Machine learning and artificial intelligence	Case base reasoning	Li and Sun (2011); Sartori et al. (2016)
		Neural network	Odom and Sharda (1990); Tsai (2009); Tsai and Wu (2008); Hosaka (2019)
		Recursively partitioned decision trees	Frydman et al. (1985)
		Rough set theory	Ahn et al. (2000); Yeh et al. (2012)
		Genetic programming	McKee and Lensberg (2002); Alfaro-Cid et al. (2007); Etemadi et al. (2009)
	Operation research	Data envelopment analysis	Yeh et al. (2010); Shetty et al. (2012); Sun et al. (2014); Ouenniche and Tone (2017)
		Multi-criteria decision making	Doumpos et al. (2002); Corazza et al. (2016); Ouenniche et al. (2017); Ouenniche et al. (2018); Michalis Doumpos and Figueira (2019); Ouenniche et al. (2019)
Parametric	1st generation	Univariate discriminant analysis (UDA)	Beaver (1966, 1968)
		Multivariate discriminant analysis (MDA)	Altman (1968); Deakin (1972); Blum. (1974); Altman et al. (1977); Altman (1982); Serrano-Cinca and Gutiérrez-Nieto (2013)
	2nd generation	Linear probability model (LPM)	Meyer and Pifer (1970)
		Logit analysis (LA)	Martin (1977); Ohlson (1980)
		Probit analysis (PA)	Zmijewski (1984)
	3rd generation	Empirical grounding: Survival analysis (SA)	Shumway (2001); Chava et al. (2004); Wu et al. (2010); Hernandez Tinoco and Wilson (2013); Bauer and Agarwal (2014)
		Theoretical grounding: Contingent claims analysis (CCA)	Hillegeist et al. (2004); Bharath and Shumway (2008)

Note: The table presents different techniques in developing distress prediction models.

Ohlson, 1980; Taffler, 1983; Zmijewski, 1984). The third generation of models, that is, hazard models, and CCA, made use of additional sources of information as features, such as market-based variables (MVs) (e.g., Aktug, 2014; Bharath & Shumway, 2008; Hillegeist et al., 2004), macroeconomic information (MIs) (e.g., Charalambakis & Garrett, 2016; Hernandez Tinoco & Wilson, 2013; Kim & Partington, 2015; Mousavi et al., 2019;, 2015), cultural dimensions (e.g., Laitinen & Suvas, 2016), and corporate governance indicators (CGIs) (e.g., Barboza et al., 2017; Brédart, 2014; Cheng et al., 2014; Hassan Al-Tamimi, 2012; Lee & Yeh, 2004; Lin et al., 2010). The alternative features used in developing distress prediction are presented in Table 3.

Furthermore, several studies have introduced a variety of strategies and methods for identifying the most representative group of features to feed failure prediction models (Balcaen & Ooghe, 2006). On one hand, feature selection strategies could be theoretically grounded such as the features used in CCA models, which are based on option pricing theory (e.g., Bharath & Shumway, 2008; Hillegeist et al., 2004), empirically grounded (e.g, Barboza et al., 2017; Neves & Vieira, 2006; Sun et al., 2011; Unler & Murat, 2010; Zhou et al., 2015), or both (e.g., Laitinen & Suvas, 2016). On the other hand, feature selection methods could be objective or subjective. Objective feature selection methods could be statistical (e.g., Tsai, 2009; Zhou et al., 2012) or non-statistical (e.g., Pacheco et al., 2007; Pacheco et al., 2009) but adopt a common approach, that is, optimizing an effectiveness criterion, whereas subjective feature selection methods make often use of a subjective decision rule including reviewing the literature and selecting the most commonly used features (e.g., Cleary & Hebb, 2016; Zhou, 2013; Zhou et al., 2015). As employing different techniques often results in different sets of selected features, most

studies use one specific feature selection technique. Classifications of alternative grounds and techniques used in feature selection of prediction models are provided in Table 4 and 5, respectively.

Apart from investigating the effect of the classification model or method, the type of information with which models are fed and the type of feature selection technique, the literature on comparative studies suggests that the type of performance criteria and measures, and the chosen evaluation framework could also have a significant impact on the performance outlook of models (Balcaen & Ooghe, 2006; Mousavi et al., 2015; Zhou, 2013). The conventional comparative analyses have used four categories of evaluation criteria including the correctness of categorical prediction, discriminatory power, information content, and calibration accuracy. In practice, the majority of these studies have applied a restricted number of performance measures and criteria (Bauer & Agarwal, 2014)—for example, *type I and type II errors* as measures of the correctness of categorical prediction (e.g., Bauer & Agarwal, 2014; Ben & Youssef, 2018; Collins & Green, 1982; Lennox, 1999; Lo, 1986; Luoma & Laitinen, 1991; Press & Wilson, 1978; Sartori et al., 2016; Theodossiou, 1991), *ROC or Gini index* as a measure of discriminatory power (e.g., Hajek & Henriques, 2017; Hernandez Tinoco & Wilson, 2013; Hillegeist et al., 2004; Hosaka, 2019; e.g., Theodossiou, 1991; Wu et al., 2010), *Pseudo-R^2* and *Log-likelihood* as measures of information content (e.g., Agarwal & Taffler, 2008; Bandyopadhyay, 2006; Bauer & Agarwal, 2014; Li et al., 2010; Theodossiou, 1991), and *Brier score (BS)* as a measure of calibration accuracy (e.g., Theodossiou, 1991)—which

TABLE 3 Alternative features used in distress prediction

Features	Example
Accounting based information	Altman (1968); Beaver (1968); Begley and Watts (1997); Ohlson (1980); Taffler (1983); Zmijewski (1984)
Market-based information	Hillegeist et al. (2004); Bharath and Shumway (2008); Aktug (2014)
Macroeconomic information	Hernandez Tinoco and Wilson (2013); Kim and Partington (2015); Charalambakis and Garrett (2016); Mousavi et al. (2015, 2019)
Cultural dimensions	Laitinen and Suvas (2016)
Corporate governance indicators	Lee and Yeh (2004); Lin et al. (2010); Al-Tamimi (2012); Brédart (2014); Cheng et al. (2014); Barboza et al. (2017)

Note: The table provides a summary of different features used in developing distress prediction models.
Source: Authors' survey of literature.

TABLE 4 Feature selection techniques: Theoretical versus empirical

Feature selection grounds	Explanation	Example
Theoretically grounded	Option pricing theory estimates a value of an options contract by allocating a price, known as a premium, based on the calculated likelihood that the contract will finish in the money at maturity.	Hillegeist et al. (2004); Bharath and Shumway (2008)
Empirically grounded	Selecting the best features using empirical techniques.	Neves and Vieira (2006); Unler and Murat (2010); Sun et al. (2011); Zhou et al. (2015); Barboza et al. (2017)

Note: The table presents two grounds in selecting the best features for distress prediction.
Source: Authors' survey of literature.

TABLE 5 Feature selection techniques: Objective versus subjective

Feature selection technique	Explanation	Example
Objective	Applying statistical techniques such as stepwise regression, factor analysis, and principal component analysis	Tsai (2009); Zhou et al. (2012)
	Using non-statistical and machine learning techniques such as random forest, SVM, and Fisher score.	Pacheco et al. (2007); Pacheco et al. (2009)
Subjective	Use of a subjective decision rule such as reviewing the literature, selecting the most used features, and experts' opinions.	Zhou (2013); Zhou et al. (2015); Cleary and Hebb (2016)

Note: The table presents two techniques in selecting the best features in developing distress prediction models.
Source: Authors' survey of literature.

results in an incomplete assessment of DPMs. Also, conventional studies are criticized for adopting a mono-criterion assessment framework (Mousavi et al., 2015) where a single measure of a single criterion is used at a time to evaluate the relative performance of models and provide mono-criterion rankings. Typically, the rankings associated with different measures and criteria are mostly inconsistent, which results in a situation where users cannot make an informed decision as to which DPM is superior in performance.

Bauer and Agarwal (2014) were the first to use several measures under three performance criteria of discriminatory power, information content, and correctness of categorical prediction to provide a comprehensive mono-criterion assessment comparing the performance of SA model of Shumway (2001), CCA model of Bharath and Shumway (2008), and MDA model of Altman (1968). Mousavi et al. (2015) was the pioneer study that suggests a multi-criteria assessment framework of prediction models of categorical variables or classifiers, that is, super-efficiency DEA to rank the performance of competing statistical models. However, a super-efficiency DEA framework is criticized to be unfair benchmarking since the reference benchmark changes from one distress prediction model evaluation to another (Xu & Ouenniche, 2011). To overcome this methodological issue, Mousavi and Ouenniche (2018) proposed a slacks-based context-dependent DEA model to assess and rank competing DPMs. Further, Mousavi and Lin (2020) applied PROMETHEE[1] multi-criteria decision aid (MCDA) framework to compare the performance of machine learning and artificial intelligence techniques fed with different types of information (i.e., accounting, market, and corporate governance) in distress prediction.

One existing gap in the literature of comparative studies of DPMs is that the application of conventional mono-criterion framework (i.e., one measure under one criterion) and the recently introduced multi-criteria frameworks are restricted by their static nature; therefore, the dynamic performance of models over time has been overlooked—a gap that we address in this paper. To be more specific, our main contribution in this paper is to propose a dynamic or multi-period, multi-criteria performance evaluation framework—see next section, which by design provides multi-criteria rankings of competing DPMs over multiple periods. In practice, such a framework will not only allow one to highlight the models' performance dynamics over time but also reveal sources of inefficiencies in these models. As far as we know, no previous study has investigated the performance of DPMs using a dynamic multi-criteria assessment framework.

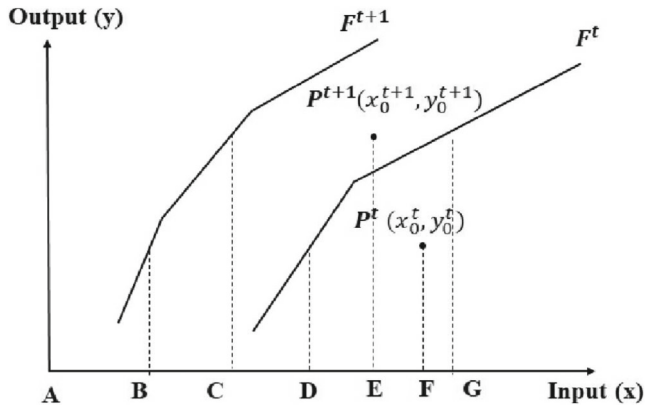

FIGURE 1 Efficiency change and efficient frontier shift

3 | A DYNAMIC FRAMEWORK FOR ASSESSING DPMS

This paper proposes a dynamic multi-criteria assessment framework, namely, Malmquist-DEA, which by design can measure and rank the relative performance of competing DPMs over time. Hereafter, we first present the Malmquist productivity index (MPI) and its estimation using DEA to measure the efficiency of decision-making units (DMUs) (see Section 3.1). Then, we present how one might customize a Malmquist-DEA framework to estimate the relative performance of financial distress prediction models (see Section 3.2).

3.1 | Malmquist productivity index

First introduced by Caves et al. (1982b, 1982a) and continuously improved by other scholars (e.g., Coelli, 1997; Färe et al., 1994), the MPI represents the growth of total factor productivity for a DMU over time—for a review of the applications of MPI, the reader is referred to Färe et al. (2012).

The concepts underlying MPI are illustrated in Figure 1 where a given *DMU*, say, DMU_0, has changed its mix of inputs and outputs or production from $P^t(x_0^t, y_0^t)$ in period t to $P^{t+1}(x_0^{t+1}, y_0^{t+1})$ in period $t+1$, and the efficient frontier, say, F, with respect to which DMU_0 is assessed has changed from the efficient frontier in period t, say, F^t, to the efficient frontier in period $t+1$, say, F^{t+1}, where x_0^t and x_0^{t+1}(respectively, y_0^t and y_0^{t+1}) represent the input (respectively, output) of DMU_0 at time t and $t+1$, respectively. Note that, from a given period t to the next period $t+1$, the efficiency of DMU_0 could change as a result of a change in its efficiency from period t to period $t+1$, as measured by the ratio of the efficiency of DMU_0^{t+1} with respect to the efficient frontier F^{t+1} in period $t+1$ to the efficiency of DMU_0^t with

respect to the efficient frontier F^t in period t. We shall denote such efficiency change from period t to period $t+1$ by $EC^{t,t+1}$, or efficiency change from period t to period $t+1$, which is commonly referred to as the *efficient frontier catch-up or recovery effect* of DMU_0 from period t to period $t+1$ and reflects the degree to which a DMU improves or worsens its efficiency; in sum, $EC^{t,t+1}>1$ indicates progress in relative efficiency from period t to $t+1$; $EC^{t,t+1}=1$ indicates no change or stability from period t to $t+1$, and $EC^{t,t+1}<1$ indicates a regress in efficiency.

