Expert Systems With Applications 94 (2018) 164–184

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

Systematic review of bankruptcy prediction models: Towards a framework for tool selection

Hafiz A. Alaka[a], Lukumon O. Oyedele[b,*], Hakeem A. Owolabi[c], Vikas Kumar[d], Saheed O. Ajayi[e], Olugbenga O. Akinade[f], Muhammad Bilal[f]

[a] Faculty of Engineering, Environment and Computing, Coventry University, Coventry, United Kingdom
[b] Bristol Enterprise Research and Innovation Centre (BERIC), University of the West of England, Bristol, United Kingdom
[c] Department of International Strategy & Business, The University of Northampton, United Kingdom
[d] Bristol Enterprise Research and Innovation Centre (BERIC), University of the West of England, Bristol, United Kingdom
[e] School of Built Environment and Engineering, Leeds Becket University, Leeds, United Kingdom
[f] Bristol Enterprise Research and Innovation Centre (BERIC), University of the West of England, Bristol, United Kingdom

ARTICLE INFO

Article history:
Received 14 November 2016
Revised 15 October 2017
Accepted 15 October 2017
Available online 26 October 2017

Keywords:
Bankruptcy prediction tools
Financial ratios
Error types
Systematic review
Tool selection framework
Artificial intelligence tools
Statistical tools

ABSTRACT

The bankruptcy prediction research domain continues to evolve with many new different predictive models developed using various tools. Yet many of the tools are used with the wrong data conditions or for the wrong situation. Using the Web of Science, Business Source Complete and Engineering Village databases, a systematic review of 49 journal articles published between 2010 and 2015 was carried out. This review shows how eight popular and promising tools perform based on 13 key criteria within the bankruptcy prediction models research area. These tools include two statistical tools: multiple discriminant analysis and Logistic regression; and six artificial intelligence tools: artificial neural network, support vector machines, rough sets, case based reasoning, decision tree and genetic algorithm. The 13 criteria identified include accuracy, result transparency, fully deterministic output, data size capability, data dispersion, variable selection method required, variable types applicable, and more. Overall, it was found that no single tool is predominantly better than other tools in relation to the 13 identified criteria. A tabular and a diagrammatic framework are provided as guidelines for the selection of tools that best fit different situations. It is concluded that an overall better performance model can only be found by informed integration of tools to form a hybrid model. This paper contributes towards a thorough understanding of the features of the tools used to develop bankruptcy prediction models and their related shortcomings.

1. Introduction

The effect of high rate of business failure can be devastating to firm owner, partners, society and the country's economy at large (Alaka et al., 2015; Edum-Fotwe, Price, & Thorpe, 1996; Hafiz et al., 2015; Xu & Zhang, 2009). The consequent extensive research into developing bankruptcy prediction models (BPM) for firms is undoubtedly justified. The performance of such models is largely dependent on, among other factors, the choice of tool selected to build it. Apart from a few studies (e.g. Altman, 1968; Ohlson, 1980), tool selection in many BPM studies is not based on capabilities of the tool; rather it is either chosen based on popularity (e.g. Abidali & Harris, 1995; Koyuncugil and Ozgulbas, 2012; Langford, Iyagba, & Komba, 1993) or based on professional background (e.g. Altman, Marco, & Varetto, 1994; Beaver, McNichols, & Rhie, 2005; Hillegeist, Keating, Cram, & Lundstedt, 2004; Lin & Mcclean, 2001; Nasir, John, Bennett, Russell, & Patel, 2000). This is because there is no evaluation material which shows and compares the relative performance of major tools in relation to the many important criteria a BPM should satisfy. Such material can provide a guideline and subsequently aid an informed and justified tool selection for BPM developers.

Most prediction tools are either statistical or artificial intelligence (AI) based (Balcaen & Ooghe, 2006; Jo & Han, 1996). The most common statistical tool is the multiple discriminant analysis (MDA) which was first used by Altman (1968) to develop a BPM popularly known as Z model, based on Beaver's (1966) recommendation in his univariate work. MDA, normally used with financial ratios (quantitative variables), subsequently became popular with accounting and finance literature (Taffler, 1982) and many

* Corresponding author.
E-mail addresses: ac7485@coventry.ac.uk (H.A. Alaka), L.Oyedele@uwe.ac.uk (L.O. Oyedele), olugbenga.akinade@uwe.ac.uk (O.O. Akinade).

https://doi.org/10.1016/j.eswa.2017.10.040

subsequent studies by finance professionals simply adopted MDA without considering the assumptions that are to be satisfied for MDA's model to be valid. This resulted in inappropriate application, causing developed models to be un-generalizable (Joy & Tollefson, 1975; Richardson & Davidson, 1984; Zavgren, 1985). Abidali and Harris (1995), for example, unscholarly employed A-score alongside Z-score (i.e. MDA) in order to involve qualitative managerial variables, alongside quantitative variables, in their analysis when logistic regression (LR) [or logit analysis] can handle both types of variables singularly.

AI tools are computer based techniques of which Artificial Neural Network (ANN or NN) is the most common for bankruptcy prediction (Aziz & Dar, 2006; Tseng & Hu, 2010). Simply because it is the most popular architecture, many studies arbitrarily employed the back-propagation algorithm of ANN for bankruptcy prediction (e.g. Boritz, Kennedy, & Albuquerque, 1995; Odom & Sharda, 1990; Tam & Kiang, 1992; Wilson & Sharda, 1994; among others) despite it having a number of relatively undesirable features which include computational intensity, absence of formal theory, "illogical network behaviour in response to different variations of the input values" etc. (Altman et al., 1994; Coats & Fant, 1993, p. 507; Zhang, Hu, Patuwo, & Indro, 1999). Further, Fletcher and Goss (1993) developed an ANN prediction model for a relatively small sample size when ANNs are known to need large samples for optimal performance (Boritz et al., 1995; Ravi Kumar & Ravi, 2007; Shin, Lee, & Kim, 2005).

These improper uses of tools regularly occur because there is no readily available evaluation material or guidelines which can help BPM developers identify which tool best suits what data/purpose/situation. As Chung, Tan, and Holdsworth (2008), p. 20) put it, "*given the variety of techniques now available for insolvency prediction, it is not only necessary to understand the uses and strengths of any prediction model, but to understand their limitations as well*". Hence to ensure a BPM performs well with regards to criteria of preference (e.g. accuracy, type I error, transparency, among others), a model developer has to understand the strength and limitations of the available tools/techniques. This will ensure that the right tool is employed for the right data characteristics, right situation and the right purpose. This study thus aims to develop a comprehensive evaluation framework for selection of BPM tools using a systematic and comprehensive review. The following objectives are needed to achieve this aim:

1. Presentation of an overview of the common tools used for bankruptcy prediction and identification of BPM studies that have used these tools
2. Identifying the key criteria BPMs need to satisfy and how each tool performs in relation to each criterion by analysing the systematic review

The scope of this study is limited to reviewing only popular and promising tools that have been employed for the development of BPMs in past studies since interest in them is high. This is because it is virtually impossible to review all the many tools that can be used for this purpose in this study. In total, two statistical and six AI tools were reviewed. The next section explains the systematic review methodology used in this study with all the inclusion and exclusion criteria. This is followed by a brief description of each of the eight tools. Section four presents the 13 identified key criteria used to assess the tools. Section five discusses the analysis and results of the review in form of tables and charts. Section six presents the proposed tabular and diagrammatic frameworks. This is followed up with a conclusion section.

2. Methodology

This study used a systematic review method to create a guideline for the selection of an appropriate tool for developing a bankruptcy prediction model (BPM). There are so many tools that can be used to develop a BPM that it is virtually impossible to review them all in one study. As a result, the two most popular statistical tools as noted by Balcaen and Ooghe (2006) in their comprehensive review of BPMs were reviewed: multiple discriminant analysis (MDA) and Logistic regression (LR). Also covered in this review are the most popular and promising artificial intelligence (AI) tools as advocated by Aziz and Dar (2006) in their comprehensive review, and Min, Lee, and Han (2006) among others: artificial neural network (ANN), support vector machines (SVM), rough sets (RS), case based reasoning (CBR), decision tree (DT) and genetic algorithm (GA). A process flow of the methodology is presented in Fig. 1.

Systematic review is a well-known method for producing valid and reliable knowledge as it minimizes bias hence its popularity in the all-important medical research world (Schlosser, 2007; Tranfield, Denyer, & Smart, 2003). The inclusion criteria for this study were carefully chosen to allow fair comparison and ensure adequate quality (Khan, Kunz, Kleijnen, & Antes, 2003). To improve validity of this study, only peer reviewed journal articles were considered since they are considered to be of high quality and their contribution considered as very valid (Schlosser, 2007).

Systematic review requires wide literature search (Smith, Devane, Begley, & Clarke, 2011) hence following Appiah, Chizema, and Arthur (2015) approach, which is the most recently published systematic review in the BPM research area, the following databases were considered: Google Scholar; Wiley Interscience; Science Direct; Web of Science UK (WoS); and Business Source Complete (BSC). However, a careful observation revealed Google scholar produced an almost endless result and did not have the required filters to make it very efficient hence it was removed as it was unmanageable. Further observation revealed that (WoS) and BSC contained all the journal articles provided in Wiley and Science Direct; this is probably because the latter two are publishers while the former two are databases with articles from various publishers including the latter two. To increase the width of the search, Engineering Village (EV) database was added to WoS and BSC databases to perform the final search. EV was chosen because articles from the engineering world usually deal with BPM tools comprehensively.

The initial searches in the three databases (WoS, BSC and EV) showed that studies tend to use bankruptcy, insolvency and financial distress as synonyms for failure of firms. A search framework which captured all these words was thus designed with the following defined string ("Forecasting" OR "Prediction" OR "Predicting") AND ("Bankruptcy" OR "Insolvency" OR "Distress" OR "Default" OR "Failure").

To ensure high consistency and repeatability of this study, and consequently reliability and quality (Stenbacka, 2001; Trochim & Donnelly, 2006), only studies that appeared in the three databases were used; this ensured the eradication of database bias (Schlosser, 2007). These databases contain studies from all over the world hence geographic bias was also eliminated. Balcaen and Ooghe (2006) in their comprehensive review of statistical tools in 2006 noted that AI tools, mainly ANN, were gradually becoming adopted in BPM studies. With new tools emerging all the time, a four-year advance from 2006, which would have seen more use of AI tools, is how a start year of 2010 was chosen for this study. The end year is the year this paper was written, 2015.

Generally, the topic of articles that emerge from the search looked okay to determine which ones were fit for this study. However, this was not the case for all articles. Where otherwise, arti-

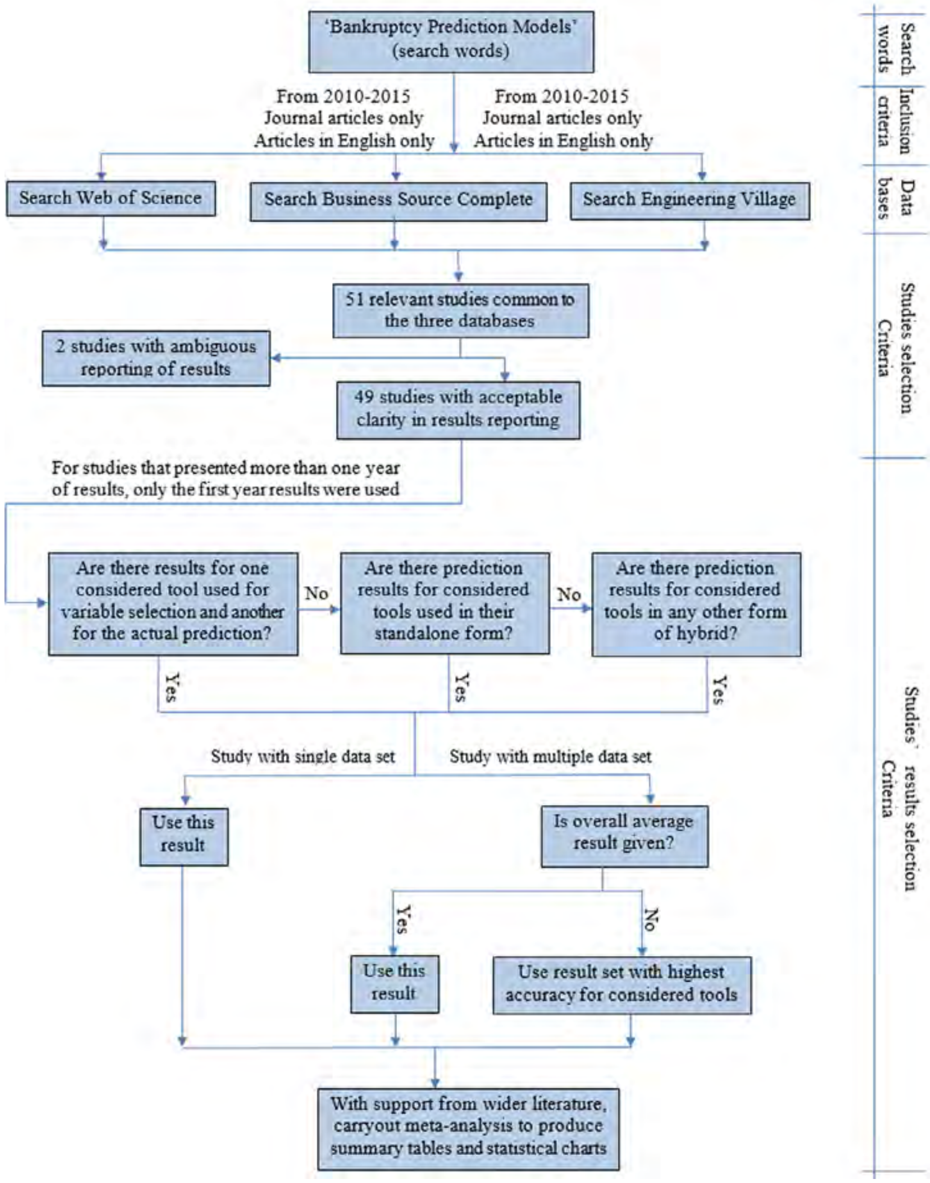

Fig. 1. Process flow of the methodology (the term 'considered tool' refers to the eight tools considered in this study).

cle's abstract was read and, if necessary, introduction and/or conclusion were read. In some cases, the complete articles had to be read. Although language constraint is not encouraged in systematic review, it is sometimes unavoidable due to lack of funds to pay for interpretation services (Smith et al., 2011) as in the case of this study. Only studies written in English were thus used.

After eliminating unrelated studies that dealt with topics like credit scoring (e.g. Martens et al., 2010), policy forecasting (e.g. Won, Kim, & Bae, 2012), or that did not use any of the tools reviewed (e.g. Martin, Manjula, & Venkatesan, 2011), only 51 studies had a presence in the three data bases. Of these, two had great ambiguity in reporting their results hence were excluded, leaving 49 studies to be used as the sample. The 'review studies' in the search results (e.g. Sun, Li, Huang, & He, 2014) were not considered since original results from tools implementation were needed.

In the final 49 studies sample (i.e. the primary studies used for this systematic review), where results of tools in their hybrid forms and the result of the tools in their standalone form were presented, results of the tools in their standalone form was used to allow fair comparisons of inter-study results. However, where any of the eight tools in consideration is used for variable selection and in turn used to hybridise the predicting tool, the result of such hybrid is used. Where the tools were used on more than one dataset and the results of each dataset was presented alongside the total average from all dataset, the total average results were used. In cases where average values were not given, the result set with the best accuracy for most/all of the tools in consideration in this study was used so as to give all tools a good chance of high accuracy. In cases where the results of more than a year of prediction were given, the results of the first year were used to allow for fair comparison since most BPM studies normally present first year results.

As required for systematic review, a meta-analysis was done with data synthesised using 'summary of findings' tables, statistical methods and charts (Higgins, 2008; Khan et al., 2003; Smith et al., 2011); with facts explained sometimes by employing quotes from especially the reviewed studies, and discussions backed up with a wider review of literature. Three summary of findings tables were provided in this study. Where there is not enough information from the reviewed studies regarding a certain criterion, results are discussed using the reviewed studies and wider literature. Opinions are only taken as facts in such cases if there are no opposing studies. This type of deviation from protocol for a valid reason is acceptable in systematic review (Schlosser, 2007). Finally, this review is used to create a guideline using a tabular framework for tool comparisons and a diagrammatic framework that clearly shows what situations/data characteristics/variable types etc. each discussed tool is best suited to. This will ensure developers can choose a tool based on what they have and/or intend to get rather than just arbitrarily.

3. The tools

This study will review eight tools used to develop bankruptcy prediction models including: multiple discriminant analysis (MDA), Logistic regression (LR), artificial neural network (ANN), support vector machines (SVM), rough sets (RS), case based reasoning (CBR), decision tree (DT) and genetic algorithm (GA).

Multiple discriminant analysis: MDA uses a linear combo of variables, normally financial ratios, that best differentiate between failing and surviving firms to classify firms into one of the two groups. The MDA function, constructed usually after variable selection, is as follows:

$$Z = c_1X1 + c_2X2 + \cdots + c_nXn.$$

Where c_1, c_2, ... c_n = discriminant coefficients; and X1, X2, ... Xn = independent variables

MDA calculates the discriminant coefficients. The function is used to calculate a Z-score. A cut-off Z score is chosen based on status of sample firms

Logistic regression: LR is a “conditional probability model which uses the non-linear maximum log-likelihood technique to estimate the probability of firm failure under the assumption of a logistic distribution” (Jackson & Wood, 2013, p. 190). The LR function, constructed after variable selection, is as follows:

$$\begin{aligned} P_1(V_i) &= 1/[1 + exp - (b_0 + b_1V_{i1} + b_2V_{i2} + \ldots\ldots + b_nV_{in})] \\ &= 1/[1 + exp - (D_i)] \end{aligned}$$

where $P_1(V_i)$ = probability of failure given the vector of attributes; V_i; V_{ij} = value of attribute or variable j (j = 1, 2, ..., n) for firm *i*; b_j = coefficient for attribute *j*; b_0 = intercept; D_i = logit of firm *i*.

The dependent variable P_1 is expressed in binary form (0,1) (Boritz & Kennedy, 1995).

Neural network: ANN was created to imitate how the neural system of the human brain works (Hertz, Krogh, & Palmer, 1991) and was first applied to bankruptcy prediction by Odom and Sharda (1990). A typical ANN is a network of nodes interconnected in layers. There are various parameters, architectures, algorithms, and training methods that can be used to develop an ANN (Jo & Han, 1996) and choosing the best combination can be demanding.

Support vector machines: SVM employs a linear model to develop an optimal separating hyperplane by using a highly nonlinear mapping of input vectors into a high-dimensional feature space (Ravi Kumar & Ravi, 2007; Shin et al., 2005). It constructs the boundary using binary class. The variables closest to the hyperplane are called support vectors and are used to define the binary outcome (failing or non-failing) of assessed firms. All other samples are ignored and are not involved in deciding the binary class boundaries (Vapnik, 1998). Like ANN, it has some parameters that can be varied for it to perform optimally (Dreiseitl & Ohno-Machado, 2002).

Rough sets: RS theory, discovered by Pawlak (1982), assumes that there is some information associated with all objects (firms) of a given universe; information which is given by some attributes (variables) that can describe the objects. Objects that possess the same attributes are indiscernible (similar) with respect to the chosen attributes. RS creates a partition in the universe that separates objects with similar attributes into blocks (e.g. failing and non-failing blocks) called elementary sets (Greco, Matarazzo, & Slowinski, 2001). Objects that fall on the boundary line cannot be classified because information about them is ambiguous. RS is used to extract the decision rules to solve classification problems (Greco et al., 2001; Ravi Kumar & Ravi, 2007).

Case based reasoning: CBR fundamentally differs from other tools in that it does not try to recognize pattern, rather it classifies a firm based on a sample firm that possess similar attribute values (Shin & Lee, 2002). It justifies its decision by presenting the used sample cases (firms) from its case library (Kolodner, 1993). It induces decision rules for classification.

Decision tree: DT became an important machine learning tool after Quinlan (1986) developed iterative dichotomiser 3 (ID3). DT uses entropy to measure the discriminant power of samples' variables and subsequently recursively partitions (RPA) the set of data for the classification of firms (Quinlan, 1986). Quinlan (1993) later developed the advanced version called Classifier 4.5 (C4.5). DT induces the decision rules. The positions of the rules in the decision tree are usually determined using heuristics (Jeng, Jeng, & Liang, 1997). For example, if profitability was found to be more important than liquidity, it will be placed above, or evaluated before, liquidity.

Genetic algorithm: GA is a searching optimization technique that imitates the Darwin principle of evolution in solving nonlinear, non-convex problems (Ravi Kumar & Ravi, 2007). It is effective at locating the global minimum in a very large space. It differs from other tools in that it simultaneously searches multiple points, works with character springs and uses probabilistic and not deterministic rules. GA can extract decision rules from data which can be used for classifying firms. It is applied to selected variables in order to find a cut-off score for each variable (Shin & Lee, 2002).

4. Important criteria required for bankruptcy prediction model tools

To be considered effective, there are many criteria that can be required to be satisfied by a tool when it is used to develop a bankruptcy prediction model (BPM). The set of criteria required usually depend on the situation and intention of the BPM developer. For example, a financier or client may simply be interested in the accuracy of a model. The BPM needed simply needs to be able to predict if a firm is financially healthy (unhealthy) enough to be granted (refused) a loan or contract, hence a highly accurate tool/technique is needed. A firm owner on the other hand is interested in result transparency as much as accuracy because he needs to know where/what the firm is going/doing wrong in order to know where rescue efforts need to be focused on. In such a case, a tool with high accuracy as well as result transparency will be needed to build the required BPM

Different researchers have used different criteria to develop their BPMs, however after a thorough and comprehensive review of the primary studies and other studies in the area, 13 criteria were identified to be the most common and important. Example of other reviewed studies include Ahn and Kim (2009), Balcaen and Ooghe (2006), Chung et al. (2008), Dreiseitl and Ohno-Machado (2002), Edum-Fotwe et al. (1996), Haykin (1994), Min and Lee (2005), Ravi Kumar and Ravi (2007), Shin et al. (2005), Tam and Kiang (1992) and Zurada, Malinowski, and Cloete (1994); to mention a few. The identified 13 criteria are as follows:

1) *Accuracy*: This relates to the percentage of firms a tool correctly classifies as failing or non-failing.

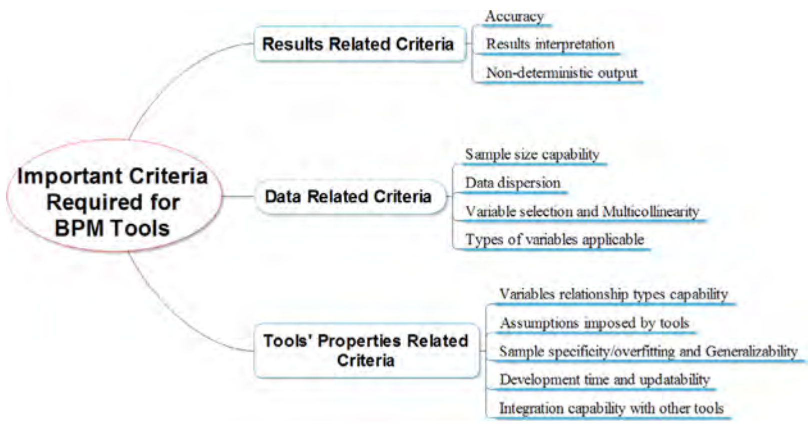

Fig. 2. Important criteria required for BPM tools.