On the other hand, from a given period t to the next period $t+1$, the efficiency of DMU_0 could change as a result of a shift in the efficient frontier from period t to period $t+1$ and the relative positions of DMU_0^t (resp. DMU_0^{t+1}) with respect to frontiers F^t and F^{t+1}—this change in efficiency is referred to as the *efficient frontier shift effect* and is computed as the square-root of the product of the efficient frontier shift for DMU_0^t, say EFS^t, and the efficient frontier shift for DMU_0^{t+1}, say, EFS^{t+1}, where EFS^t is measured by the ratio of the efficiency of DMU_0^t with respect to the efficient frontier F^t in period t to the efficiency of DMU_0^t with respect to the efficient frontier F^{t+1} in period $t+1$, and EFS^{t+1} is measured by the ratio of the efficiency of DMU_0^{t+1} with respect to the efficient frontier F^t in period t to the efficiency of DMU_0^{t+1} with respect to the efficient frontier F^{t+1} in period $t+1$. For more specifics about MPI, the reader is referred to Tone (2004) and Färe et al. (1992, 1994).

Hereafter, we outline the procedure for computing the MPI:

Step 1: *Estimating the efficiency change* (*EC*) *of each DMU*

In Figure 1, $\frac{AE}{AC}$ represents the efficiency of DMU_0 with input x^{t+1} and output y^{t+1} with respect to the efficient frontier F^{t+1}, and $\frac{AF}{AD}$ represents the efficiency of DMU_0 with input x^t and output y^t with respect to the efficient frontier F^t. The efficiency change of DMU_0 is measured as follows:

$$EC^{t,t+1}=\frac{AE/AC}{AF/AD}=\frac{\text{Efficiency of } DMU_0^{t+1} \text{ wrt } F^{t+1}}{\text{Efficiency of } DMU_0^{t} \text{ wrt } F^{t}}$$

where $EC>1$, $EC=1$, and $EC<1$ refer to progress, stability, and regress in the relative efficiency, respectively.

Step 2: *Estimating the efficient frontier shift* (*EFS*)

In Figure 1, $\frac{AD}{AF}$ (respectively, $\frac{AB}{AF}$) represents the efficiency of DMU_0 with input x^t and output y^t at time t (respectively, $t+1$) with respect to frontier F^t (respectively, frontier F^{t+1}); therefore, the efficient frontier shift for t, say EFS_t, is measured as follows:

$$EFS^t=\frac{AD/AF}{AB/AF}=\frac{AD}{AB}=\frac{\text{Efficiency of } DMU_k^t \text{ wrt } F^t}{\text{Efficiency of } DMU_k^t \text{ wrt } F^{t+1}}$$

Equivalently, the efficient frontier shift at time $t+1$, say, EFS^{t+1}, is

$$EFS^{t+1}=\frac{AC/AE}{AG/AE}=\frac{AC}{AG}=\frac{\textit{Efficiency of } DMU_k^{t+1} \textit{ wrt } F^t}{\textit{Efficiency of } DMU_k^{t+1} \textit{ wrt } F^{t+1}},$$

where the ratio AC/AE (respectively, AG/AE) demonstrates the efficiency DMU_0 at time $t+1$ with input x^{t+1} and output y^{t+1} with respect to F^{t+1} (respectively, F^t) frontier.

The $EFS^{t,t+1}$ is the efficient frontier shift between time t and $t+1$, which is calculated using the geometric mean of EFS^t and EFS^{t+1}, as follows:

$$EFS^{t,t+1}=\left[EFS^t\times EFS^{t+1}\right]^{\frac{1}{2}}.$$

Step 3: *Estimating the MPI*

MPI is the product of EC and EFS:

$$MPI_0^{t,t+1}=EC^{t,t+1}\times EFS^{t,t+1}.$$

3.2 | Adaptation of MPI for our purpose

MPI is a generic framework and as such its application for our purpose, that is, a dynamic relative performance evaluation of competing DPMs under multiple criteria, requires several key specifications.

First, the choice *of decision-making units or DMUs*: In this study, we are concerned with the performance of distress prediction models or DPMs; thus, *DMUs* are the DPMs (see Section 4.4).

Second, the choice of *inputs and outputs*: In this study, the inputs and outputs are the measures of the performance criteria under consideration; namely, discriminatory power, the correctness of categorical prediction, calibration accuracy, and the information content (for details about performance measures of different criteria, the reader is referred to Mousavi et al., 2015). The outputs (respectively, inputs) are designated based on the principle that the more (respectively, the less), the better; thus, performance measures to be maximized (respectively, minimized) are set as outputs (respectively, inputs).

TABLE 6 Applying SBM-DEA to estimate MPI

Description	Formula
The SBM-DEA model of (Kaoru Tone, 2001, 2011) measures the efficiency of a DMU in a static or single period context; in sum, it measures the distance between a DMU and the efficient frontier that envelops all other DMUs under evaluation. The following SBM model is our proposal for measuring the distance $D^{[1,T]}\left(x_0^{t+1}, y_0^{t+1}\right)$ between DMU_0^{t+1} (DMU_0 observed in period $t+1$) and the global frontier, say $F^{[1,T]}$, where the global frontier $F^{[1,T]}$ envelops all the data points or DMUs observed over the whole horizon of analysis T:	Minimize $D^{[1,T]}\left(x_0^{t+1}, y_0^{t+1}\right) = \frac{\left(1-\frac{1}{m}\sum_{i=1}^{m}\frac{s_{i,0}^{t+1-}}{x_{i,0}^{t+1}}\right)}{\left(1+\frac{1}{s}\sum_{r=1}^{s}\frac{s_{r,0}^{t+1+}}{y_{r,0}^{t+1}}\right)}$ s.t.: $\sum_{j=1}^{n}\lambda_j^q x_{i,j}^q + s_{i,0}^{q-} = x_{i,0}^{t+1};\ i=1,...,m, q=1,...,T$ $\sum_{j=1}^{n}\lambda_j^q y_{r,j}^q - s_{r,0}^{q+} = y_{r,0}^{t+1};\ r=1,...,s, q=1,...,T$ $\sum_{j=1}^{n}\lambda_j^q = 1;\ q=1,...,T$ $\lambda_j^q \geq 0;\ j=1,...,n, q=1,...,T$ $s_{i,0}^{q-} \geq 0;\ i=1,...,m;\ s_{r,0}^{q+} \geq 0;\ r=1,...,s, q=1,...,T$
Through a change of variables similar to the one proposed by Tone (2001) this nonlinear program can be transformed into an equivalent linear program for solution purposes.	

Note: The table explains SBM-DEA method used to estimate MPI.

Third, the choice of the model to estimate efficiency scores or distances from the frontiers: In this study, we opted for the slacks-based measure (SBM) model of Tone (2001) to estimate $EC^{t,t+1}$, EFS^t, and EFS^{t+1}, because of its advantages compared to radial models such as CCR and BCC. For more details on the SBM model, we refer the reader to the eminent work of Tone (2002). Table 6 presents the application of SBM-DEA to estimate MPI.

Fourth, the choice of *reference index*: We aim to measure and compare the relative performance of DPMs over a period; however, the contemporaneous $MPI^{t\ t+1}$ is sensitive to linear programming (LP) infeasibility and represents the performance of DMU_0 from time t to $t+1$; thus, it should be adjusted for our purpose. To deal with this issue, we followed Pastor and Lovell (2005) in estimating the global MPI, which represents the best DMUs over the concerned period (i.e., t_1, t_2, ..., t_n)[2]. Also, in a situation of crossing of the efficient frontiers of different time periods (e.g., F^{t_1} and F^{t_2} in Figure 2), the global MPI can be used as the benchmark reference frontier for all DMUs over the concerned period.For illustration, as Figure 2 indicates, the relative efficiency of DMU_0 can be estimated considering the efficient frontier of period 1 (a combination of DMU_1, DMU_2, DMU_3, DMU_4, and DMU_5) or the efficient frontier of period 2 (combination of DMU_6, DMU_7, DMU_8, DMU_9, and DMU_{10}). Alternatively, the performance of DMU_0 can be measured considering the global frontier, which consists of the best *DMUs* in the past, that is, DMU_3, DMU_4, DMU_5, DMU_6, and DMU_7.

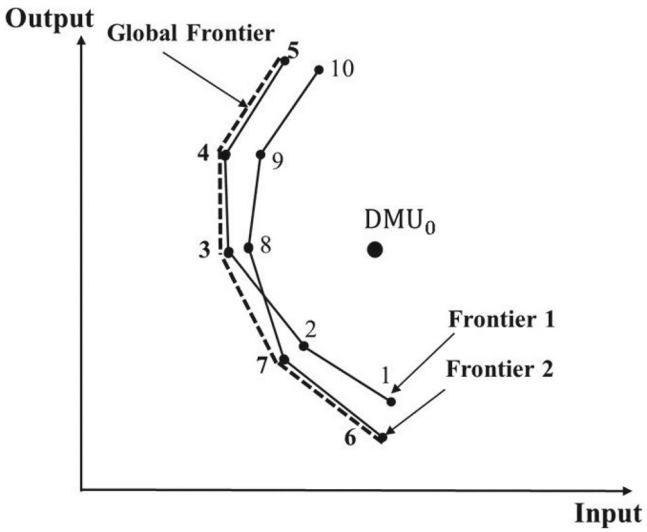

FIGURE 2 Global frontier

4 | EMPIRICAL INVESTIGATION

This section provides details on the data (see Section 4.1), sampling (see Section 4.2), features and features selection procedure (see Section 4.3), and the choice of DPMs (see Section 4.4) for our comparative study to illustrate the use of the proposed dynamic multi-criteria methodology for the relative performance evaluation of classifiers.

4.1 | Data

We took the following steps to select the data set for our empirical analysis. First, we collected the financial accounting and market-based information of all listed companies (except for utility and financial ones) on the London Stock Exchange (LSE) between 1995 and 2014. Second, since developing some models needs minimum historical data, we excluded companies that have been listed for less than 2 years. Third, to minimize any bias related to excluding companies with missing data (Platt & Platt, 2012; Zavgren, 1983; Zmijewski, 1984), we only discarded those companies with missing values for

TABLE 7 Basic sample statistics

Observation (1990–2014)	#	%
Distressed company-year observations (D)	1414	3.2%
Healthy company-year observations	35,570	96.18%
Total company-year observation	36,984	100%

Note: This table presents the total number of distressed companies versus healthy ones for the period between 1990 and 2014.

the main accounting variables (e.g., total assets and sales) and market-based variables (e.g., price), which are essential for computing a variety of accounting and market-based ratios (Lyandres & Zhdanov, 2013; Mousavi & Ouenniche, 2018). The residual missing values of each variable were replaced by its most recently observed ones for each company (Mousavi & Ouenniche, 2018; Zhou et al., 2012). Fourth, to deal with the extreme values of any variable, the outlier values are winsorized between the 1st and 99th percentile of each variable (Shumway, 2001). Fifth, we lagged the data set to make sure that the essential accounting variables are existing in the year in which distress is observed (Bauer & Agarwal, 2014; Mohammadi et al., 2011).

To mark distress firms, we defined a binary variable, say, D, that equals 1 for financially distressed firms and 0 otherwise. We followed Pindado et al. (2008) in classifying a firm as financially distressed if (1) for two succeeding years, the company's interest expense is more than its earnings before interest, taxes, depreciation, and amortization (EBITDA), and (2) for two succeeding years, the company suffers a negative growth in market value. The final sample consists of 3389 firms and 36,984 firm-year observations, of which 1414 firm-year observations qualified as distressed resulting in a distress rate average of 3.82% per year. Table 7 summarizes the number and the proportion of healthy and distress firms in the sample.