2) *Result transparency*: This has to do with interpretability of a tool's result.
3) *Non-deterministic*: The case where a tool cannot successfully classify a firm
4) *Sample size*: This refers to the sample size(s) suitable for a tool to perform optimally.
5) *Data dispersion*: This refers to ability of a tool to handle equally or unequally dispersed data
6) *Variable selection*: This refers to the variables selection methods required for optimum tool performance.
7) *Multicollinearity*: This refers to sensitivity of a tool to collinear variables.
8) *Variable types*: A tool's capability to analyse quantitative and/or qualitative variables.
9) *Variable relationship*: This explains a tool's limitation in analysing linear or non-linear variables
10) *Assumptions imposed by tools*: Requirements a sample data has to satisfy for a tool to perform optimally.
11) *Sample specificity/overfitting*: This is when the model developed from a tool performs well on sample firms but badly on validation data.
12) *Updatability*: The ease with which a tool's model can be updated with new sample firms and its effectiveness afterwards
13) *Integration capability*: the ease with which a tool is hybridisable.

These 13 criteria can be grouped into three main categories as shown in Fig. 2. These categories are:

1) Results related criteria
2) Data related criteria
3) Tool's properties related criteria

5. Results and discussion

This section presents the results, analysis and discussion of the systematic review in form of summary of findings tables and statistical charts. The results are presented in relation to each identified criterion. The tables and charts compare the performance/ability of the tools as deduced from all the reviewed studies. The outcome is used to judge each tool based on the criterion in question. The criteria were assessed and discussed in the context of bankruptcy prediction models (BPM). For example, to assess the accuracy criteria of the tools, error cost had to be considered since it is an important aspect of accuracy assessment in the BPM research area.

Not all the reviewed studies provided necessary tool information required to assess a tool under each criterion. Where a tool had too few studies providing information regarding a criterion, the tool was excluded from the statistical analysis regarding that criterion. The measure for exclusion was determined by calculating the average number of studies that provided information on the tools regarding the criterion in consideration; the tools with numbers well below average were excluded. This process, where employed, is clearly explained.

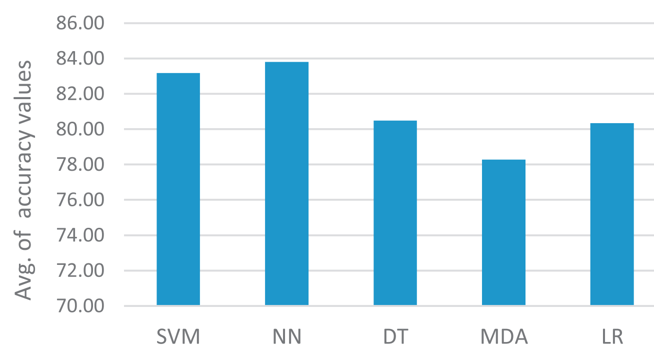

Fig. 3. Overall average accuracy chart for each tool.

5.1. Results related criteria

5.1.1. Accuracy

One of the main essences of using varying tools to develop BPMs is to increase accuracy of prediction. Fig. 3 shows the mean average accuracy chart for each tool calculated from all the studies that gave an accuracy reading for the tool. The chart includes only the tools that had up to 17 studies that reported an accuracy value for them since the mean average of number of studies that reported accuracy value is 17 (Table 1). The chart clearly shows ANN and SVM to be the most accurate while MDA appears to be the least accurate. Table 2 is the first summary of findings table. It shows the accuracy value of each tool as reported in each study.

The chart in Fig. 4 shows a more direct comparison between pair of tools. It shows the average accuracy value calculated from studies that directly compared any pair of tools. The chart includes only the pairs that were compared in five or more studies since the mean average of the number of times any two tools were directly compared is 5.5 (see Table 3). To be more objective and fair in analysis, and to make a good critique, the pie charts in Fig. 5 is produced to compare the percentage of studies that rated one tool as being more accurate than the other. It contains exactly the same pairs as Fig. 4.

Figs. 3–5 all show that AI tools are more accurate than statistical tools except in Fig. 5j where the number of studies that indi-

Table 1
Summary statistics of the accuracy and error types of the tools.

Tool	No. of authors that used tool	No. of authors that reported an accuracy value	No. of authors that reported Type I error	No. of authors that reported Type II error
SVM	24	22	10	11
ANN	38	37	18	18
DT	19	17	5	6
RS	4	4	1	1
GA	10	6	4	4
CBR	4	3	0	0
MDA	21	19	7	7
LR	31	28	12	12
Total	Fu = 151	fac = 136	fo = 57	ft = 59
Mean		Σ fac/Σf = 17	Σ fo/Σf = 7.1	Σ ft/Σf = 7.35

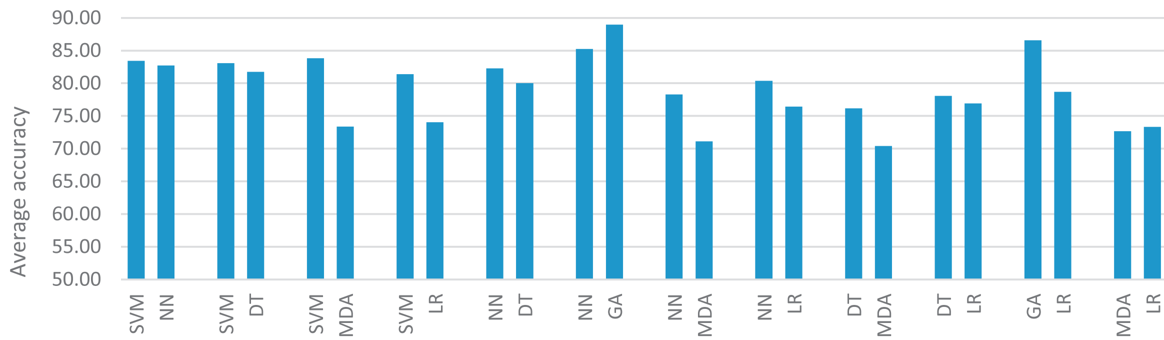

Fig. 4. Average accuracy results only from studies that directly compared pair of tools.

cated that DT is more accurate than LR and vice versa are equal. Figs. 4 and 5a–d clearly show SVM to be more accurate than any directly comparing tool though Fig. 3, which is just the average accuracy of each tool, shows ANN to be more accurate. SVM apart, Figs. 4 and 5e–h similarly show ANN to be more accurate than any comparing tool except for GA; with three against two studies confirming GA to be more accurate.

A further examination of the five studies that compared the pair (ANN and GA) revealed they were written by two main set of authors. Of the five studies, only Chen, Yang et al. (2011), which reported ANN and SVM as being more accurate than GA (see Table 2) could be said to have done a fair comparison since it produced the results for ANN and GA using the same features. Chen, Ribeiro, Vieira, Duarte, and Neves (2011) in their study developed a robust hybrid of GA and K-nearest neighbour (KNN) and compared it with other tools including ANN and SVM in their standalone form thus giving GA the advantage. Divsalar, Firouzabadi, Sadeghi, Behrooz, and Alavi (2011) 'unfairly' used a special version of GA called linear genetic programming (LGP) for comparison with normal ANN. Also, Divsalar, Roodsaz, Vahdatinia, Norouzzadeh, and Behrooz (2012) proposed a special version of GA called gene expression programming (GEP), thoroughly developed its BPM using all possible enhancements, and proved it was more accurate than models from other tools, including ANN developed with default settings.

Similarly, Kasgari, Divsalar, Javid, and Ebrahimian (2013), which included Divsalar as the second author, used the same data as Divsalar et al. (2012), proposed ANN for developing BPMs, thoroughly developed its model and proved it was more accurate than other tools, including GA. Besides, GA is well known to be more suited to the process of feature/variable selection because of its powerful global search hence its relatively infrequent use to develop BPMs (see Fig. 6); it was used for this purpose in at least four of the primary studies (Chen, 2011; Jeong, Min, & Kim, 2012; Liang, Tsai, & Wu, 2015; Zhou, Lai, & Yen, 2014) and other studies. Further "GA is a stochastic one. So, when using GA-based models on the same training samples twice, we may get two different models, and the decision on the same test sample may also be different. This stochastic characteristic of this method may be unacceptable for the decision makers or the analysts" (Zhou et al., 2014, p, 252). SVM and ANN can thus be claimed to be more accurate. As noted in some of the reviewed studies (e.g. Iturriaga & Sanz, 2015; Virág & Nyitrai, 2014), this is in line with literature as it is mostly agreed that SVM and ANN are the most accurate tools for developing BPMs.

No study compared RS directly with GA. However, RS and ANN as well as SVM were compared directly in two studies and RS gave a slightly better result. Like the unfair cases with GA, Yeh, Chi, and Lin (2014) thoroughly developed many RS hybrid models using various enhancements and compared their average accuracy value to single accuracy values of separate hybrids of ANN and SVM. In the second study, Virág and Nyitrai (2014), working further from their previous study which confirmed SVM and ANN to be most accurate tools, decided to check why non-transparent tools (i.e. SVM and ANN) were more accurate than transparent tools like RS. They (Virág & Nyitrai, 2014) initially concluded "there seems to be a kind of trade-off between the interpretability and predictive power of bankruptcy models" (p.420) so they tried to find out what to use with "RST technique in order to maximise the predictive power of the constructed model?" In other words, special effort was made to improve RS accuracy while SVM and ANN were used at default level; this obviously resulted in a biased result. Despite the effort, RS was only able to achieve the same accuracy as SVM and only a slightly higher accuracy than ANN (Table 3). Besides, Ravi Kumar and Ravi (2007) showed in their review that RS is not as accurate as claimed in many studies and Mckee (2003) reported a significantly reduced accuracy, compared to his previous study, when used with what was termed a 'more realistic' data. RS theory is difficult to implement hence its sparse usage (see Fig. 6).

While DT's average accuracy appears slightly higher than LR's in Fig. 4, Fig. 5j shows that the number of studies that indicated DT to be more accurate than LR and vice versa are the same. DT has generally been confirmed to be less accurate than other AI tools

Table 2
Summary of reviewed studies aims, variable selection methods, sample characteristics and accuracy values.

S/N	Author year	Aim of study	Variable selection method	Sample size	% of Exist firms	% of Fail firms	% for val.	Accuracy values							
								SVM	ANN	DT	RS	GA	CBR	MDA	LR
1.	Tseng and Hu (2010)	**Comparing models**	Literature rev. (stepwise)	77	58.4	41.6	20		93.75						86.25
2.	Cho et al. (2010)	Propose new CBR approach	*t*-test& decision tree	1000	50	50	20		71.8	65.7			73.7		72.2
3.	Kim and Kang (2010)	Check enhanced ANN against ord. ANN	Cumulative accuracy profiles	1458	50	50	10		71.02						
4.	Yoon and Kwon (2010)	Use credit card data for small bus. & compare techniques	*t*-test & chi-square	10000	50	50	30	74.2	73.1					70.1	70.1
5.	Du Jardin (2010)	To Check variable selection methods effect	Error backward-order (stepwise)	1020	50	50	49		94.03					87.20	92.01
6.	Lin, Liang, and Chu (2010)	Use SVM with ratios & non-fin. variables	Stepwise regression	108	50	50		94.44						90.74	
7.	Gepp, Kumar, and Bhattacharya (2010)	Compare DT & MDA	Lit. rev (stepwise)	200	71	29	20			87.6				84.5	
8.	De Andrés, Lorca, de Cos Juez, and Sánchez-Lasheras (2011)	Propose hybrid model (C-means & MARS)	Altman's ratios (stepwise)	59474	99.77	0.23	20		92.38					91.44	86.56
9.	Kim (2011)	Compare techniques	Stepwise	66	50	50		95.95	91.6					72.6	80
10.	Yang et al. (2011)	Propose hybrid model (PLS & SVM)	Pearson cor. & PLS	120	53.3	46.7	100	79	78.33						
11.	Chen (2011)	Use PSO with SVM	Lit. rev. (stepwise), GA	80	50	50	20								
12.	Divsalar et al. (2011)	Use GA & NN	SFS	150	51.4	48.6			82.5			95			80
13.	Du Jardin and Séverin (2011)	Use self-organizing map	Error backward-order (stepwise)	2360	50	50	37.3		82.61					81.93	81.14
14.	Chen, Ribeiro et al. (2011)	Integrate error cost into model		1200	50	50	20	90	90.6			86.7			
15.	Chen, Yang et al. (2011)	Propose FKNN		240	53.3	46.7	10	76.67	79.58			83.33			
16.	Li et al. (2011)	Propose Random subspace LR (RSBL)	Stepwise & *t*-test	370	50	50	30				88.46			88.26	87.50
17.	Divsalar et al. (2012)	Use new type of GA called GEP	SFS	136	52.5	47.5	33.3		79.41			91.18			76.47
18.	Huang et al. (2012)	Propose hybrid KLFDA & MR-SVM					10	86.61	83.67	83.24					77.9
19.	Tsai and Cheng (2012)	Check effect of outlier on BPMs		653	45.3	54.7	10	86.37	86.06	84.69					86.37
20.	Shie et al. (2012)	Proposed enhanced PSO-SVM	Factor analysis & PCA	54	55	44.4		81.82	75.76	77.77					72.73
21.	Kristóf and Virág (2012)			504	86.7	13.3	25		88.7	88.8					88.5
22.	Jeong et al. (2012).	To fine-tune ANN factors	GAM	2542	50	50	20	79	81	76			73	73.5	76.48
23.	Du Jardin and Séverin (2012)	To use Kohonen map to stabilize temporal accuracy							81.3					81.2	81.6
24.	De Andrés et al. (2012)	To improve performance of classifiers		122	50	50	19.6		76.03					74.87	
25.	Zhou, Lai, and Yen (2012)	To find the best variables for accuracy	Spearman correlation		50	50	10.8	71.1	67.8	75.6				64.4	54.4
26.	Xiong, Wang, Mayers, and Monga (2013)	Use sequence on credit card data						70.94							
27.	Lee and Choi (2013)	To do multi industry investigation	*t*-test &correlation analysis	1775	66.2	33.8	4.2		92					82.01	
28.	Tsai and Hsu (2013)	Present met-learning framework (hybrid)	MC	Avg. many			20		78.82	77.29					79.11
29.	Callejón, Casado, Fernández, and Peláez (2013)	To increase predictive power of ANN		1000	50	50	20		92.11						
30.	Chuang (2013)	To Hybridise CBR	Multiple	321	86.9	13.1							90.1		

(*continued on next page*)

Table 2 (*continued*)

S/N	Author year	Aim of study	Variable selection method	Sample size	% of Exist firms	% of Fail firms	% for val.	Accuracy values							
								SVM	ANN	DT	RS	GA	CBR	MDA	LR
31.	Ho, McCarthy, Yang, and Ye (2013)	Develop BPM for US paper companies	Lit rev (stepwise)	366	66.7	33.3	20								93
32.	Arieshanti, Purwananto, Ramadhani, Nuha, and Ulinnuha (2013)	To compare techniques	Lit rev. (stepwise)	240	53.3	46.7	20	70.42	71						
33.	Kasgari et al. (2013)	Compare ANN to other techniques	Garson's algorithm	135	52.5	47.5	25		94.11			88.57			91.43
34.	Zhou et al. (2014)	Propose new feature selection method	GA	2010	50	50			75.6	50.67				71.72	73.99
35.	Tsai (2014)	To compare hybrids	SOM	690	44.5	55.5	20		91.61	86.83					87.28
36.	Yeh et al. (2014)	To increase accuracy using RF&RS	RF	220	75	25	33	94.58	92.95	91.55	96.99				
37.	Wang, Ma, and Yang (2014)	Inject feature selection into boosting		132	50	50	10	79.99	75.69	75.99					73.90
38.	Abellán and Mantas (2014)	To correctly use bagging scheme	Lit. rev. (stepwise)	690			30				93.64				
39.	Tserng, Chen, Huang, Lei, and Tran (2014)	To use LR to predict contractors default		87	66.7	33.3									79.18
40.	Yu, Miche, Séverin, and Lendasse (2014)	Produce BPM using ELM		500	50	50	33.3	93.2						86.5	
41.	Gordini (2014)	Test GA accuracy & compare to other techniques	VIF & stepwise	3100	51.6	48.4	30	69.5				71.5			66.8
42.	Heo and Yang (2014)	To prove AdaBoost is right for Korean construction firms		2762	50	50	20	73.3	77.1	73.1				51.3	
43.	Tsai, Hsu, and Yen (2014)	To compare classifier ensembles		690	44.5	55.5	10	86.37	84.38	86.37					
44.	Virág and Nyitrai (2014)	To show RS accuracy is competitive with SVM & ANN		156	50	50	25	89.32	88.03		89.32				
45.	Liang et al. (2015)	To compare feature selections	GA	688	50	50	10	91.77	91.63	92.98					
46.	Iturriaga and Sanz (2015)	To develop ANN BPM for US banks	Mann-Whintney test & Gini index	772	50	50	13.5	89.42	93.27					77.88	81.73
47.	Du Jardin (2015)	To improve BPM accuracy beyond one year		16880	50	50	50		80.8					80.1	80.6
48.	Bemš et al. (2015).	Introduce new scoring method called Gini index	Gini index	459	67	33	579		0.291				0.199	0.207	0.301
49.	Khademolqorani, Zeinal Hamadani, and Mokhatab Rafiei (2015)	To develop a novel hybrid	Factor analysis	180		58			94	94				77	80

Cor.: correlation ELM: extreme learning machine Exist firms: non bankrupt firms Fail firms: bankrupt firms FKNN: fuzzy *k*-nearest neighbour GAM: generalized additive model GEP: gene expression programming Lit. Literature KLFDA: kernel local fisher discriminant analysis MARS: Multivariate Adaptive Regression Splines MC: meta classifier Rev.: review MR: manifold-regularized PCA: principal component analysis PLS: partial least squares PSO: particle swarm optimization RF: random forest RSBL: random subspace binary logit SFS: sequential feature selection SOM: self-Organising maps Val.: Validation VIF: variance Inflation Factor

Note: Bemš et al. (2015) scoring methods results are not used as they will act as outliers in the computations of mean average and disadvantage accuracy results of tools that have them. Chen (2011) results were not clear enough to be included for computational analysis

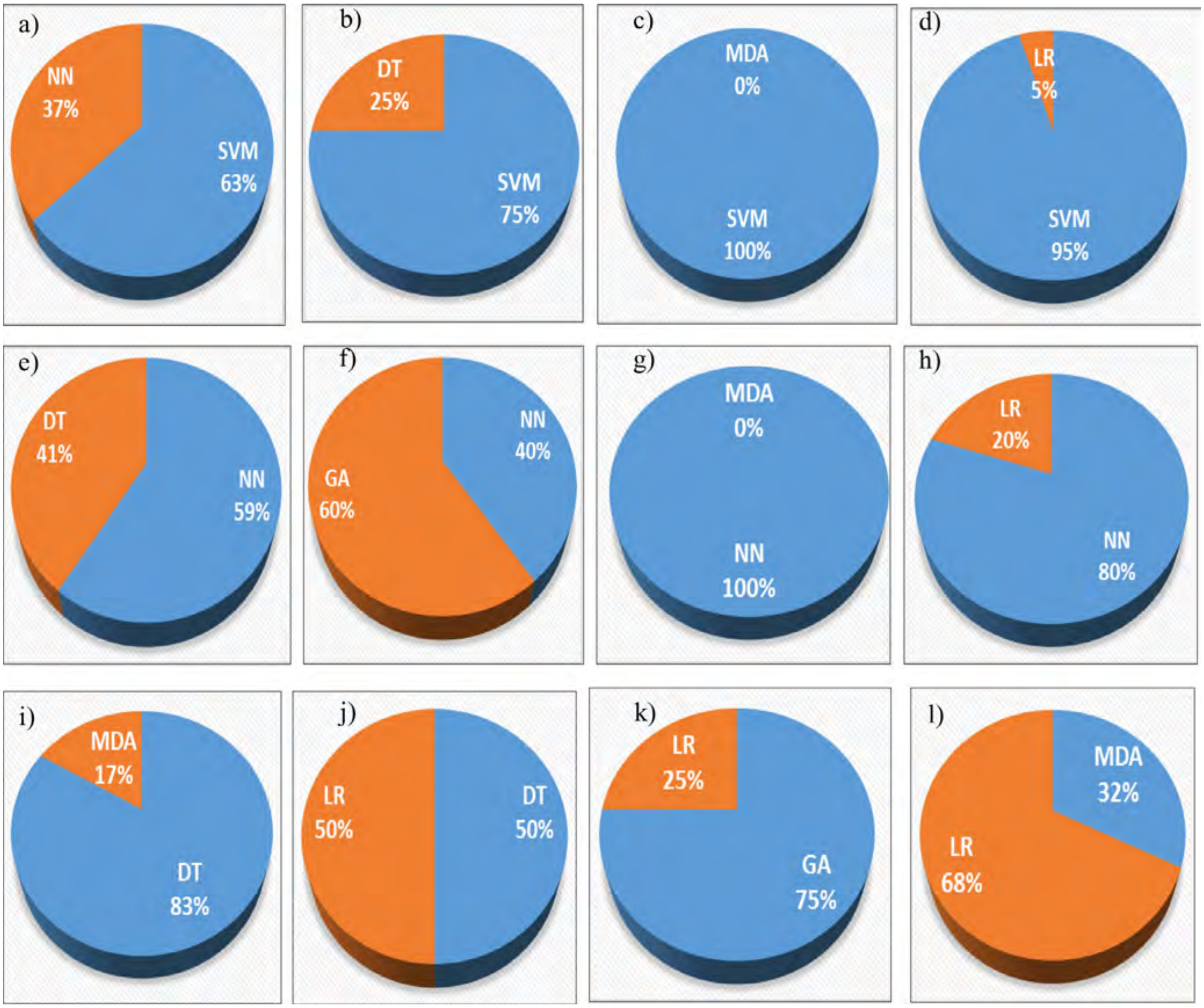

Fig. 5. Pie charts that compare the percentage of studies that indicated one tool as being more accurate than the other.

Table 3
Matrix of number of times studies directly compared pair of tools. The tools compared above the average number of comparisons are in bold.