4.2 | Sampling

In this study, we test the performance of DPMs out-of-sample. To investigate the effect of the length of the training period on the performance of DPMs, we developed models using two different lengths of sample period: namely, 3 and 5 years. In addition, we applied the moving origin rolling horizon sampling technique (Mousavi & Ouenniche, 2018) to update the parameters of the models as time goes by and thus get rid of the less relevant historical data and include the most recent data to improve the predictive ability of models and reduce any bias due to events that are no more relevant.

To be clearer, we used firm-year observations for n years from $t = n+1$ to t (where $n = 3$, 5 and $1999 < t < 2013$) as training samples to develop models, which then are used to predict distress in year $t+1$ as the holdout sample year. For example, we used firm-year observations for 3 years from 1997 to 1999 as a training sample to predict the distress probability of firms in the year 2000. Considering the 15 years period of our data set and two lengths of the sample period, we end up with 30 training samples. Table 8 presents the details about the proportion of distressed firms for 30 training samples and 15 holdout samples.

4.3 | Features and feature selection procedure

There is a variety of strategies and methods for identifying the most effective group of features to feed failure prediction models (Balcaen & Ooghe, 2006; Neves & Vieira, 2006; Sun et al., 2013; Unler & Murat, 2010). Feature selection strategies could be theoretically grounded, empirically grounded, or both (e.g., Laitinen & Suvas, 2016). On the other hand, feature selection methods could be objective or subjective. Objective feature selection methods could be statistical (e.g., Tsai, 2009; Zhou et al., 2012) or non-statistical (e.g., Pacheco et al., 2007, 2009) but adopt a common approach; that is, optimizing an effectiveness criterion. Whereas subjective feature selection methods make often use of a subjective decision rule including reviewing the literature and selecting the most used features (Mousavi et al., 2015; Ravi Kumar & Ravi, 2007; Zhou et al., 2015). In this research, we used a statistical objective feature selection method.

To be more specific, we reduced our very large initial set of accounting-based ratios (i.e., 83 accounting-based ratios) using factor analysis, where factors are selected based on two criteria: (1) absolute values of loading are more than 0.5 and (2) communities are more than 0.8. This process continued until either no improvement is seen in the total explained variance or no more variables are excluded. Principal component analysis with factor extraction of VARIMAX is used to run this factor analysis (Chen, 2011; Mousavi et al., 2015). Finally, 31 accounting-based ratios with high commonality values and high factor loading, five frequently used market-based information and two mixed ratios (interaction effect of macroeconomic indicators and accounting-based information) were retrained as explanatory variables to be used as inputs into a stepwise procedure of each statistical model. Table 9 represents the final explanatory variables selected for our analysis.

TABLE 8 The proportion of distress firms (*D*) in training and holdout samples

3-year training sample	5-year training sample			Hold out sample	
Years	**Year**	**Year**	***D* %**	**Year**	***D* %**
1997–1999	2.32%	1995–1999	1.79%	2000	1.60%
1998–2000	2.32%	1996–2000	1.96%	2001	1.39%
1999–2001	2.15%	1997–2001	1.99%	2002	6.22%
2000–2002	3.04%	1998–2002	2.89%	2003	11.78%
2001–2003	6.42%	1999–2003	4.82%	2004	3.21%
2002–2004	6.97%	2000–2004	4.77%	2005	2.00%
2003–2005	5.37%	2001–2005	4.76%	2006	3.06%
2004–2006	2.75%	2002–2006	4.99%	2007	4.25%
2005–2007	3.13%	2003–2007	4.62%	2008	5.86%
2006–2008	4.37%	2004–2008	3.69%	2009	10.18%
2007–2009	6.59%	2005–2009	4.94%	2010	4.15%
2008–2010	6.77%	2006–2010	5.41%	2011	1.96%
2009–2011	5.66%	2007–2011	5.37%	2012	5.21%
2010–2012	3.76%	2008–2012	5.63%	2013	8.12%
2011–2013	4.99%	2009–2013	6.01%	2014	5.56%

Note: This table presents the yearly proportion of distress in our training and hold-out samples. The percentage of distress is presented based on the definition of distress (*D*) and two different lengths of the training period.

TABLE 9 List of features

Category	Ratio or item	Category	Ratio or item
Liquidity	Current assets/assets Cash and equivalent/current liabilities Current assets/current liabilities Sales/inventory Current assets/sales Quick assets/current liabilities Current assets/total liabilities Quick assets/inventory Quick assets/assets Inventory/assets Inventory/current assets Net fixed assets/assets Current liabilities/assets	Solvency	Liabilities/sales Long-term and current liabilities/assets Equity/capital The book value of equity/liabilities Net worth/total debt Shareholder's capital/total capital ABD = |1− (fixed assets/equity)|
Market information	Lag (excess return) Lag (sigma) Ln (price) Real size The distress rate in last year	Cash flow	Operating cash flow/liabilities Cash and equivalent/sales Funds provided by operations/liabilities
Asset utilization	Working capital/sales Quick assets/sales	Profitability	Net income/capital Net income/long-term funding Net worth/liabilities ROI × average payment period
Mixed	GDP × Sales Interest rate × Income	Firm characteristics	Ln (age) Log (total assets/GNP price level index)

Note: The table presents the primary features that are used in developing distress prediction models.

To examine the impact of the type of information on the performance of DPMs, we trained models with different combinations of available information sources, that is, financial accounting (FA), market variables (MV), and macroeconomic indicators (MI). Also, note that for a better model fit, we considered the interaction effect of macroeconomic indicators on financial accounting items (Mousavi et al., 2019).

4.4 | The choice of DPMs for evaluation

To evaluate the dynamic performance of the competing DPMs in predicting distress, we considered two general categories of models, namely, static and dynamic, and implemented the most cited models of each category. In terms of static models, we developed two models of MDA and LA. In terms of dynamic models, we followed Nam et al. (2008) in considering two subcategories of models: namely, duration-dependent (DD) and duration-independent (DI) models.

Note that duration-dependent DPMs comprise a time-dependent baseline rate, which could be estimated using historical information of the firm or be represented by a time-dummy or by a time-varying feature of the firm. To this end, we used Kim and Partington's (2015) approach in estimating the time-dependent baseline hazard rate for each firm using firms' historical information in a Cox hazard duration model. We refer to this model as duration-dependent with the firm's specific baseline rate (DDWFSB). Also, since employing an indirect measure of the baseline rate such as time dummies (Beck et al., 1998) is less efficient, we followed Gupta et al. (2015), Nam et al. (2008), and Shumway (2001) in using the time-varying feature of firm's age to proxy the time-dependent baseline rate in a discrete-time duration model. We refer to this model as duration-dependent with a time-dependent baseline rate (DDWTDB).

Also, note that based on containing a constant (time-independent) baseline rate, the duration-independent models are classified into two subgroups, namely duration-independent without baseline (DIWOTIB) and duration-independent with time-independent baseline (DIWTIB). In this study, we use the natural logarithm of firm age, *ln* (*age*), as the time-independent baseline rate.

Considering two static and four dynamic modeling frameworks, two lengths of the training samples, 15 hold-out samples, and three combinations of features, we

TABLE 10 New developed models fed with 3-year training samples using FAMVMI information

Explanatory variables	Model 1	Model 2	Model 3	Model 4*	Model 5	Model 6
Intercept	−2.94	−3.44	−1.53	$-1.53+ln(age)_i$	−1.9749	
Current assets to liabilities	−0.002					
Net income to long term funding	−0.014					0.0022
Current assets to sales	−.0002					
Total liabilities to assets	0.066					−0.0079
Cash and equivalent to sales	0.006					0.0003
Inventory to assets						−0.9997
Equity to sales	−0.0003					
Lag of excess return	−1.289	−0.832***	−0.987***	−0.987***	−1.001***	−0.985***
Lag of sigma	2.865					
Ln (price)	−3.842	−0.281***	−0.217***	−0.217***	−0.218***	−0.226***
Equity to capital		0.045				
Current liabilities to assets		0.038				
Real size		−0.2562***				−0.111
Ohlson size			−0.197***	−0.197***	−0.201***	
Interest rate × net income	−0.0001	−0.0006***	−0.00007***	−0.00007***	−0.00007***	−0.00004***
GDP × sales	−0.0001	−0.0004***	−0.0001***	−0.0001***	−0.0001***	−0.0001***
Ln (age)					0.185	

Note: The table presents the features and coefficients of the new models, namely, MDA (1), LA (2), DIWOTIB (3), DIWTIB-ln (age) (4), DDWTDB-ln (age) (5), and DDWFSB (6). In model 4, *ln* (*age*) of firm *i* is added to the intercept as the baseline hazard rate.

***1% significance level.

**5% significance level.

ended up with 540 developed models. A set of newly developed models using 3-year (from 2011 to 2013) FAMVMI information is presented in Table 10. Appendix A provides more details about the models developed in this study.

5 | DYNAMIC AND MULTI-CRITERIA ASSESSMENT OF DISTRESS PREDICTION MODELS

This section summarizes our implementation decisions of the proposed dynamic multi-criteria framework for assessing the performance of DPMs, which we used to address our research questions—see Section 1.

To illustrate the use of the proposed framework and highlight its advantages, we developed six statistical frameworks, namely, MDA, LA, DIWOTIB, DIWTIB-ln (age), DDWTDB-ln (age), and DDWFSB using three combinations of information, that is, FA, MVMI, FAMVMI, and two lengths of training periods, that is, 3- and 5-year training samples.

We tested the prediction accuracy of the DPMs developed on 15 consecutive hold-out samples (from the year 2000 to 2014) using measures under four criteria, that is, the correctness of categorical prediction, discriminatory power, information content, and calibration accuracy. For a detailed presentation of performance criteria and their measures, the reader is referred to Mousavi et al. (2015).

The estimated performance measures are used as inputs and outputs in Malmquist DEA. The Malmquist DEA framework provides the global efficiency scores for each model each year. We used the average scores of 15 hold-out samples as the overall efficiency score of each model. Also, to answer the research question "Which DPMs perform better in predicting distress over the years with high distress rate (HDR)?," we used the average scores of the HDR years, that is, 2003, 2008, and 2013.

As mentioned above, the advantage of the multi-criteria framework is that it facilitates taking multiple performance criteria into account, which results in a comprehensive performance evaluation, provides a single multi-criterion ranking in each period, and facilitates the presentation and monitoring of the performance of models over time. Depending on practitioners' preferences, alternative measures could be selected for each criterion. Section 5.1 presents the yearly and total 15-year rankings of prediction models extracted from the first round (i.e., a combination of measures of the criteria under consideration) of assessment—the other 11 rounds of ranking are not presented here for the sake of saving space. Section 5.2 presents the overall rankings of prediction models considering 12 rounds of assessments.

5.1 | Round I of rankings (inputs: T1, BS; outputs: Pseudo-R^2, ROC)

In the first round of dynamic multi-criteria assessment, we use Type 1 (T1) error (to measure the accuracy of categorical prediction) and BS (to measure the calibration accuracy) as inputs, and Pseudo-R^2 (to measure the information content) and ROC (to measure the discriminatory power) as outputs. Table 11 presents the first-round rankings of DPMs based on the estimated efficiency scores for each year (i.e., 2000 to 2014). Using these dynamic rankings or the corresponding scores, one could analyze the performance of a DPM overtime. To compare the overall or aggregate performance of each model over the 15-year period, we calculate an aggregate ranking based on the average scores over 15 years. Also, to compare the performance of DPMs over years with high distress rates, we calculate an aggregate HDR ranking based on the average scores over years 2003, 2008, and 2013. In sum, Table 11 provides the rankings of DPMs for each year referred to as yearly rankings, the aggregate rankings over 5-year periods (i.e., 2000–2004; 2005–2009; 2010–2014) and over 15-year period (i.e., 2000–2014), and the aggregate rankings over HDR years (i.e., 2003, 2008 and 2013) using Dynamic Multi-Criteria Evaluation Framework, where the aggregate rankings are based on averages of the Malmquist DEA scores of the models.