	SVM	ANN	DT	RS	GA	CBR	MDA	LR	Total
SVM	–	**18**	10	2	3	1	8	10	
ANN	**18**	–	16	2	5	3	16	25	
DT	**10**	**16**	–	1	0	2	6	12	
RS	2	2	1	–	0	0	1	1	
GA	3	5	0	0	–	0	0	4	
CBR	1	3	2	0	0	–	2	3	
MDA	**8**	**16**	**6**	1	0	2	–	14	
LR	**10**	**25**	**12**	1	**4**	3	**14**	–	
Total	53	67	21	2	4	5	14	0	165
Mean average	165/30 = 5.5								

like ANN and RS (Tam & Kiang, 1992; Chung & Tam, 1993; McKee, 2000; Ravi Kumar & Ravi, 2007) except CBR; it (DT) has been classified as a somewhat weak classifier in one of the reviewed studies (Heo & Yang, 2014). CBR is the overall least accurate tool. Of the four studies that used it, Chuang (2013) used it alone without comparison to any other tools. Jeong et al. (2012) and Bemš, Starý, Macaš, Žegklitz, and Pošík (2015) showed that it was the least accurate when compared to SVM, DT, ANN, MDA and LR (Table 2). Only Cho, Hong, and Ha (2010), who presented an enhanced and hybridised CBR using DT and Mahalanobis distance, which was the aim of the study, was able to get a better accuracy figure for CBR (hybrid) than ANN, MDA and LR (Table 2). CBR's low accuracy is a consequence of it not being able to handle non-linear problems and has been deemed by some as not suitable for bankruptcy prediction (e.g. Bryant, 1997; Ravi Kumar & Ravi, 2007). In the reviewed studies, Chuang (2013) noted that "one major factor for the poorer performance of a stand-alone CBR model lies in its failure to separate the more important "key" attributes from those

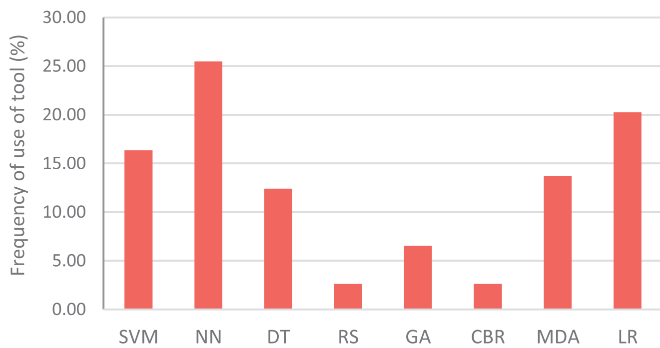

Fig. 6. Percentage frequency of use of each tool.

less significant common attributes and to assign each key attribute with a different, corresponding weight" (p.184). No wonder it is very scarcely used for BPMs (see Fig. 6). Of the two statistical models, LR is clearly the more accurate tool.

5.1.1.1. Error cost. For accuracy ratings, error cost is a very important concept in bankruptcy prediction hence the tools must be appraised with regards to it. There are two types of error in bankruptcy prediction: type I and type II. Type I error is when a tool misclassifies a potentially bankrupt firm as being healthy. This is costlier as it could cause a financier to loan money to a failing firm and eventually lose the money, or it could make a firm relax when it is supposed to take active steps against insolvency. Type II error is when a tool misclassifies non-bankrupt firm as potentially bankrupt/failing. This error is less costly. This means a tool with relatively lesser type I error is more accurate. This, however, does not imply that type II error is unimportant as it could cost the firm its eligibility for loans, for example.

Since the mean average of frequency of reported types I and II errors are 7.36 and 7.63 respectively (Table 1), only the six tools that had up to 4 reporting studies and above were compared in the average types I and II error of tools chart in Fig. 7. Four is deemed not too far from seven in this case so as to allow more tools to be compared. All error values are presented in Table 4. No study reported an error value for CBR while only one study reported for RS. Fig. 7 shows that ANN has the least average type I error followed by SVM. Coupled with their high normal accuracy performance, they can be concluded to be the most accurate tools for bankruptcy prediction, followed by GA. DT and LR errors are again as close as their accuracies hence their total accuracy can be regarded to be of the same rank. However, MDA appears to be very poor with type I error hence its accuracy can be regarded as low. ANN, DT, GA and LR have better type I errors than type II errors and vice versa for SVM and MDA

5.1.2. Results interpretation

For financiers and potential clients, it is enough for a firm to be predicted as healthy or about to bankrupt. However, for firm owners to appreciate a prediction model, the model must give an indication of where a firm is going wrong if the firm is classified as a failing firm so that necessary steps can be taken to avoid total failure if possible. In this context, about a quarter of the studies that used SVM and ANN highlighted the 'black box' nature of the tools as a major problem (Table 5 and Fig. 8). Other studies have also pointed out that the results of ANN and SVM models are quite hard to understand in that weightings/coefficients they assign to the variables are illogical and very hard to interpret (Ahn & Kim, 2009; Chung et al., 2008; Shin et al., 2005; Tam & Kiang, 1992; Tseng & Hu, 2010).

As noted by at least five of the reviewed studies (Table 5) and older studies (e.g. Balcaen & Ooghe, 2006; Boritz & Kennedy, 1995; Ohlson, 1980; Tam & Kiang, 1992; among others), the variable coefficients in LR represent the importance of variables thus its result is transparent and help users identify key areas of problem of a failing firm. As noted by at least five of the reviewed studies (Table 5) and some previous studies (Greco et al., 2001; McKee, 2000; Ravi Kumar & Ravi, 2007; Shin & Lee, 2002; Shin et al., 2005; Tam & Kiang, 1992), AI tools that generate decision rules for classification (i.e. RS, CBR, GA and DT) all produce explanatory results that can be easily interpreted and understood. It appears that for AI tools, the more accurate the tool, the less transparent the result (Fig. 9). Nonetheless, McKee (2000) once spotted an inconsistency in a set of rules generated by RS in one of his previous co-authored studies.

Kim (2011) and older studies (Altman, 1968; Balcaen & Ooghe, 2006; Taffler, 1983; Tam & Kiang, 1992) noted that although the MDA function makes MDA result look easily interpretable, the truth is that the variables' coefficients in the function do not represent their importance, hence results are hard to interpret. Further, MDA sometimes yields a model with counter intuitive signs (Balcaen & Ooghe, 2006; Edum-Fotwe et al., 1996). One of many example models is Mason and Harris' (1979) in which a negative sign was assigned to the profit before tax variable while representing firms with scores above cut-off as being healthy. This means profit is bad for a firm's health! This is obviously incomprehensible.

One relatively popular approach to transparency problem has been to use decision rules-generating tools to select variables and decide the importance of variables before using the very accurate black box tool for prediction. This, from the way it is explained, is obviously not the perfect answer as it sounds like using two separate tools for two different criteria. Kasgari et al. (2013, p.930) suggested that "to overcome this difficulty, the weights and biases are frozen after the network was well trained and then the trained MLP models are translated into explicit forms", but did not explain how this is done.

5.1.3. Non-deterministic output

Unlike statistical tools and non-decision rules AI tools (i.e. ANN and SVM), AI tools that induce decision rules for classification can produce some non-deterministic rules i.e. rules that cannot be applied to a new object (firm) being assessed. The presence of non-deterministic rules for a new object can result into no classification (Ahn, Cho, & Kim, 2000; McKee, 2000; Ravi Kumar & Ravi, 2007; Shin & Lee, 2002). Of the reviewed studies, Gordini (2014) highlighted GA as a tool that is synonymous with this problem. According to Shin and Lee (2002), as much as 46% of new cases might not be classified by these tools (GA was used in their study).

The non-deterministic problem is encountered in this group of tools because the set of rules extracted work like a multiple univariate system rather than a multivariate system. As a result, when any new case being assessed cannot satisfy any or all of the rules for one reason or the other, the non-deterministic problem arises. To curtail this problem, some studies have "reported that reduced data set (horizontally or vertically) is fed into neural network for complementing the limitation of RS, which finally produces full prediction of new case data" (Ahn et al., 2000, p. 68). Shin and Lee (2002) suggested the integration of multiple rules to solve the problem. For instance, if two of eight rules (two deterministic and six non-deterministic) show a new object as unhealthy, then the object is classified as unhealthy. Conclusively, it appears that there is no tool that clearly outperforms all other tools in relation to all result related criteria (Fig. 10).

Table 4
Summary of error types as reported for the tools by some of the authors.

S/N	Author year	SVM		ANN		DT		RS		GA		MDA		LR	
		Type I error	Type II error	Type I error	Type II error	Type I error	Type II error	Type I error	Type II error	Type I error	Type II error	Type I error	Type II error	Type I error	Type II error
1.	Kim and Kang (2010)			17.23	30.83										
2.	Yoon and Kwon (2010)	11.34	25.14												
3.	Du Jardin (2010)			4.72	7.22							16.8	8.8	9.58	6.4
4.	Lin et al. (2010)	5.56	5.56												
5.	Kim (2011)			4.8	12.1							47.6	27.4	22	18.4
6.	Yang et al. (2011)	8.93	17.2	16.07	26.56										
7.	Du Jardin and Séverin (2011)			17.95	16.82							18.41	17.73	18.18	19.55
8.	Chen, Ribeiro et al. (2011)	15.7	4.3	12.2	6.7					17.1	9.7				
9.	Chen, Yang et al. (2011)	26.55	18.96	18.52	21.71					14.94	17.02				
10.	Divsalar et al. (2012)			15.79						9.52	7.69			20	
11.	Tsai and Cheng (2012)	19.9	6.1	12.1	16.2	13.5	17.6							17.4	9.1
12.	Shie et al. (2012)		16.7		17.65		22.23								25
13.	Du Jardin and Séverin (2012)			20.1	17.4							22.1	15.5	20.1	16.6
14.	De Andrés et al. (2012)			26.52	21.71							28.7	22.08	25.67	21.35
15.	Lee and Choi (2013)			12.0	6.0							24	14		
16.	Tsai and Hsu (2013)			20.19	28.63	21.57	33.02							17.87	30.67
17.	Kasgari et al. (2013)			5.0	7.14										
18.	Tsai (2014)			6.87	10.09	9.21	17.82							13.79	11.36
19.	Yeh et al. (2014)	11.02	3.74	18.02	4.32	26.0	1.90	10.6	3.5						
20.	Wang et al. (2014)	21.55	18.19	20.62	27.69	23.10	24.74							26.38	25.38
21.	Gordini (2014)	22.9	38.1							21.1	35.8			23.3	43.1
22.	Iturriaga and Sanz (2015)	11.54	9.62	5.77	7.69							23.08	21.15	19.23	17.31

Table 5
Summary of tools that have been highlighted to be transparent, usable as hybrid, updatable, have overfitting problems etc., and year, country industry of the samples used in the reviewed studies.

S/N	Author	Journal of publication	Req. small sample size	Hybrid	Updateable	Black box	Transparent	Multicollinearity check	Overfit reported	Country
1.	Tseng and Hu (2010)	Expert Systems with Applications							NN	England
2.	Cho et al. (2010)	Expert Systems with Applications					DT			Korea
3.	Kim and Kang (2010)	Expert Systems with Applications		NN				NN	NN	Korea
4.	Yoon and Kwon (2010)	Expert Systems with Applications							NN, SVM	
5.	Du Jardin (2010)	.Neurocomputing								France
6.	Lin et al. (2010)	Journal of Marine Science and Technology								Taiwan
7.	Gepp et al. (2010)	Journal of forecasting					DT			US
8.	De Andrés et al. (2011)	Knowledge-Based Systems		NN						Spain
9.	Kim (2011)	Expert Systems with Applications				MDA	LR	MDA, LR	NN	Korea
10.	Yang et al. (2011)	Expert Systems with Applications	SVM	SVM		SVM, NN		SVM, NN		
11.	Chen (2011)	Neural Network World	SVM	SVM, GA				SVM, GA		Taiwan
12.	Divsalar et al. (2011)	Expert Systems			GA	NN	GA	GA		Iran
13.	Du Jardin and Séverin (2011)	Decision Support Systems						MDA, LR		France
14.	Chen, Ribeiro et al. (2011)	Knowledge-Based Systems							GA	France
15.	Chen, Yang et al. (2011)	Expert Systems with Applications		GA						
16.	Li et al. (2011)	Knowledge-Based Systems								China
17.	Divsalar et al. (2012)	Journal of Forecasting				NN	GA, LR		GA	Iran
18.	Huang et al. (2012)	Expert Systems with Applications	SVM	SVM, MDA					SVM, NN, DT	
19.	Tsai and Cheng (2012)	Knowledge-Based Systems								Japan
20.	Shie et al. (2012)	Neural Computing and Applications		SVM, GA			LR	DT		USA
21.	Kristóf and Virág (2012)	Acta Oeconomica				NN	DT, LR	LR	DT, LR	
22.	Jeong et al. (2012).	Expert Systems with Applications		SVM, NN, GA		SVM, NN			NN	Korea
23.	Du Jardin and Séverin (2012)	European Journal of Operational Research						MDA	NN	France
24.	De Andrés et al. (2012)	Knowledge-Based Systems				NN		MDA		Spain
25.	Zhou et al. (2012)	Computers & Mathematics with Applications								USA

(*continued on next page*)

Table 5 (*continued*)

S/N	Author	Journal of publication	Req. small sample size	Hybrid	Updateable	Black box	Transparent	Multicollinearity check	Overfit reported	Country
26.	Xiong et al. (2013)	Expert Systems with Applications		SVM						
27.	Lee and Choi (2013)	Expert Systems with Applications							NN	Korea
28.	Tsai and Hsu (2013)	Journal of Forecasting		NN, DT, LR						
29.	Callejón et al. (2013)	International Journal of Computational Intelligence Systems								Multiple
30.	Chuang (2013)	Information Sciences		CBR, RS, DT	CBR					
31.	Ho et al. (2013)	Empirical Economics								US
32.	Arieshanti et al. (2013)	TELKOMNIKA (Telecommunication Computing Electronics and Control)		SVM, NN, LR						
33.	Kasgari t al. (2013).	Neural Computing and Applications				NN			NN	Iran
34.	Zhou et al. (2014)	International Journal of Systems Science	SVM	SVM, NN, DT, GA		SVM, NN			SVM, NN, DT	US
35.	Tsai (2014)	Information Fusion		NN, DT, LR					NN	Australia
36.	Yeh et al. (2014)	Information Sciences		SVM, NN, DT, RS						
37.	Wang et al. (2014)	Expert Systems with Applications								US
38.	Abellán and Mantas (2014)	Expert Systems with Applications								
39.	Tserng et al. (2014)	Journal of Civil Engineering and Management							LR	US
40.	Yu et al. (2014)	Neurocomputing		SVM						France
41.	Gordini (2014)	Expert Systems with Applications					LR		SVM, GA	Italy
42.	Heo and Yang (2014)	Applied Soft Computing								Korea
43.	Tsai et al. (2014)	Applied Soft Computing								Japan
44.	Virág and Nyitrai (2014)	Acta Oeconomica							SVM, NN, RS	
45.	Liang et al. (2015)	Knowledge-Based Systems		SVM, NN, DT, GA						China
46.	Iturriaga and Sanz (2015)	Expert Systems with Applications				SVM, NN			SVM, NN	US
47.	Du Jardin (2015)	European Journal of Operational Research								France
48.	Bemš et al. (2015).	Expert Systems with Applications								
49.	Khademolqorani et al. (2015)	Mathematical Problems in Engineering				NN			NN, DT	Iran

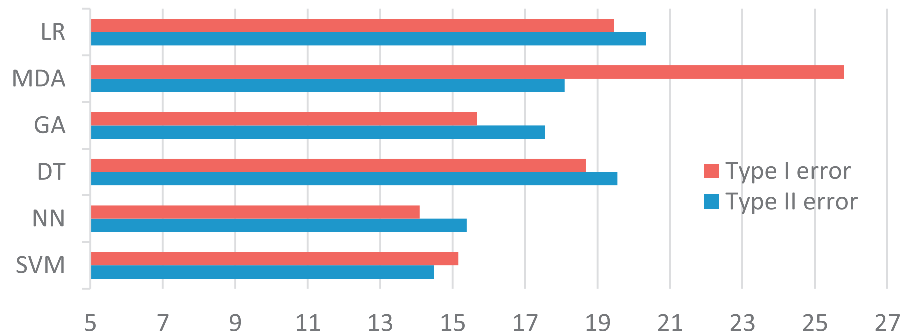

Fig. 7. Type I versus Type II error for each tool.

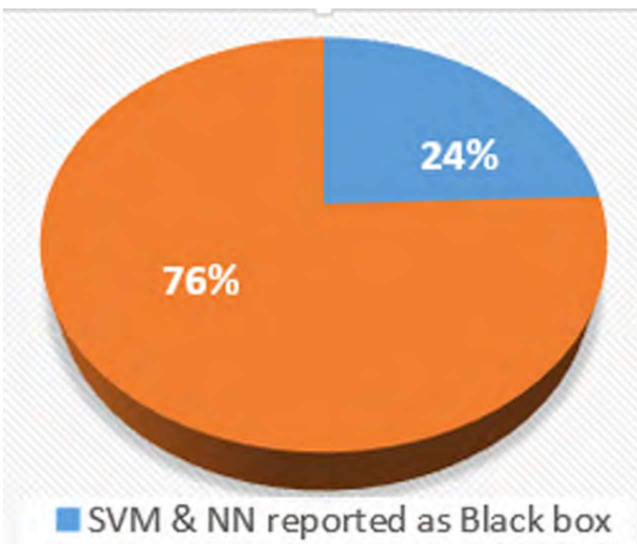

Fig. 8. Percentage of studies that complained/noted the non-transparent nature of SVM and ANN.

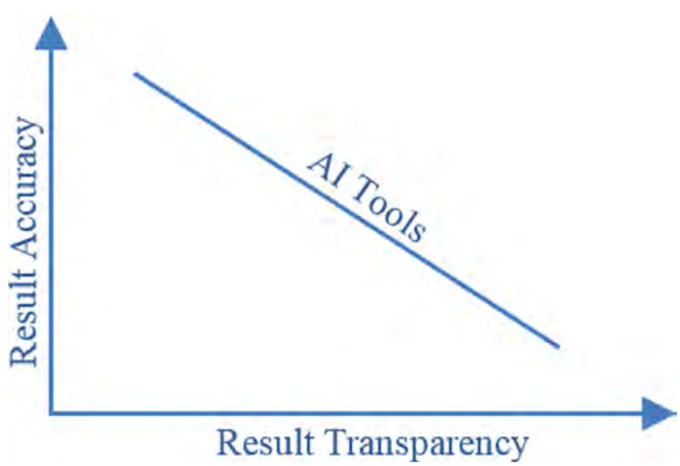

Fig. 9. Relationship between the accuracy and transparency of results of AI tools.

5.2. Data related criteria

5.2.1. Data dispersion and sample size capability

Data dispersion, i.e. ratio of number of non-failing sample firms to failing sample firms, is known to be key to performance; the relative ease with which data on existing firms can be gathered usually makes them dominate data and reduce performance. According to Du Jardin (2015), this normally means that "data that

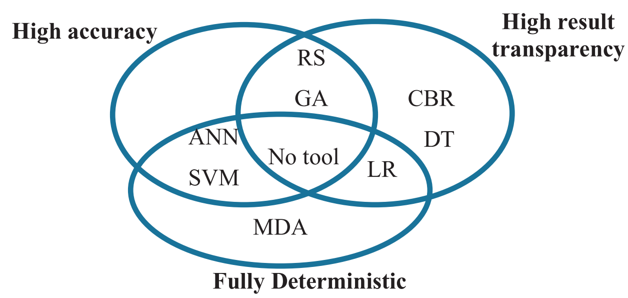

Fig. 10. Performance of tools in relation to results related criteria. There is no one tool that satisfies all the results related criteria required to develop a robust prediction model.

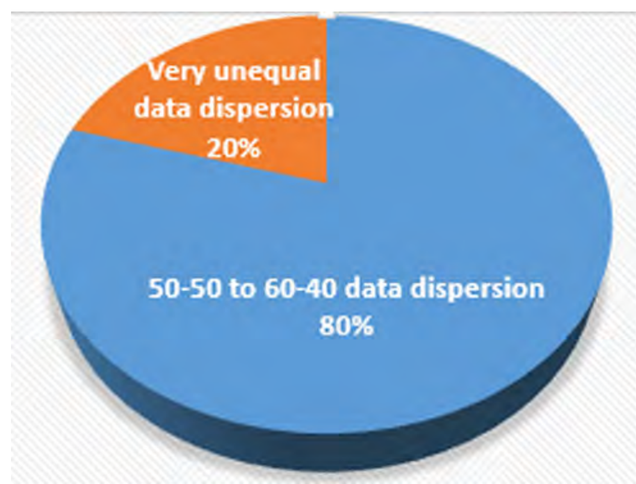

Fig. 11. Proportion of studies that used equal or almost equal data dispersion.

characterized failed firms would be hidden by those that represent non-failed firms, and therefore would become rather useless" (p.291) hence it is best to have equal dispersion (Jo, Han, & Lee, 1997).

MDA is quite sensitive to unequal dispersion (Balcaen & Ooghe, 2006). Compared to MDA, LR and Optimal Estimation Theory of ANN, are better with dispersion but ANN require the least dispersion at 20% failed firms before it could recognize pattern (Boritz et al., 1995; Du Jardin, 2015). However, no tool can perform reasonably well at this level of dispersion i.e. 20:80 (Boritz et al., 1995). The best option is to use equally dispersed data as most studies do. Most of the review studies have data dispersion ranging between 50–50 and 60–40 (Fig. 11) with nearly half using equally dispersed data (Table 2).

The sample size available for analysis can also influence the performance of a tool and should thus be given serious considera-

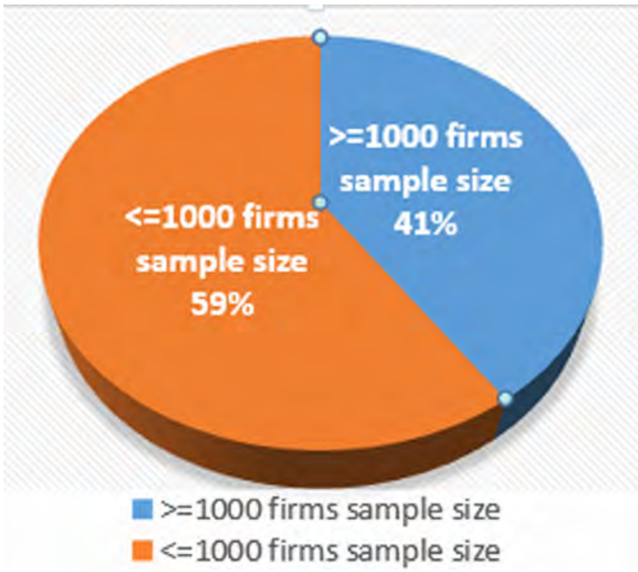

Fig. 12. Proportion of studies that used less or more than the 1000 firms sample size for ANN.

tion before selecting a tool. At least three of the reviewed studies (De Andrés, Landajo, & Lorca, 2012; Tseng & Hu, 2010; Zhou et al., 2014), and other studies (Haykin, 1994; Min & Lee, 2005; Ravi Kumar & Ravi, 2007; Shin et al., 2005) clearly indicated that ANNs and MDAs need a large training sample in order to reasonably recognize pattern and provide highly accurate classification. According to Haykin (1994), the minimum number of sample firms required to train an ANN network is ten times the weights in the network with an allowable error margin of 10%, i.e. over 1000 sample firms will be required to properly train a standard ANN to make it fit for generalization. This is not too commonly implemented in many ANN studies (Shin et al., 2005) as is evident in this study (Fig. 12). However, Lee, Booth, and Alam (2005) were able to show that ANNs can still perform reasonably well (better than statistical models) with a small number of sample firms provided 'a target vector is available'. Like with ANN, a primary study (Tseng & Hu, 2010) and another study (Ravi Kumar & Ravi, 2007) have reported DT and LR to require a large data set to perform well.