Figure 3 illustrates the difference in overall performance between static and dynamic DPMs as well as the effect on the performance of these models resulting from the use of different categories of information or features, different rolling horizon lengths and a combination of these. Notice that Panel 1 of Figure 3 shows the superiority of dynamic models' performance over static ones. The results also suggest that the models developed in duration-dependent frameworks, that is, DDWFSB and DDWTDB-ln (age), outperform duration-independent and static models, and among static models, LA outperforms MDA. On the other hand, Panel 2 of Figure 3 indicates that models fed with FAMVMI features outperform those fed with MVMI and FA, respectively. These results are consistent across all models. In addition, Panel 3 of Figure 3 shows that the dynamic (respectively, static) models developed using 5-year (respectively, 3-year) information outperform the dynamic (respectively, static) ones using 3-year (respectively, 5-year) information. Finally, Panel 4 of Figure 3 indicates that most of the dynamic models fed with 5-year features of FAMVMI and FA outperform those fed with 3-year features. On the contrary, most of the dynamic models fed with 3-year MVMI outperform 5-year MVMI. For the robustness of our findings, we performed 12 other experiments or rounds using a variety of combinations of inputs and

TABLE 11 Ranking of models using one combination of inputs and outputs

			Yearly rankings								
Framework	**Feature**	**Training period**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
DDWFSB	FAMVMI	5	1	9	4	6	2	8	1	2	11
DDWFSB	FAMVMI	3	4	23	12	21	1	14	3	23	22
DDWTDB-ln (age)	FAMVMI	3	2	16	14	13	7	1	2	6	8
DDWTDB-ln (age)	FAMVMI	5	6	5	3	2	4	4	14	3	1
DIWTIB-ln (age)	FAMVMI	3	3	10	10	7	15	5	4	8	7
DDWFSB	MVMI	3	8	22	21	22	3	21	10	24	17
DIWTIB-ln (age)	FAMVMI	5	7	6	1	3	9	3	13	5	4
DIWOTIB	FAMVMI	5	18	1	5	4	6	2	19	14	5
DDWFSB	MVMI	5	10	2	36	1	36	32	9	1	13
DDWTDB-ln (age)	MVMI	3	14	13	14	13	10	9	5	9	12
DIWTIB-ln (age)	MVMI	3	5	3	10	7	23	7	7	11	14
LA	FAMVMI	3	11	18	13	15	11	10	6	13	16
DDWTDB-ln (age)	MVMI	5	9	15	6	11	5	15	17	4	3
LA	FAMVMI	5	12	14	7	12	8	13	16	12	2
DIWOTIB	FAMVMI	3	13	19	16	16	16	16	8	15	15
DIWTIB-ln (age)	MVMI	5	17	7	2	5	19	6	15	7	6
DIWOTIB	MVMI	3	15	20	16	16	14	11	12	19	19
LA	MVMI	3	15	20	16	16	17	11	11	18	18
LA	MVMI	5	19	11	8	9	12	17	22	16	9
DIWOTIB	MVMI	5	19	11	8	9	12	18	21	16	9
MDA	FAMVMI	3	22	24	22	23	20	19	18	22	24
MDA	FAMVMI	5	21	25	23	24	18	20	23	21	20
MDA	MVMI	5	23	26	19	19	21	22	24	10	21
MDA	MVMI	3	24	27	20	20	22	23	20	20	23
DDWTDB-ln (age)	FA	5	26	8	24	25	24	24	25	25	25
DIWOTIB	FA	5	28	17	26	26	25	25	26	27	26
DDWFSB	FA	5	25	4	35	29	35	36	32	36	36

TABLE 11 (Continued)

Framework	Feature	Training period	Yearly rankings								
			2000	2001	2002	2003	2004	2005	2006	2007	2008
DIWTIB-ln (age)	FA	5	27	28	25	27	27	26	27	26	27
MDA	FA	3	29	30	28	28	28	30	31	28	34
LA	FA	3	30	32	29	34	31	31	34	32	29
DDWTDB-ln (age)	FA	3	33	33	32	30	29	28	29	29	28
DIWOTIB	FA	3	34	34	33	31	30	29	30	30	29
LA	FA	5	31	29	31	36	34	35	36	31	31
DDWFSB	FA	3	35	36	27	35	26	27	28	34	32
MDA	FA	5	32	31	30	33	32	33	33	33	33
DIWTIB-ln (age)	FA	3	36	35	34	32	33	34	35	35	35

Note: The table presents the rankings of DPMs yearly, aggregately over 5- and 15-year periods, and aggregately over HDR years using dynamic multi-criteria evaluation framework for one combination of inputs and outputs (inputs: T1 and BS; outputs: ROC and Pseudo-R^2).

TABLE 11 (Continued)

Framework	Yearly rankings						Aggregate rankings over 5, 15, and HDR year periods				
	2009	2010	2011	2012	2013	2014	2000–2004	2005–2009	2010–2014	15 years	HDR years
DDWFSB	2	18	1	1	11	2	1	2	1	1	4
DDWFSB	11	23	3	2	12	10	2	9	4	2	19
DDWTDB-ln (age)	3	10	10	5	1	9	8	1	8	3	2
DDWTDB-ln (age)	5	5	13	10	13	17	3	4	15	4	5
DIWTIB-ln (age)	4	15	20	7	6	18	5	3	17	5	2
DDWFSB	16	24	4	3	7	1	7	20	3	6	20
DIWTIB-ln (age)	6	17	21	13	19	19	4	5	21	7	8
DIWOTIB	15	2	8	16	16	15	10	6	12	8	11
DDWFSB	1	19	2	4	10	3	20	17	2	9	1
DDWTDB-ln (age)	8	14	15	9	3	7	12	8	10	10	6
DIWTIB-ln (age)	10	22	23	19	9	16	6	7	22	11	7
LA	14	13	16	8	8	6	16	10	9	12	13
DDWTDB-ln (age)	7	9	17	14	15	13	9	13	16	13	10

TABLE 11 (Continued)

	Yearly rankings						Aggregate rankings over 5, 15, and HDR year periods				
Framework	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2000–2004**	**2005–2009**	**2010–2014**	**15 years**	**HDR years**
LA	12	1	7	15	14	14	11	16	11	14	14
DIWOTIB	13	6	9	6	2	8	17	12	7	15	9
DIWTIB-ln (age)	9	20	22	20	21	20	13	11	23	16	11
DIWOTIB	19	7	5	11	4	4	18	15	5	17	15
LA	20	8	6	11	5	4	19	14	6	18	16
LA	17	3	11	17	17	11	14	18	13	19	17
DIWOTIB	17	3	11	17	17	11	14	19	13	20	17
MDA	23	21	24	21	20	21	21	21	24	21	22
MDA	22	11	19	23	23	23	22	22	20	22	24
MDA	21	12	18	24	24	24	25	23	19	23	21
MDA	24	16	14	22	22	22	26	24	18	24	22
DDWTDB-ln (age)	25	25	25	25	25	25	23	25	25	25	25
DIWOTIB	27	26	26	26	26	26	27	26	26	26	26
DDWFSB	32	27	31	28	32	29	24	36	28	27	30
DIWTIB-ln (age)	26	28	27	27	27	27	28	27	27	28	27
MDA	34	32	34	34	31	32	29	31	34	29	30
LA	31	30	30	31	34	35	31	32	33	30	33
DDWTDB-ln (age)	29	33	28	32	28	28	33	29	29	31	28
DIWOTIB	30	31	29	33	29	31	34	30	30	32	29
LA	33	29	33	30	30	33	30	34	32	33	33
DDWFSB	28	34	32	29	33	30	35	28	31	34	32
MDA	36	35	36	35	35	34	32	33	35	35	36
DIWTIB-ln (age)	35	36	35	36	36	36	36	35	36	36	35

Note: The table presents the rankings of DPMs yearly, aggregately over 5- and 15-year periods, and aggregately over HDR years using dynamic multi-criteria evaluation framework for one combination of inputs and outputs (inputs: T1 and BS; outputs: ROC and Pseudo-R^2).

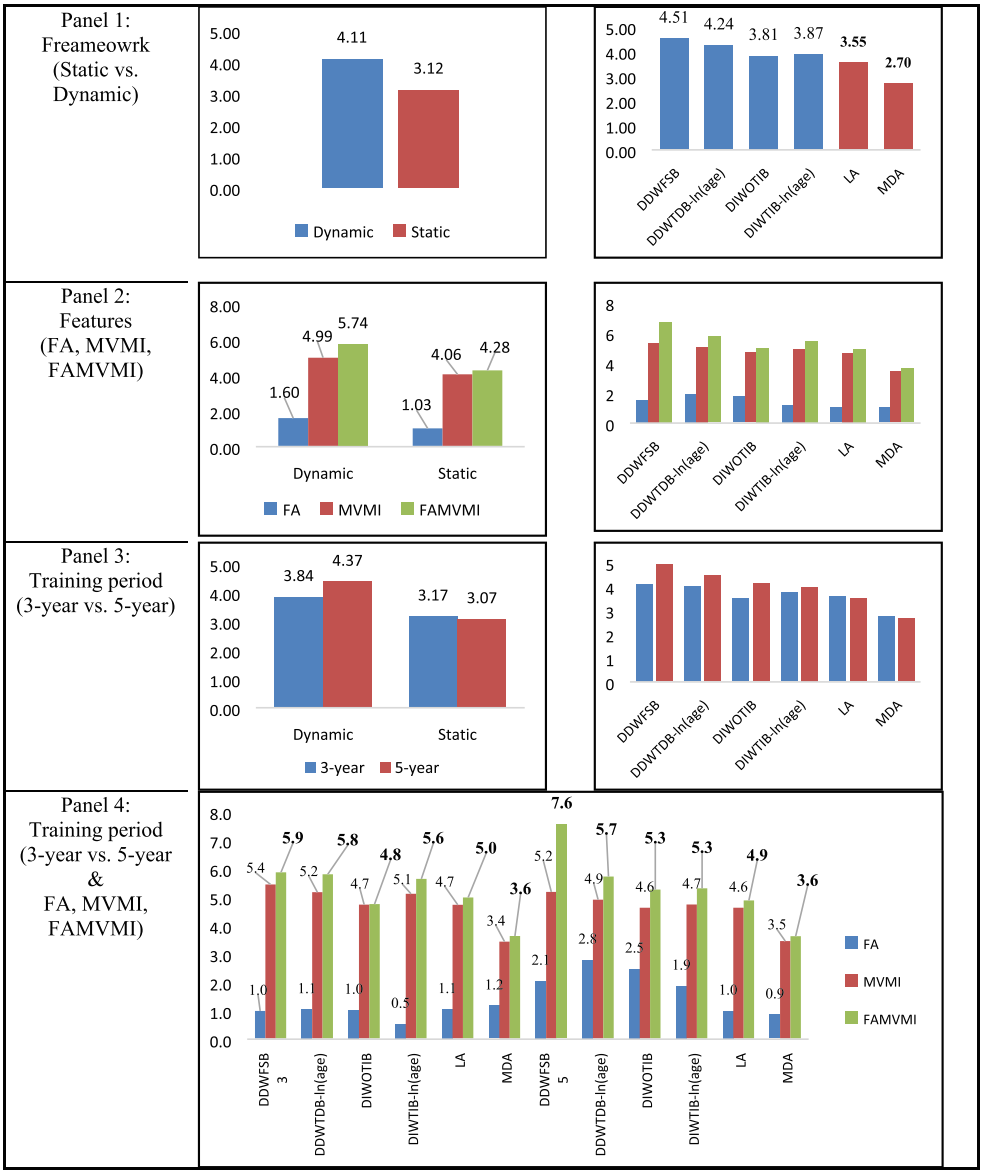

FIGURE 3 Dynamic rank score of models in the first round of assessment using multi-criteria evaluation framework (inputs: T1, BS, outputs: ROC, Pseudo-R^2)

output or equivalently measures of the four performance criteria under consideration. The findings based on Panel 3 of Figure 3 of this combination of inputs and outputs are rather an outlier compared to the findings of the remaining combinations, which are summarized in the next subsection.

5.2 | The average of rankings for all rounds (combination of different in- puts and outputs)

We expanded our dynamic multi-criteria assessment to 12 different rounds (i.e., combinations of measures of the criteria under consideration), where a single measure is used for each criterion to void any implicit assignment of a higher weight to any of the criteria. Table 12 summarizes the inputs and outputs used in each round and provides aggregate rankings over 15-year period for each combination of inputs and outputs—referred to as AIR-15; the aggregate ranking of these individual combinations of inputs and outputs—referred to as AR-15; and the aggregate ranking of individual combinations of inputs and outputs over HDR years—referred to as AR-HDR—of DPMs using the proposed dynamic multi-criteria evaluation framework, where the aggregate rankings are based on averages of the Malmquist DEA scores of the models. In addition, Figure 4 illustrates the difference in overall performance between static and dynamic DPMs based on the AR-15 rankings of Table 12 as well as the effect on the performance of these models resulting from the use of different categories of information or

TABLE 12 Aggregate rankings over 15-year period and HDR years for each individual combination of inputs and outputs

			AIR-15													
Framework	**Feature**	**Training period**	**T1, BS H, R^2**	**T2, BS H, R^2**	**BS H, R^2, OCC**	**MR, BS H, R^2**	**T1, BS ROC, R^2**	**T2, BS ROC, R^2**	**BS ROC, R^2, OCC**	MR, BS ROC, R^2	**T1, BS KS, R^2**	**T2, BS KS, R^2**	**BS KS, R^2, OCC**	**MR, BS KS, R^2**	**AIR-15**	**AR-HDR**
DDWFSB	FAMVMI	3	2	1	1	1	2	1	1	1	4	1	1	1	1	14
DDWTDB-ln (age)	FAMVMI	3	4	3	5	3	3	3	2	2	3	3	3	2	2	1
DDWFSB	MVMI	3	5	2	3	2	6	2	4	4	2	2	4	3	3	13
DIWTIB-ln (age)	FAMVMI	3	6	4	2	4	5	4	3	3	6	4	2	4	4	5
DDWTDB-ln (age)	FAMVMI	5	3	6	6	6	4	5	5	5	5	7	6	7	5	9
DIWTIB-ln (age)	FAMVMI	5	8	5	4	5	7	6	6	6	8	8	5	9	6	15
DDWTDB-ln (age)	MVMI	3	10	7	7	7	10	8	8	8	9	6	8	6	7	2
DIWOTIB	FAMVMI	3	17	8	8	8	15	7	9	7	13	5	9	5	8	3
LA	FAMVMI	3	11	9	9	9	12	11	11	10	10	9	10	8	9	4
DIWTIB-ln (age)	MVMI	3	14	11	10	10	11	9	7	9	12	10	7	10	10	10
DDWTDB-ln (age)	MVMI	5	13	12	12	12	13	12	12	12	11	12	12	12	11	8
DIWTIB-ln (age)	MVMI	5	19	10	11	11	16	10	10	11	18	11	11	11	12	18
LA	FAMVMI	5	12	13	13	13	14	13	13	13	15	13	13	13	13	11
DIWOTIB	FAMVMI	5	7	16	14	16	8	16	15	16	7	16	14	16	14	12
LA	MVMI	3	16	14	15	14	18	14	14	14	17	14	15	14	15	7
DIWOTIB	MVMI	3	15	15	16	15	17	15	16	15	16	15	16	15	16	6
LA	MVMI	5	20	18	18	18	19	17	17	17	19	17	17	17	17	17
DDWFSB	FAMVMI	5	1	24	23	24	1	23	23	23	1	24	23	24	17	21
DIWOTIB	MVMI	5	18	17	17	17	20	18	18	18	21	18	18	18	19	15
MDA	FAMVMI	3	21	19	19	19	21	19	19	19	20	19	19	19	20	23
MDA	FAMVMI	5	22	21	21	21	22	21	21	21	22	20	20	20	21	24
MDA	MVMI	3	24	20	20	20	24	20	20	20	24	21	21	21	22	19

(Continues)

TABLE 12 (Continued)

Framework	Feature	Training period	AIR-15												AIR-15	AR-HDR
			T1, BS H, R^2	T2, BS H, R^2	BS H, R^2, OCC	MR, BS H, R^2	T1, BS ROC, R^2	T2, BS ROC, R^2	BS ROC, R^2, OCC	MR, BS ROC, R^2	T1, BS KS, R^2	T2, BS KS, R^2	BS KS, R^2, OCC	MR, BS KS, R^2		
MDA	MVMI	5	23	22	22	22	23	22	22	22	23	22	22	22	23	22
DDWFSB	MVMI	5	9	26	26	26	9	27	27	27	14	27	26	27	24	20
DDWTDB-ln (age)	FA	5	25	25	24	25	25	25	24	25	25	25	24	25	25	25
DIWOTIB	FA	5	26	27	25	27	26	26	25	26	26	26	25	26	26	26
DDWFSB	FA	3	30	23	28	23	34	24	28	24	32	23	28	23	27	29
DIWTIB-ln (age)	FA	5	28	28	27	28	28	28	26	28	28	28	27	28	28	27
MDA	FA	3	29	29	29	29	29	29	29	29	29	29	29	29	29	33
LA	FA	3	31	30	30	30	30	30	30	30	30	30	30	30	30	32
DDWTDB-ln (age)	FA	3	33	33	32	32	31	31	31	31	31	32	31	32	31	30
LA	FA	5	32	31	31	31	33	33	33	33	33	31	32	31	32	31
DIWOTIB	FA	3	34	32	33	33	32	32	32	32	34	33	33	33	33	28
DDWFSB	FA	5	27	35	35	35	27	35	36	35	27	35	36	35	34	34
MDA	FA	5	35	34	34	34	35	34	34	34	35	34	34	34	35	35
DIWTIB-ln (age)	FA	3	36	36	36	36	36	36	35	36	36	36	35	36	36	36

Note: The table presents the DPMs' aggregate ranking of individual combinations of Inputs and Outputs (AR-15), and aggregate ranking of individual combinations of inputs and outputs over HDR years (AR-HDR) using dynamic multi-criteria evaluation framework.

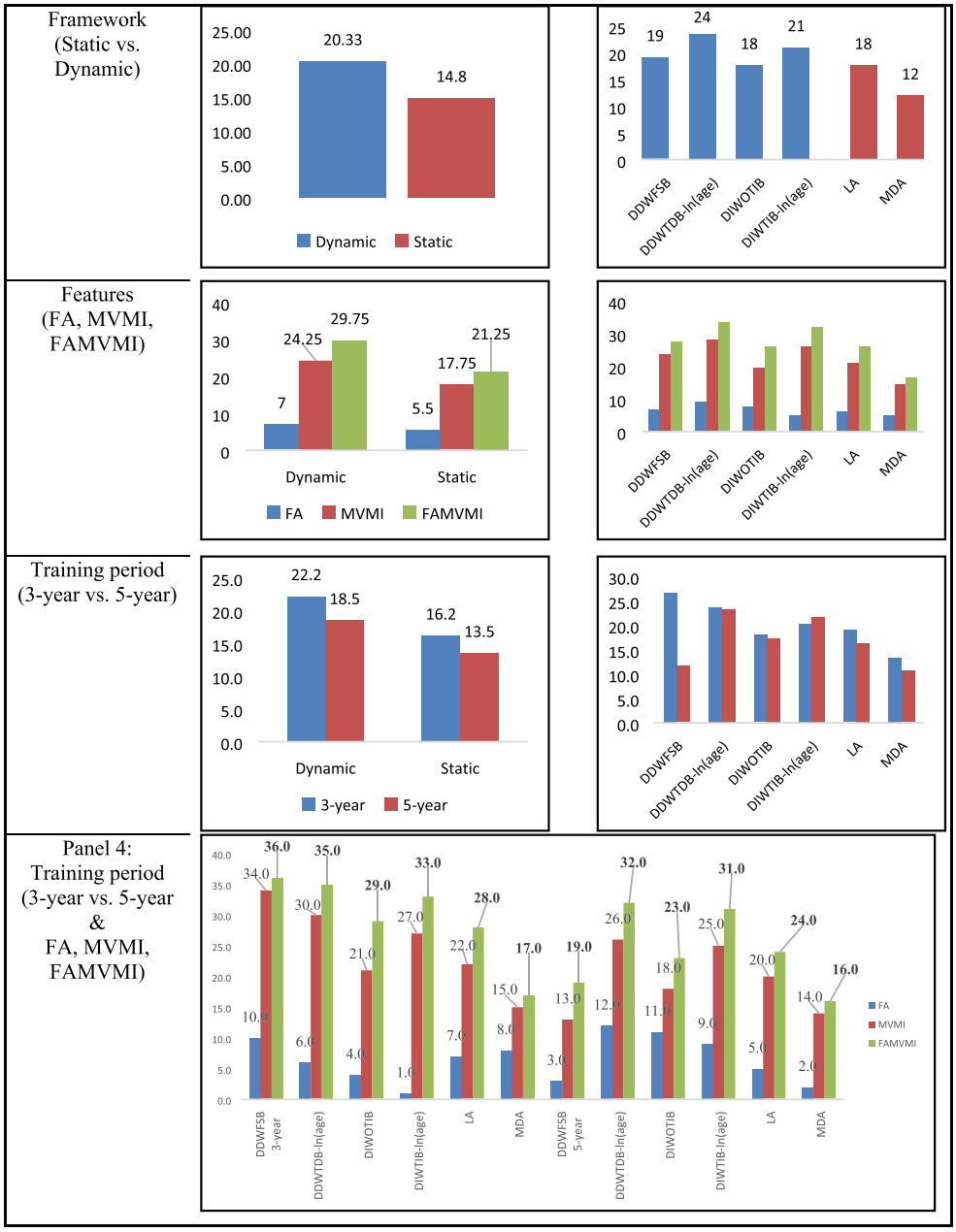

FIGURE 4 Total dynamic rank score of models using 12 rounds of assessment (alternative combinations of inputs and outputs)

features, different rolling horizon lengths and a combination of these. Our main findings based on the analysis of this information could be summarized as follows.

First, Panel 1 shows the superiority of dynamic models' performance over static ones, where the performance of the models is aggregated across all categories of information and time horizons. The results also suggest that the models developed in duration-dependent frameworks, that is, DDWFSB and DDWTDB-ln (age), outperform duration-independent and static models, and among static models, LA outperforms MDA. These findings answer the question: What is the effect of the modeling framework design on the performance of models? Our findings are consistent with other studies (Bauer & Agarwal, 2014; Mousavi & Ouenniche, 2018; Nam et al., 2008; Shumway, 2001), which indicate that the duration models utilize information more effectively in DPMs. Further, the results suggest that models developed in the duration-dependent framework of DDWFSB followed by DDWTDB-ln (age) outperform other dynamic and static frameworks. Our finding is consistent with Nam et al. (2008) and Kim and Partington (2015) proposed that using the baseline rate in the distress models improves their performance. Further, compatible with other research such as Bauer and Agarwal (2014) and Wu et al. (2010), our results suggest that the LA model outperforms MDA.

Second, Panel 2 indicates that the models fed with FAMVMI features outperform those fed with either MVMI or FA, regardless of the frameworks under which they are developed. In addition, models fed with MVMI features outperform those fed with FA. These findings answer the question: What is the effect of the type of information with which models are fed on their performance? Our findings support Shumway (2001), Agarwal and Taffler (2008) and Bauer and Agarwal (2014), who suggest that using more information enhances the performance of failure prediction models.

Third, Panel 3 shows that both static and dynamic models developed using 3-year information outperform the ones using 5-year information, where the performance of the models is aggregated across all categories of information, except for DIWTIB-ln (age) for which the nature of the performance is reversed because of not taking account of time dynamics. These findings answer the question: What is the effect of the length of the training sample on the models' performance?

Fourth, Panel 4 indicates that static models fed with 3-year features—whether they belong to category FA, category MVMI or category FAMVMI—outperform those fed with 5-year features; that is, shorter horizons seem to enhance their aggregate performance, and this is consistent across all categories of information with which the static models are fed. As to the dynamic models, when fed with features belonging to category MVMI or category FAMVMI, shorter horizons seem to enhance their aggregate performance. However, when dynamic models are fed with features belonging to the FA category, longer horizons seem to enhance their aggregate performance except for model DDWFSB. These findings answer the question: What is the effect of the length of the training sample on the models' performance?

Finally, regarding which DPMs perform better in predicting distress over the years with high distress rate (HDR), examination of the rankings of models in HDR years, that is, 2003, 2009, and 2013, year by year and over all these years, we found out that DDWTDB-ln (age) fed with FAMVMI and FAMV, respectively, are the best models for predicting distress during HDR years. Furthermore, LA fed with FAMVMI shows very good performance during HDR years despite its static nature.

6 | CONCLUSION

Prediction of corporate financial distress is crucial for many stakeholders and decision-makers in finance and investment. Although many models have been developed to forecast bankruptcy and financial distress, the relative performance assessment of competing DPMs remained in practice an exercise, which is both mono-criterion and static. The mono-criterion nature of comparative studies is criticized because of conflicts in rankings of models from one performance criterion or measure to another. The static framework of comparative studies does not support any monitoring of the performance of models over time.