CBR, RS and SVM can handle small data size (Jo et al., 1997; Olmeda & Fernández, 1997; Ravi Kumar & Ravi, 2007). Although Buta (1994) claimed that CBR's accuracy increases with increase in data size, Ravi Kumar and Ravi (2007) made it clear that it cannot handle very large data. At least four of the reviewed studies confirmed SVM's special ability to perform well with a small training dataset (Table 5), with Zhou et al. (2014) noting in their wide experiment that "most SVM-based models can still keep higher performance as the size of training samples decreases. It demonstrates that SVM models can keep good performance with small training samples, which has been proved in many other applications also" (p.248). In Yang, You, and Ji (2011) experiment, they showed that for their SVM, "the support vector number is 33 and 35 … This shows that only 33 and 35 samples from the total of 120 samples are required to achieve the appropriate identification" (p.8340). In fact, Shin et al. (2005) did prove that SVM performs better and optimally with small training data sets as against a large one and fairs better than ANN only when a small data set is used to train both. This SVM's advantage is confirmed in older studies as well (e.g. Min & Lee, 2005; Ravi Kumar & Ravi, 2007; Shin et al., 2005).

5.2.2. Variable selection, multicollinearity and outliers

Statistical tools, especially LR, are highly sensitive and reactive to multicollinearity hence an effective method of choosing non-collinear variables is normally employed for them (Back, Laitinen, & Sere, 1996; Balcaen & Ooghe, 2006; Edmister, 1972; Joy & Tollefson, 1975; Lin & Piesse, 2004). Multicollinearity can easily lead to unstable performance and inaccurate results (Balcaen & Ooghe, 2006; Edmister, 1972; Joy & Tollefson, 1975). Before the emergence of AI tools, the most common variable selection method is the stepwise method because of its effectiveness in avoiding collinear variables (Altman, 1968; Back et al., 1996; Jo et al., 1997; Lin & Piesse, 2004). Its common use, over quarter of the studies used it, is usually to allow fair comparison with statistical tools.

The reviewed studies (Chen, 2011; Chen, Ribeiro et al., 2011; Liang et al., 2015; Yang et al., 2011) and other previous studies (Altman et al., 1994; Chung et al., 2008; Jo & Han, 1996) clearly indicate that AI tools, apart from CBR, are less sensitive to multicollinearity and can perform well with almost any variable selection method. CBR's performance decreases with increased number of variables (Chuang, 2013). On the other hand, some studies have claimed the higher the number of variables (usually when the multitude of variables available are used without selecting special ones), the better for ANN and GA (Chen, 2011; Chen, Ribeiro et al., 2011; Liang et al., 2015). In fact, Liang et al. (2015), who particularly investigated the effect of variable selection, concluded that "performing feature [variable] selection does not always improve the prediction performance" (p.289) of AI tools. However, Huang, Tang, Lee, and Chang (2012) feel removing irrelevant variables' can improve performance. Although Liang et al. (2015) found no best variable selection method in their study, they and Back et al. (1996) recommended GA as the best selection method for AI tools. Overall, it is not uncommon to use a decision rule generating AI tool to select variables for another AI tool as in some of the reviewed studies (Chen, 2011; Jeong et al., 2012; Liang et al., 2015; Zhou et al., 2014) and older studies (Ahn & Kim, 2009; Back et al., 1996; Wallrafen, Protzel, & Popp, 1996).

Although outliers can cause problems for any tool, LR has been particularly noted to be extremely sensitive to outliers in at least two of the reviewed studies (Kristóf & Virág, 2012; Tsai & Cheng, 2012). Outlier effects are normally reduced by normalising variables by industry average (McKee, 2000). Such normalization has however been found to reduce accuracy of models (Jo et al., 1997; Tam & Kiang, 1992).

5.2.3. Types of variables applicable

This criterion was not explicitly considered by the primary studies hence only the wider literature was used to discuss it. Although the vast majority of BPM studies use quantitative variables, usually in form of financial ratios, the need for qualitative/explanatory/managerial variables use, as noted in many studies, cannot be overemphasized (Abidali & Harris, 1995; Alaka et al., 2016; Argenti, 1980; Keasey & Watson, 1987; Zavgren, 1985; among others). MDAs can use only quantitative variables (Agarwa & Taffler, 2008; Altman, 1968; Bal, Cheung, & Wu, 2013; Chen, 2012; Odom & Sharda, 1990; Taffler, 1982 and more) while LR can use both (Cheng, Chen, & Fu, 2006; Keasey & Watson, 1987; Lin & Piesse, 2004; Ohlson, 1980; Tseng & Hu, 2010).

ANNs and SVMs can use mainly quantitative variables but can also use qualitative variables converted to quantitative variables using means such as the Likert scale (Cheng et al., 2006; Lin, 2009; StatSoft, 2014). All AI tools that yield the 'if… then,' decision rules for bankruptcy prediction, inclusive of RS, DT, CBR and GA, use qualitative variables and need quantitative variables to be converted to qualitative such as 'low, medium, high' etc. before they can be analysed making them suitable for use of combined variables (Dimitras, Slowinski, Susmaga, & Zopounidis, 1999; Martin, Balaji, & Venkatesan, 2012; Quinlan, 1986; Ravi Kumar & Ravi, 2007; Shin & Lee, 2002). The conversion is however not carried out by the AI and "involves dividing the original domain into subintervals which appropriately reflect theory and knowledge of the domain" (McKee, 2000, p. 165).

5.3. Tools' properties related criteria

5.3.1. Variables relationship capability and assumptions imposed by tools

Many independent variables used with BPM tools do not possess a linear relationship with the dependent variable (Balcaen & Ooghe, 2006; Keasey & Watson, 1991). Three of the reviewed studies (Divsalar et al., 2012; Du Jardin & Séverin, 2011; Du Jardin and Séverin, 2012) highlighted that MDA and LR require a linear and logistic relationship respectively between dependent and independent variables. This means important predictor variables with non-linear relationship to dependent variable will cause MDA to perform poorly. LR can solve logistic and non-linear problems (Jackson & Wood, 2013; Tam & Kiang, 1992). From this review, it appears all AI tools, except CBR (Chuang, 2013), can solve non-linear problems as identified by about a quarter of the reviewed studies (e.g. Chen, Ribeiro et al., 2011; Divsalar et al., 2011; Du Jardin & Séverin, 2011, 2012; Kasgari et al., 2013; Shie, Chen, & Liu, 2012; Yeh et al., 2014; Zhou et al., 2014; among others).

Du Jardin and Séverin (2011) and other studies (Balcaen & Ooghe, 2006; Chung et al., 2008; Coats & Fant, 1993; Lin & Piesse, 2004; among others) have shown that statistical tools require data to satisfy certain restrictive assumptions for optimal performance. Some of these assumptions include multivariate normality of independent variables, equal group variance-covariance, groups are discrete and non-overlapping etc. (Altman, 1993; Balcaen & Ooghe, 2006; Ohlson, 1980 Joy & Tollefson, 1975). All these restrictive assumptions can barely be satisfied together by one data set hence are violated in many studies (Chung et al., 2008; Richardson & Davidson, 1984; Zavgren, 1985). Nonetheless LR is deemed relatively less demanding compared to MDA (Altman, 1993; Balcaen & Ooghe, 2006; Jackson & Wood, 2013). On the other hand, none of the reviewed studies noted any restrictive assumptions on data for AI tools. This is because they look to extract knowledge from training samples or directly compare a new case to cases in the case library (Coats & Fant, 1993; Jackson & Wood, 2013; Lin, 2009; Shin & Lee, 2002).

5.3.2. Sample specificity/overfitting tendency and generalizability of tools

The common use of stepwise variable selection method and mainly financial ratios as variables for statistical tools sometimes lead to a sample specific model where the model performs excellently on the samples used to build it but woefully on hold out samples thereby possessing low generalizability (Agarwal & Taffler, 2008; Edmister, 1972; Lovell, 1983; Zavgren, 1985). LR nonetheless has a relatively reasonable generalizability (Dreiseitl & Ohno-Machado, 2002).

The equivalent of sample specificity in AI tools is called overfitting and is a common problem. There is also underfitting which is vice versa of overfitting. It is now a norm to avoid this problem (in statistical and AI tools) by testing models on a validation sample (and re-model if necessary) as indicated in most of the reviewed studies (Fig. 13a). Over a third of the reviewed studies also proactively identified this problem early (Fig. 13b) and considered it from the initial model development stage. Overfitting and underfitting are not necessarily caused by variable selection method or variable types in the case of AI tools. Apart from the case of CBR, it is generally known that the longer (shorter) the decision rules, the more the possibility of overfitting (underfitting) (Brodley & Utgoff, 1995; Clark & Niblett, 1989; Ravi Kumar & Ravi, 2007; Ren, 2012). CBRs tend not to overfit because they simply match a new case to one or more very similar cases in their library (Watson, 1997). CBR however has poor generalization but that is due to its poor accuracy (Ravi Kumar & Ravi, 2007).

Overfitting is a known problem of ANN and is as a result of overtraining the network (Ahn & Kim, 2009; Cheng et al., 2006; Jackson & Wood, 2013; Min & Lee, 2005; Tseng & Hu, 2010). Suggestions on how to construct more generalizable networks in ANN are given by Hertz et al. (1991). Overfitting (underfitting) in SVM is caused by a too large (small) upper bound value, usually denoted with '*C*' (Min & Lee, 2005; Shin et al., 2005). Thus, finding the optimum number of training and optimum *C* value for ANN and SVM respectively is key to their optimum performances. The notion that the structural risk minimization (SRM) used by SVM helps it to reduce the possibility of overfitting and increases generalization is not well proven according to Burges (1998). However, the tendency of overfitting in SVM is lower than in ANN and MDA (Cristianini & Shawe-Taylor, 2000; Kim, 2003; Shin et al., 2005).

5.3.3. Model development time, updatability and integration capability with other tools

Although the reviewed studies did not really touch on training times, past studies have noted that training AI tools, especially ANN and GA, can take a relatively longer time compared to statistical tools. This is because of the iterative process of finding the best parameters for AI tools (Jo & Han, 1996; Min & Lee, 2005; Ravi Kumar & Ravi, 2007). ANN architectures normally require many training cycles and GAs search for global optimum, while locating and negating local minima, make them (ANN and GA) take time for model development (Chung et al., 2008; Fletcher & Goss, 1993; Ravi Kumar & Ravi, 2007; Shin & Lee, 2002). For SVM, the polynomial function takes a long time but its RBF function is quicker (Huang, Chen, Hsu, Chen, & Wu, 2004; Kim, 2003). RS however does not take very long to train (Dimitras et al., 1999).

As noted in the reviewed studies, CBR and GA create the most updatable BPMs (Table 5). CBR is easy to update and quite effective after an update since all it takes is to simply add new cases to its case library and prediction of a new case is done by finding the most similar cases(s) among all cases, old and new, in the library (Ahn & Kim, 2009; Bryant, 1997). An attempted update of a statistical BPM can lead to much reduced accuracy (Charitou, Neophytou, & Charalambous, 2004; Mensah, 1984). ANNs can be adaptively updated with new samples since they are known to be robust on sample variations (Altman, 1993; Tam & Kiang, 1992; Zhang et al., 1999). However, if the situations of the new cases are significantly different for the ones used to build the model, then a new model must be developed (Chung et al., 2008). RS is particularly very sensitive to changes in data and can really be ineffective after an update with data that has serious sample variations (Ravi Kumar & Ravi, 2007).