In this study, we proposed a dynamic multi-criteria framework, based on Malmquist DEA, to evaluate the performance of DPMs. This framework provides a multi-criteria ranking per period that allows the monitoring of DPMs' performance over time. The multi-period ranking takes account of the variations of inputs and outputs over time (e.g., trends) and as such provides a fair evaluation of models. Also, a multi-period ranking of DPMs allows one to assess the robustness of models' predictive power and accuracy during different periods and business cycles. In sum, the proposed performance evaluation framework of DPMs is a powerful tool for monitoring the performance of models over time, highlighting any impact of specific events (e.g., economic recessions) on the performance of DPMs, and suggesting how the DPMs, as specified by specific sets of explanatory variables, that became outdated or less relevant could be changed.

Also, we performed a comparative analysis of the most cited static and dynamic DPMs, which are developed using different combinations of information. Different rounds of evaluation are conducted using several combinations of measures in the four categories of performance criteria, that is, discriminatory power, information content, calibration accuracy, and correctness of categorical predictions. Our main findings could be summarized as follows.

First, the aggregate performance, across all categories of information and time horizons, shows that (a) dynamic models are superior to static ones and (b) dynamic models developed under duration-dependent frameworks outperform dynamic models developed under duration-independent frameworks and static models. These findings answer one of our research questions, namely, what is the effect of the modeling framework design on the performance of models?

Second, the aggregate performance, across all time horizons, shows that models fed with FAMVMI features outperform those fed with either MVMI or FA, regardless of the frameworks under which they are developed. In addition, models fed with MVMI features outperform those fed with FA. These findings answer another one of our research questions; namely, what is the effect of the type of information with which models are fed on their performance?

Third, the aggregate performance of models, across all categories of information, shows that shorter horizons

seem to enhance the aggregate performance of both static and dynamic models, except for DIWTIB-ln (age) for which an excessively poor performance when fed with FA information compared to when fed with either MVMI or FAMVMI information explains its poor aggregated performance over a shorter horizon. These findings answer another one of our research questions, namely, what is the effect of the length of the training sample on the models' performance?

Fourth, the aggregate performance of models, across all combinations of inputs and outputs, shows that (a) shorter horizons seem to enhance the aggregate performance of static models and this behavior is consistent across all categories of information with which the static models are fed and (b) shorter horizons seem to enhance the aggregate performance of dynamic models fed with features belonging to category MVMI or category FAMVMI; however, when fed with features belonging to the FA category, longer horizons seem to enhance their aggregate performance with the exception of model DDWFSB. These findings also answer our research question: what is the effect of the length of the training sample on the models' performance?

Finally, regarding which DPMs perform better in predicting distress over the years with a high distress rate (HDR), we found out that DDWTDB-ln (age) fed with FAMVMI and FAMV, respectively, are the best models for predicting distress during HDR years. Furthermore, LA fed with FAMVMI shows very good performance during HDR years despite its static nature.

One of the limitations of this research is the space constraint and as such, we restricted this study to financial distress as an event. Future research could investigate other definitions of failure such as bankruptcy and debt restructuring. Moreover, this study is restricted in terms of data (i.e., listed companies on LSE), and types of models (i.e., statistical models). Future studies could incorporate machine learning and artificial intelligence techniques. Also, future studies could analyze the extent to which failure prediction models are generalizable by considering the data from other countries or stock exchanges.

ACKNOWLEDGMENTS

The authors would like to thank reviewers for their thoughtful comments and efforts towards improving our paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Mohammad Mahdi Mousavi https://orcid.org/0000-0001-9721-5072

ENDNOTE

[1] Preference Ranking Organization METHod for Enrichment of Evaluation.

REFERENCES

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2–3), 59–88. https://doi.org/10.1002/isaf.325

Adnan Aziz, M., & Dar, H. A. (2010). Predicting corporate bankruptcy: Where we stand? *Corporate Governance: The International Journal of Business in Society*, *6*(1), 18–33. https://doi.org/10.1108/14720700610649436

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking and Finance*, *32*(8), 1541–1551. https://doi.org/10.1016/j.jbankfin.2007.07.014

Ahn, B. S., Cho, S. S., & Kim, C. Y. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Applications*, *18*(2), 65–74. https://doi.org/10.1016/S0957-4174(99)00053-6

Aktug, R. E. (2014). A critique of the contingent claims approach to sovereign risk analysis. *Emerging Markets Finance and Trade*, *50*(sup 1), 294–308. https://doi.org/10.2753/REE1540-496X5001S118

Alfaro-Cid, E., Sharman, K., & Esparcia-Alcazar, A. (2007). A genetic programming approach for bankruptcy prediction using a highly unbalanced database. In M. Giacobini (Ed.), *Workshops on applications of evolutionary computation* (pp. 169–178). Springer. https://doi.org/10.1007/978-3-540-71805-5_19

Alizadeh, R., Gharizadeh Beiragh, R., Soltanisehat, L., Soltanzadeh, E., & Lund, P. D. (2020). Performance evaluation of complex electricity generation systems: A dynamic network-based data envelopment analysis approach. *Energy Economics*, *91*, 104894. https://doi.org/10.1016/j.eneco.2020.104894

Allen, L., & Saunders, A. (2004). Incorporating systemic influences into risk measurements: A survey of the literature. *Journal of Financial Services Research*, *26*(2), 161–191. https://doi.org/10.1023/B:FINA.0000037545.38154.8a

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I. (1982). Accounting implications of failure prediction models. *Journal of Accounting, Auditing & Finance*, *6*(1), 4–19.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETATManalysis a new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, *1*(1), 29–54. https://doi.org/10.1016/0378-4266(77)90017-6

Andersen, P. K. (1992). Repeated assessment of risk factors in survival analysis. *Statistical Methods in Medical Research*, *1*(3), 297–315.

Back, P. (2005). Explaining financial difficulties based on previous payment behavior, management background variables and financial ratios. *European Accounting Review*, *14*(4), 839–868. https://doi.org/10.1080/09638180500141339

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *British Accounting Review*, *38*(1), 63–93. https://doi.org/10.1016/j.bar.2005.09.001

Bandyopadhyay, A. (2006). Predicting probability of default of Indian corporate bonds: Logistic and Z-score model approaches. *Journal of Risk Finance*, *7*(3), 255–272. https://doi.org/10.1108/15265940610664942

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Bauer, J., & Agarwal, V. (2014). Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. *Journal of Banking and Finance*, *40*(1), 432–442. https://doi.org/10.1016/j.jbankfin.2013.12.013

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, *4*, 71–111. https://doi.org/10.2307/2490171

Beaver, W. H. (1968). Alternative accounting measures as predictors of failure. *The Accounting Review*, *43*(1), 113–122.

Beck, N., Katz, J. N., & Tucker, R. (1998). Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science*, *42*(4), 1260–1288. https://doi.org/10.2307/2991857

Begley, J., & Watts, S. (1997). Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models. *Review of Accounting Studies*, *1*(812), 267–284.

Beiragh, R. G., Alizadeh, R., Kaleibari, S. S., Cavallaro, F., Zolfani, S. H., Bausys, R., & Mardani, A. (2020). An integrated multi-criteria decision making model for sustainability performance assessment for insurance companies. *Sustainability*, *12*(3), 789. https://doi.org/10.3390/su12030789

Ben, S., & Youssef, J. (2018). Forecasting financial distress for French firms: A comparative study. *Empirical Economics*, *54*(3), 1173–1186. https://doi.org/10.1007/s00181-017-1246-1

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies*, *21*(3), 1339–1369. https://doi.org/10.1093/rfs/hhn044

Bhimani, A., Gulamhussen, M. A., & Lopes, S. d. R. (2013). The role of financial, macroeconomic, and non-financial information in Bank loan default timing prediction. *European Accounting Review*, *22*(4), 739–763. https://doi.org/10.1080/09638180.2013.770967

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654. https://doi.org/10.1086/260062

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research*, *12*(1), 1–25. https://doi.org/10.2307/2490525

Brédart, X. (2014). Financial distress and corporate governance: The impact of board configuration. *International Business Research*, *7*(3), 72–80. https://doi.org/10.5539/ibr.v7n3p72

Caves, D. W., Christensen, L. R., & Diewert, W. E. (1982a). Multilateral comparisons of output, input, and productivity using superlative index numbers. *The Economic Journal*, *92*(365), 73–86. https://doi.org/10.2307/2232257

Caves, D. W., Christensen, L. R., & Diewert, W. E. (1982b). The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica*, *50*(6), 1393. https://doi.org/10.2307/1913388

Çelik, Ş., Aktan, B., & Burton, B. (2021). Firm dynamics and bankruptcy processes: A new theoretical model. *Journal of Forecasting*, *41*, 567–591. https://doi.org/10.1002/for.2826

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Charalambakis, E. C., & Garrett, I. (2016). On the prediction of financial distress in developed and emerging markets: Does the choice of accounting and market information matter? A comparison of UK and Indian firms. *Review of Quantitative Finance and Accounting*, *47*(1), 1–28. https://doi.org/10.1007/s11156-014-0492-y

Charitou, A., Neophytou, E., & Charalambous, C. (2004). Predicting corporate failure: Empirical evidence for the UK. *European Accounting Review*, *13*(3), 465–497. https://doi.org/10.1080/0963818042000216811

Chava, S., Jarrow, R. A., & August, R. (2004). Bankruptcy prediction with industry effects. *Financial Management*, *8*(4), 537–569.

Chen, L. S., Yen, M. F., Wu, H. M., Liao, C. S., Liou, D. M., Kuo, H. S., & Chen, T. H. H. (2005). Predictive survival model with time-dependent prognostic factors: Development of computer-aided SAS Macro program. *Journal of Evaluation in Clinical Practice*, *11*(2), 181–193.

Chen, M. Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications*, *38*(9), 11261–11272. https://doi.org/10.1016/j.eswa.2011.02.173

Cheng, B., Ioannou, I., & Serafeim, G. (2014). Corporate social responsibility and access to finance. *Strategic Management Journal*, *35*(1), 1–23. https://doi.org/10.1002/smj.2131

Cleary, S., & Hebb, G. (2016). An efficient and functional model for predicting bank distress: In and out of sample evidence. *Journal of Banking and Finance*, *64*, 101–111. https://doi.org/10.1016/j.jbankfin.2015.12.001

Coelli, T. J. (1997). Total factor productivity growth in Australian coal-fired electricity generation: A Malmquist index approach. In international conference on public sector efficiency, UNSW, Sydney (*27*, p. 28).

Collins, R. A., & Green, R. D. (1982). Statistical methods for bankruptcy forecasting. *Journal of Economics and Business*, *34*(4), 349–354. https://doi.org/10.1016/0148-6195(82)90040-6

Corazza, M., Funari, S., & Gusso, R. (2016). Creditworthiness evaluation of Italian SMEs at the beginning of the 2007-2008 crisis: An MCDA approach. *North American Journal of Economics and Finance*, *38*, 1–26. https://doi.org/10.1016/j.najef.2016.05.008

Crapp, H. R., & Stevenson, M. (1987). Development of a method to assess the relevant variables and the probability of financial distress. *Australian Journal of Management*, *12*(2), 221–236. https://doi.org/10.1177/031289628701200205

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, *10*(1), 167. https://doi.org/10.2307/2490225

Demyanyk, Y., & Hasan, I. (2010). Financial crises and bank failures: A review of prediction methods. *Omega*, *38*(5), 315–324. https://doi.org/10.1016/j.omega.2009.09.007

Doumpos, M., Kosmidou, K., Baourakis, G., & Zopounidis, C. (2002). Credit risk assessment using a multicriteria hierarchical discrimination approach. *European Journal of Operational Research*, *138*, 392–412. https://doi.org/10.1016/S0377-2217(01)00254-5

Doumpos, M., & Figueira, J. R. (2019). A multicriteria outranking approach for modeling corporate credit ratings: An application of the Electre tri-nC method. *Omega*, *82*, 166–180. https://doi.org/10.1016/j.omega.2018.01.003

Du Jardin, P., & Séverin, E. (2012). Forecasting financial failure using a Kohonen map: A comparative study to improve model stability over time. *European Journal of Operational Research*, *221*(2), 378–396. https://doi.org/10.1016/j.ejor.2012.04.006

Etemadi, H., Anvary Rostamy, A. A., & Dehkordi, H. F. (2009). A genetic programming model for bankruptcy prediction: Empirical evidence from Iran. *Expert Systems with Applications*, *36*(2), 3199–3207. https://doi.org/10.1016/j.eswa.2008.01.012

Färe, R., Grosskopf, S., Norris, M., & Zhang, Z. (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *The American Economic Review*, *84*(1), 66–83. https://www.jstor.org/stable/2117971

Färe, R., Grosskopf, S., & Russell, R. R. (2012). *Index numbers: Essays in honour of Sten Malmquist*. Springer Science & Business Media.

Färe, R., Lindgren, B., & Roos, P. (1992). Productivity changes in Swedish pharmacies 1980-1989: A non- parametric malmquist approach. *The Journal of Productivity Analysis*, *3*(1–2), 85–101. https://doi.org/10.1007/BF00158770

Fedorova, E., Gilenko, E., & Dovzhenko, S. (2013). Bankruptcy prediction for Russian companies: Application of combined classifiers. *Expert Systems with Applications*, *40*(18), 7285–7293. https://doi.org/10.1016/j.eswa.2013.07.032

Fich, E. M., & Slezak, S. L. (2008). Can corporate governance save distressed firms from bankruptcy? An empirical analysis. *Review of Quantitative Finance and Accounting*, *30*(2), 225–251. https://doi.org/10.1007/s11156-007-0048-5

Frydman, H., Altman, E. I., & Kao, D. L. (1985). Introducing recursive portioning for financial classification: The case of financial distress. *Journal of Finance*, *40*(1), 269–291. https://doi.org/10.1111/j.1540-6261.1985.tb04949.x

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, *241*(1), 236–247. https://doi.org/10.1016/j.ejor.2014.08.016

Gilbert, L. R., Menon, K., & Schwartz, K. B. (1990). Predicting bankruptcy for firms in financial distress. *Journal of Business Finance & Accounting*, *17*(1), 161–171. https://doi.org/10.1111/j.1468-5957.1990.tb00555.x

Grice, J. S., & Ingram, R. W. (2001). Tests of the generalizability of Altman's bankruptcy prediction model. *Journal of Business Research*, *54*(1), 53–61. https://doi.org/10.1016/S0148-2963(00)00126-0

Gupta, J., Gregoriou, A., & Healy, J. (2015). Forecasting bankruptcy for SMEs using hazard function: To what extent does size matter? *Review of Quantitative Finance and Accounting*, *45*(4), 845–869. https://doi.org/10.1007/s11156-014-0458-0

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001

Hassan Al-Tamimi, H. A. (2012). The effects of corporate governance on performance and financial distress: The experience of UAE national banks. *Journal of Financial Regulation and Compliance*, *20*(2), 169–181. https://doi.org/10.1108/13581981211218315

Hernandez Tinoco, M., & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, *30*, 394–419. https://doi.org/10.1016/j.irfa.2013.02.013

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, *9*(1), 5–34. https://doi.org/10.1023/B:RAST.0000013627.90884.b7

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications*, *117*, 287–299. https://doi.org/10.1016/j.eswa.2018.09.039

Houshyar, E., Sheikh Davoodi, M. J., & Nassiri, S. M. (2010). Energy efficiency for wheat production using data envelopment analysis (DEA) technique. *Journal of Agricultural Technology*, *6*(4), 663–672.

Hwang, R. C., Chung, H., & Ku, J. Y. (2013). Predicting recurrent financial distresses with autocorrelation structure: An empirical analysis from an emerging market. *Journal of Financial Services Research*, *43*(3), 321–341. https://doi.org/10.1007/s10693-012-0136-0

Kim, M. H., & Partington, G. (2015). Dynamic forecasts of financial distress of Australian firms. *Australian Journal of Management*, *40*(1), 135–160. https://doi.org/10.1177/0312896213514237

Laitinen, E. K., & Suvas, A. (2016). Financial distress prediction in an international context: Moderating effects of Hofstede's original cultural dimensions. *Journal of Behavioral and Experimental Finance*, *9*, 98–118. https://doi.org/10.1016/j.jbef.2015.11.003

Lane, W. R., Looney, S. W., & Wansley, J. W. (1986). An application of the cox proportional hazards model to bank failure. *Journal of Banking and Finance*, *10*(4), 511–531. https://doi.org/10.1016/S0378-4266(86)80003-6

Lau, A. H.-L. (1987). A five-state financial distress prediction model. *Journal of Accounting Research*, *25*(1), 127–138. https://doi.org/10.2307/2491262

Lee, T.-S., & Yeh, Y.-H. (2004). Corporate governance and financial distress: Evidence from Taiwan. *Corporate Governance*, *12*(3), 378–388. https://doi.org/10.1111/j.1467-8683.2004.00379.x

Lennox, C. (1999). Identifying failing companies: A re-evaluation of the logit, probit and DA approaches. *Journal of Economics and Business*, *51*(4), 347–364. https://doi.org/10.1016/S0148-6195(99)00009-0

Li, H., & Sun, J. (2011). Predicting business failure using forward ranking-order case-based reasoning. *Expert Systems with Applications*, *38*(4), 3075–3084. https://doi.org/10.1016/j.eswa.2010.08.098

Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications*, *37*(8), 5895–5904. https://doi.org/10.1016/j.eswa.2010.02.016

Li, L., & Faff, R. (2019). Predicting corporate bankruptcy: What matters? *International Review of Economics & Finance*, *62*, 1–19. https://doi.org/10.1016/j.iref.2019.02.016

Lin, F., Liang, D., & Chu, W. S. (2010). The role of non-financial features related to corporate governance in business crisis prediction. *Journal of Marine Science and Technology*, *18*(4), 504–513. https://doi.org/10.51400/2709-6998.1901

Lo, A. W. (1986). Logit versus discriminant analysis. A specification test and application to corporate bankruptcies. *Journal of Econometrics*, *31*(2), 151–178. https://doi.org/10.1016/0304-4076(86)90046-1

Luoma, M., & Laitinen, E. (1991). Survival analysis as a tool for company failure prediction. *Omega*, *19*(6), 673–678. https://doi.org/10.1016/0305-0483(91)90015-L

Lyandres, E., & Zhdanov, A. (2013). Investment opportunities and bankruptcy prediction. *Journal of Financial Markets*, *16*(3), 439–476. https://doi.org/10.1016/j.finmar.2012.10.003

Martin, D. (1977). Early warning of bank failure. A logit regression approach. *Journal of Banking and Finance*, *1*(3), 249–276. https://doi.org/10.1016/0378-4266(77)90022-X

McDonald, R. L. (2006). *Derivative Markets, 2nd*. Addison-Wesley, Pearson Education.

McKee, T. E., & Lensberg, T. (2002). Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European Journal of Operational Research*, *138*(2), 436–451. https://doi.org/10.1016/S0377-2217(01)00130-8

Mensah, Y. M. (1984). An examination of the stationarity of multivariate bankruptcy prediction models: A methodological study. *Journal of Accounting Research*, *22*(1), 380. https://doi.org/10.2307/2490719

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, *29*(2), 449–470.

Meyer, P. A., & Pifer, H. W. (1970). Prediction of Bank failures. *The Journal of Finance*, *25*(4), 853–868. https://doi.org/10.1111/j.1540-6261.1970.tb00558.x

Mohammadi, A., Rafiee, S., Mohtasebi, S. S., Avval, S. H. M., & Rafiee, H. (2011). Energy efficiency improvement and input cost saving in kiwifruit production using data envelopment analysis approach. *Renewable Energy*, *36*(9), 2573–2579. https://doi.org/10.1016/j.renene.2010.10.036

Mousavi, M. M., Ouenniche, J., & Tone, K. (2019). A comparative analysis of two-stage distress prediction models. *Expert Systems with Applications*, *119*, 323–341. https://doi.org/10.1016/j.eswa.2018.10.053

Mousavi, M. M., & Ouenniche, J. (2018). Multi-criteria ranking of corporate distress prediction models: Empirical evaluation and methodological contributions. *Annals of Operations Research*, *271*, 853–886. https://doi.org/10.1007/s10479-018-2814-2

Mousavi, M. M., Ouenniche, J., & Xu, B. (2015). Performance evaluation of bankruptcy prediction models: An orientation-free super-efficiency DEA-based framework. *International Review of Financial Analysis*, *42*, 64–75. https://doi.org/10.1016/j.irfa.2015.01.006

Mousavi, M. M., & Lin, J. (2020). The application of PROMETHEE multi-criteria decision aid in financial decision making: Case of distress prediction models evaluation. *Expert Systems with Applications*, *159*, 113438. https://doi.org/10.1016/j.eswa.2020.113438

Nam, C. W., Kim, T. S., Park, N. J., & Lee, H. K. (2008). Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting*, *27*(6), 493–506. https://doi.org/10.1002/for.985

Neves, J. C., & Vieira, A. (2006). Improving bankruptcy prediction with hidden layer learning vector quantization. *European Accounting Review*, *15*(2), 253–271. https://doi.org/10.1080/09638180600555016

Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. In 1990 IJCNN International Joint Conference on Neural Networks (pp. 163–168 vol.2).

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131. https://doi.org/10.2307/2490395

Ouenniche, J., Bouslah, K., Cabello, J. M., & Ruiz, F. (2017). A new classifier based on the reference point method with application in bankruptcy prediction. *Journal of the Operational Research Society*, *69*(10), 1653–1660. https://doi.org/10.1057/s41274-017-0254-z

Ouenniche, J., Bouslah, K., Perez-Gladish, B., & Xu, B. (2019). A new VIKOR-based in-sample-out-of-sample classifier with application in bankruptcy prediction. *Annals of Operations Research*, *296*, 495–512. https://doi.org/10.1007/s10479-019-03223-0

Ouenniche, J., Pérez-Gladish, B., & Bouslah, K. (2018). An out-of-sample framework for TOPSIS-based classifiers with application in bankruptcy prediction. *Technological Forecasting and Social Change*, *131*, 111–116. https://doi.org/10.1016/j.techfore.2017.05.034

Ouenniche, J., & Tone, K. (2017). An out-of-sample evaluation framework for DEA with application in bankruptcy prediction. *Annals of Operations Research*, *254*(1–2), 235–250. https://doi.org/10.1007/s10479-017-2431-5

Pacheco, J., Casado, S., & Nuñez, L. (2007). Use of VNS and TS in classification: Variable selection and determination of the linear discrimination function coefficients. *IMA Journal of Management Mathematics*, *18*(2), 191–206. https://doi.org/10.1093/imaman/dpm012

Pacheco, J., Casado, S., & Núñez, L. (2009). A variable selection method based on Tabu search for logistic regression models. *European Journal of Operational Research*, *199*(2), 506–511. https://doi.org/10.1016/j.ejor.2008.10.007

Pastor, J. T., & Lovell, C. A. K. (2005). A global Malmquist productivity index. *Economics Letters*, *88*(2), 266–271. https://doi.org/10.1016/j.econlet.2005.02.013

Pindado, J., Rodrigues, L., & de la Torre, C. (2008). Estimating financial distress likelihood. *Journal of Business Research*, *61*(9), 995–1003. https://doi.org/10.1016/j.jbusres.2007.10.006

Platt, H., & Platt, M. (2012). Corporate board attributes and bankruptcy. *Journal of Business Research*, *65*(8), 1139–1143. https://doi.org/10.1016/j.jbusres.2011.08.003

Platt, H. D., Platt, M. B., & Pedersen, J. G. (1994). Bankruptcy discrimination with real variables. *Journal of Business Finance & Accounting*, *21*(4), 491–510. https://doi.org/10.1111/j.1468-5957.1994.tb00332.x

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*(364), 699–705. https://doi.org/10.1080/01621459.1978.10480080

Ravi Kumar, P., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review. *European Journal of Operational Research*, *180*(1), 1–28. https://doi.org/10.1016/j.ejor.2006.08.043

Sartori, F., Mazzucchelli, A., & Gregorio, A. D. (2016). Bankruptcy forecasting using case-based reasoning: The CRePERIE approach. *Expert Systems with Applications*, *64*, 400–411. https://doi.org/10.1016/j.eswa.2016.07.033

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2013). Partial Least Square discriminant analysis for bankruptcy prediction. *Decision Support Systems*, *54*(3), 1245–1255. https://doi.org/10.1016/j.dss.2012.11.015

Shafiei Kaleibari, S., Gharizadeh Beiragh, R., Alizadeh, R., & Solimanpur, M. (2016). A framework for performance evaluation of energy supply chain by a compatible network data envelopment analysis model. *Scientia Iranica*, *23*(4), 1904–1917. https://doi.org/10.24200/sci.2016.3936

Shetty, U., Pakkala, T. P. M., & Mallikarjunappa, T. (2012). A modified directional distance formulation of DEA to assess bankruptcy: An application to IT/ITES companies in India. *Expert Systems with Applications*, *39*(2), 1988–1997. https://doi.org/10.1016/j.eswa.2011.08.043

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple Hazard model. *The Journal of Business*, *74*(1), 101–124. https://doi.org/10.1086/209665

Sun, J., Jia, M. Y., & Li, H. (2011). AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. *Expert Systems with Applications*, *38*(8), 9305–9312. https://doi.org/10.1016/j.eswa.2011.01.042

Sun, J., Li, H., Huang, Q. H., & He, K. Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, *57*, 41–56. https://doi.org/10.1016/j.knosys.2013.12.006

Sun, X., Liu, Y., Xu, M., Chen, H., Han, J., & Wang, K. (2013). Feature selection using dynamic weights for classification. *Knowledge-Based Systems*, *37*, 541–549. https://doi.org/10.1016/j.knosys.2012.10.001

Taffler, R. J. (1983). The assessment of company solvency and performance using a statistical model. *Accounting and Business Research*, *13*(52), 295–308. https://doi.org/10.1080/00014788.1983.9729767

Theodossiou, P. (1991). Alternative models for assessing the financial condition of business in Greece. *Journal of Business Finance & Accounting*, *18*(5), 697–720. https://doi.org/10.1111/j.1468-5957.1991.tb00233.x

Tian, S., & Yu, Y. (2017). Financial ratios and bankruptcy predictions: An international evidence. *International Review of Economics & Finance*, *51*, 510–526. https://doi.org/10.1016/j.iref.2017.07.025

Tone, K. (2004). Malmquist Productivity Index: Efficiency Change Over Time. In W. W. Cooper, L. M. Seiford, & J. Zhu (Eds.), 2004*Handbook on data envelopment analysis*. Kluwer Academic Publishers.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, *130*(3), 498-509. https://doi.org/10.1016/S0377-2217(99)00407-5

Tone, K. (2002). A slacks-based measure of super-efficiency in data envelopment analysis. *European Journal of Operational Research*, *143*, 32–41. https://doi.org/10.1016/S0377-2217(01)00324-1

Tone, K. (2011). Slacks-based measure of efficiency. *International Series in Operations Research and Management Science*, *164*, 195–209. https://doi.org/10.1007/978-1-4419-6151-8_8

Trujillo-Ponce, A., Samaniego-Medina, R., & Cardone-Riportella, C. (2014). Examining what best explains corporate credit risk: Accounting-based *versus* market-based models. *Journal of Business Economics and Management*, *15*(2), 253–276. https://doi.org/10.3846/16111699.2012.720598

Tsai, C. F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, *22*(2), 120–127. https://doi.org/10.1016/j.knosys.2008.08.002

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(4), 2639–2649. https://doi.org/10.1016/j.eswa.2007.05.019

Unler, A., & Murat, A. (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research*, *206*(3), 528–539. https://doi.org/10.1016/j.ejor.2010.02.032

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, *41*(5), 2353–2361. https://doi.org/10.1016/j.eswa.2013.09.033

Wanke, P., Barros, C. P., & Faria, J. R. (2015). Financial distress drivers in Brazilian banks: A dynamic slacks approach. *European Journal of Operational Research*, *240*(1), 258–268. https://doi.org/10.1016/j.ejor.2014.06.044

Wu, Y., Gaunt, C., & Gray, S. (2010). A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting and Economics*, *6*(1), 34–45. https://doi.org/10.1016/j.jcae.2010.04.002

Xu, B., & Ouenniche, J. (2011). A multidimensional framework for performance evaluation of forecasting models: Context-dependent DEA. *Applied Financial Economics*, *21*(24), 1873–1890. https://doi.org/10.1080/09603107.2011.597722

Yeh, C. C., Chi, D. J., & Hsu, M. F. (2010). A hybrid approach of DEA, rough set and support vector machines for business failure prediction. *Expert Systems with Applications*, *37*(2), 1535–1541. https://doi.org/10.1016/j.eswa.2009.06.088

Yeh, C.-C., Lin, F., & Hsu, C.-Y. (2012). A hybrid KMV model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems*, *33*, 166–172. https://doi.org/10.1016/j.knosys.2012.04.004

Zavgren, C. V. (1983). *Corporate failure prediction: The state of the art*. Institute for Research in the behavioral, economic, and

management sciences, Krannert graduate School of Management, Purdue University.

Zeng, Y., Guo, W., Wang, H., & Zhang, F. (2020). A two-stage evaluation and optimization method for renewable energy development based on data envelopment analysis. *Applied Energy*, *262*, 114363. https://doi.org/10.1016/j.apenergy.2019.114363

Zeng, Y., Guo, W., & Zhang, F. (2019). Comprehensive evaluation of renewable energy technical plans based on data envelopment analysis. *Energy Procedia*, *158*, 3583–3588. https://doi.org/10.1016/j.egypro.2019.01.907

Zhao, L., & Huchzermeier, A. (2019). Managing supplier financial distress with advance payment discount and purchase order financing. *Omega*, *88*, 77–90. https://doi.org/10.1016/j.omega.2018.10.019

Zhou, L. (2013). Predicting the removal of special treatment or delisting risk warning for listed company in China with adaboost. In. *Procedia Computer Science*, *17*, 633–640. https://doi.org/10.1016/j.procs.2013.05.082

Zhou, L., Lai, K. K., & Yen, J. (2012). Empirical models based on features ranking techniques for corporate financial distress prediction. *Computers and Mathematics with Applications*, *64*(8), 2484–2496. https://doi.org/10.1016/j.camwa.2012.06.003

Zhou, L., Lu, D., & Fujita, H. (2015). The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. *Knowledge-Based Systems*, *85*, 52–61. https://doi.org/10.1016/j.knosys.2015.04.017

Zhou, P., Ang, B. W., & Poh, K.-L. (2008). A survey of data envelopment analysis in energy and environmental studies. *European Journal of Operational Research*, *189*(1), 1–18. https://doi.org/10.1016/j.ejor.2007.04.042

Zmijewski, M. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 59–82. https://doi.org/10.2307/2490859

AUTHOR BIOGRAPHIES

Dr Mohammad Mahdi Mousavi is an Associate Professor of Finance at the University of Bradford (UK). His research interest consists of a variety of topics, including the design and performance evaluation of bankruptcy prediction models, credit scoring, corporate finance, application of AI, and machine learning in Finance, and international finance.

Prof. Jamal Ouenniche is a Professor of Business Analytics at the University of Edinburgh Business School (UK). His research portfolio encompasses a range of applications and a variety of research methodologies in descriptive, predictive, and prescriptive analytics and tackles important managerial issues in energy, manufacturing, transport, banking, and public sector policy.

Prof. Kaoru Tone is a Professor Emeritus at GRIPS in Japan. He has served as a professor at Saitama University and Keio University for over 40 years. He was the President of the Operations Research Society of Japan, from 1996 to 1998. His contribution to DEA has a variety of attainments

How to cite this article: Mousavi, M. M., Ouenniche, J., & Tone, K. (2023). A dynamic performance evaluation of distress prediction models. *Journal of Forecasting*, *42*(4), 756–784. https://doi.org/10.1002/for.2915

APPENDIX A

Table below provides details of statistical models developed in this study.

Framework	Explanation
Multiple discriminant analysis (MDA)	The common form of discriminant analysis (DA) model for the group k (assuming n groups) could be revealed as follows. $z_k = f\left(\sum_{j=1}^{p} \beta_{kj} x_j\right)$ Equation 1 where x_j denotes the discriminant features j, β_{kj} denotes the discriminant coefficients of feature j for group k, z_k denotes the score of group k, and f is the classifier (linear or nonlinear) that maps the estimated scores of $\beta^t x$ onto a set of real numbers. We followed Hillegeist et al. (2004) to convert estimated scores to the probability of distress using a logit link as follows. $P(distress)_i = \frac{e^z}{1+e^z}$ Equation 2
Logit analysis (LA)	The common model of logit analysis for binary variables could be defined as follows: $\begin{cases} P(distress)_i = P(Y=1) \\ P(distress)_i = G(\beta, X) \end{cases}$ Equation 3 where Y represents the binary dependent variable, X represents the vector of features, β is the vector of coefficients of X in the model, and $G(.)$ represents a link function that maps the estimated scores of $\beta^t x$, onto a probability. Practically, the link function determines the type of probability model. For example, the link function for a logit model (respectively, probit model) is the cumulative logistic distribution (respectively, cumulative standard normal distribution) function.
Discrete-time hazard framework: Duration-dependent hazard model (DD)	Shumway (2001) applied an estimation procedure like the one used in estimating the parameters of a multi-period logit model and proposed a (Shumway, 2001) discrete-time hazard model as follows: $P(y_{i,t}=1 \mid x_{i,t}) = h(t \mid x_{i,t}) = \frac{e^{(\alpha_t + x_{i,t}\beta)}}{1+e^{(\alpha_t + x_{i,t}\beta)}}$ Equation 4 where $h(t \mid x_{i,t})$ denotes the hazard rate of firm i at time t, $X_{i,t}$ represents the vector of features of firm i at time t. α_t denotes the time-variant baseline hazard function, which could be associated with the firm, for example, ln (age) or associated with macroeconomic conditions, for example, exchange rate volatility (Nam et al., 2008). Note that Shumway used $ln\ (age)$ as a constant time-variant baseline rate. The notation of the duration-dependent hazard model is as follows: $h(t \mid x_{i,t}) = h_0(t).e^{x_{i,t}.\beta}$ Equation 5 $p(y_{i,t}=1) = \frac{1}{1+e^{-(\alpha_t + x_{i,t}.\beta)}}$ Equation 6
Discrete-time hazard framework: Duration-independent model with time-invariant baseline (DIWTIB) and duration-independent model without baseline hazard rate (DIWOTIB)	The coefficients of the features for the duration independent hazard models are estimated using the multi-period logit framework. However, on the contrary of duration-dependent models, the baseline hazard rate of DIWTIB is a time-invariant term, which could be represented by firm related features such as ln (age), 1/ln (age) or macroeconomic features such as exchange rate volatility. The notation of duration-independent hazard model is as follows: $h(t \mid x_{i,t}) = h_0.e^{x_{i,t}.\beta}$ Equation 7 $p(y_{i,t}=1) = \frac{1}{1+e^{-x_{i,t}.\beta}}$ Equation 8

(Continues)

Framework	Explanation
	DIWOTIB used the multi-period logit framework to estimate the coefficients of the features, however, contrary to DD and DIWTIB, it does not use any baseline hazard rate.
Cox hazard framework	The Cox-hazard model (Cox, 1972) is another duration-dependent model that can take account of time-varying covariates of a firm. This model could be presented as follows: $h(t\|x_{i,t}) = h_0(t).e^{x_{i,t}.\beta}$ Equation 9 The vector of coefficients β are estimated using a partial likelihood function on the training sample, as follows: $PL(\beta) = \prod_{i=1}^{m}\left[\frac{exp\left(\sum_{j=1}^{p}\beta_j x_j^i(t)\right)}{\sum_{k\in R_t(t)} exp\left(\sum_{j=1}^{p}\beta_j x_j^k(t)\right)}\right]$ Equation 10 where i denotes the firm in the event of distress; p is the number of features; k is the firm in the risk set at time t. Note that this equation estimates the vector of β without estimating the baseline hazard rate (Hosmer & Lemeshow, 1999). However, to apply the developed model to estimate the probability of distress, the baseline hazard rate term is required. Therefore, we followed Chen et al. (2005) in estimating the integrated baseline hazard function with time-varying covariates base on Andersen (1992) as follow: $\widehat{H}_0(t) = \sum_{\dot{T}_i \le t} \frac{D_i}{\sum_{j\in\left(\dot{T}_i\right)} \exp\left(\widehat{\beta}.x_j\left(\dot{T}_i\right)\right)}$ Equation 11 where D_i denotes a dummy variable equals to 0 if firm i faces the distress, and 0 otherwise; $\widehat{T}_i$ is the distress time for the ith firm; $\widehat{\beta}$ is the vector of estimated coefficients. Using Equations 10 and 11, we estimate the probability of distress for individual firms in Equation 9.