AI tools are more flexible and allow integration with other tools better than statistical tools do. This is evident from the reviewed studies as more of the studies that used AI tools produced hybrids with them than those that used statistical tools (Figs 14a and b). The review clearly indicates that effective hybrids perform better than standalone tools (Iturriaga & Sanz, 2015; Tsai, 2014; Zhou et al., 2014), and "usually outperforms even the MLP [a type of ANN] and SVM procedure" (Iturriaga & Sanz, 2015, p.2866). This is also confirmed in older studies (Ahn & Kim, 2009; Ahn et al., 2000; Jeng et al., 1997; Jo & Han, 1996). "However, these hybrid models consume more computational time" (Zhou et al., 2014, p.251) and "it is unknown which type of the prediction models by classifier ensembles and hybrid classifiers can perform better". (Tsai, 2014, p.50).

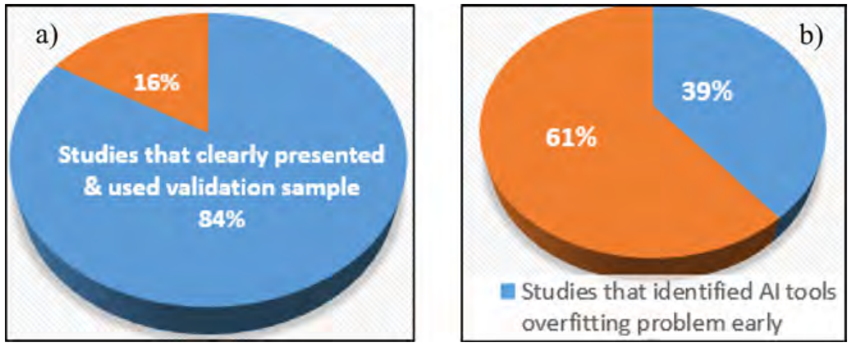

Fig. 13. Proportion of studies that identified overfitting problem early and those that solved the problem using validation sample.

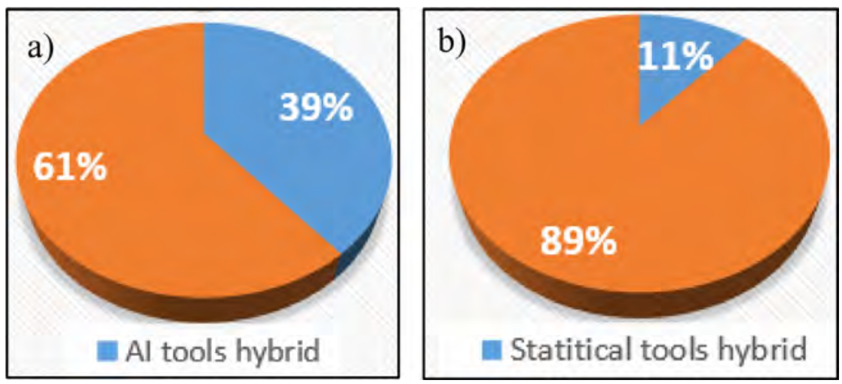

Fig. 14. Proportion of studies that integrated AI or statistical tools to form a hybrid.

Table 6
A tabular framework of tools' performance in relation to important BPMs criteria.

Important criteria/Tools	Tools category							
	Statistical		AI tools					
	MDA	LR	ANN	SVM	RS	GA	DT	CBR
Accuracy	Low	Mod.	V. High	V. High	High	High	Mod.	Low
Result transparency	Low	High	Low	Low	High	High	High	High
Can be Non-deterministic	No	No	No	No	Yes	Yes	Yes	Yes
Ability to use small Samples size	Low	Low	Low	V. high	high	NR	low	high
Data dispersion sensitivity	High	Normal	High	NR	NR	NR	NR	NR
Suitable variable selection	SW	SW	Any	Any	Any	Any	Any	Any
Multicollinearity Sensitivity	High	V. High	Low	Low	Low	Low	Low	Low
Sensitivity to outlier	Mod.	High	Mod.	Mod.	Mod.	Mod.	Mod.	Mod.
Variable type used	QN	Both	QN (both)	QN (both)	QL (both)	(both)	(both)	QL (both)
Variable relationship required	Linear	Logistic	Any	Any	Any	Any	Any	Linear
Other Assumptions to be satisfied	Many	Some	None	None	None	None	None	None
Overfitting possibility	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Updatability	Poor	Poor	OK	–	Poor	OK/good	Poor	Good
Ways to integrate to give hybrid	Few	Few	Many	Many	Many	Many	Many	Many
Output Mode	Cut-off	Binary	Binary	Binary	DR	DR	DR	DR

Note: All rankings are relative. NR: Not Reported SW: Stepwise V.: Very Mod: moderate QN: Quantitative QL: Qualitative DR: Decision rules.

6. The proposed model

Fig. 15 presents a diagrammatic framework, gotten from the result of this review, which serves as a guideline for a BPM developer to select the right tool(s) that is best suited to available data and BPM preference criteria. Virtually all tools that are used for developing BPMs can successfully make predictions. However, some tools are more powerful in relation to certain criteria than others (see Table 6).

The framework clearly shows that to get the best performance from a BPM, the developing tool should be selected based on the output criteria preferences and the characteristics of data available. The framework is a very good starting point for any BPM developer and will ensure tools are not selected arbitrarily to the disadvantage of the developer. It will also ensure the final user of the BPM, having communicated his requirements to the model developer, gets the most appropriate BPM. For example, a BPM developer that considers accuracy as the highest preference because of his client's requirements, but has a very small dataset will not be wrongly choosing the highly accurate ANN for his model if this framework is used; SVM will be the right tool in such a circumstance.

The implication of using this framework on practice is that it will allow tools to be used to the best of their strengths and encourage BPMs to be developed in a more customized way to customer/client requirements. This is better than to continue with the

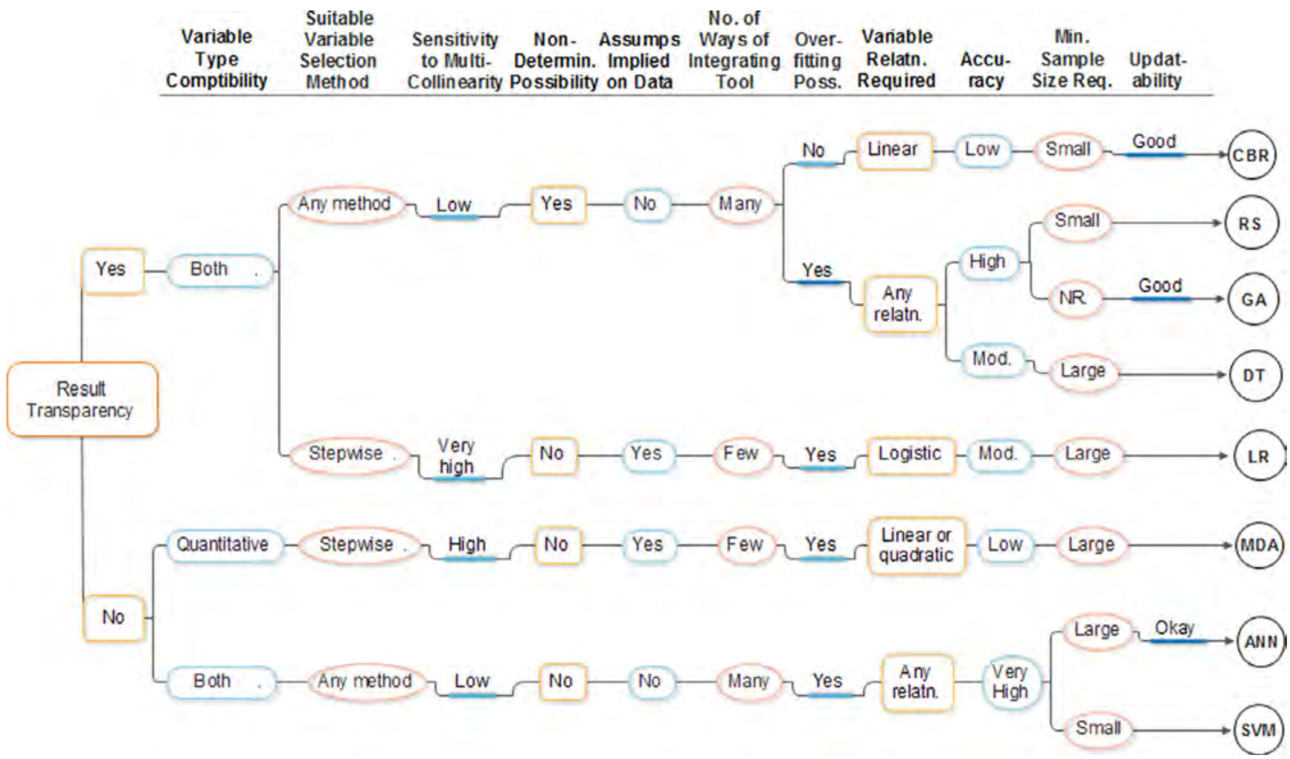

Fig. 15. A framework for selection of the most suitable tools for various situations
Determin: Deterministic Assumps: Assumptions No.: Number Poss: Possibility Relatn: Relationship Min: Minimum Req: Required Mod.: Moderate.

present trend of 'one size fits all' where a BPM is assumed to be good enough for the very different users/clients e.g. financiers, clients, owners, government agencies, auditors etc. It will also eliminate the time-wasting process of developing multiple BPMs with multiple tools in order to select the best after a series of test. The implication of this work on research is that it will guide researchers in selecting the best tool for their data and situation and help avoid the arbitrary selection of tools or selection simply based on popularity. It will also inform researchers on the need to use hybrid tools the more if the 'one size fits all' tool has to be achieved. It will hence invoke the development of new hybrid models.

7. Conclusion

The bankruptcy prediction research domain continues to evolve with many new models developed using various tools. Yet many of the tools are used with the wrong data conditions or for the wrong situation. This study used a systematic review, to reveal how eight popular and promising tools (MDA, LR, ANN, SVM, RS, CBR, DT and GA) perform with regard to various important criteria in the bankruptcy prediction models (BPM) study area. Overall, it can be concluded that there is no singular tool that is predominantly better than all other tools in relation to all identified criteria. It is however clear that each tool has its strengths and weaknesses that make it more suited to certain situations (i.e. data characteristics, developer aim, among others) than others. The framework presented in this study clearly provides a platform that allows a well-informed selection of tool(s) that can best fit the situation of a model developer.

The implication of this study is that BPM developers can now make an informed decision when selecting a tool for their model rather than make selection based on popularity or other unscholarly factors. In essence, tools will be more regularly selected based on their strength. Another implication is that BPMs with better performance with regards to end users' requirement will be more commonly developed. This is better than to continue with the present trend of 'one size fits all' where a BPM tool is assumed to be good enough for the very different users/clients (e.g. financiers, clients, owners, government agencies, auditors, among others) that need them. The framework in this study will also reduce the time-wasting process of developing many BPMs with different tools in order to select the best after a series of test; only the tools that best fit a developer's situation will be used and compared.

Future studies should consider possibilities of making ANN and SVM results interpretable since they appear to be the most accurate tools and satisfy a number of criteria for BPM. The very best overall model that will outperform all others in relation to all or most criteria, though not yet found, might come in the form of a hybrid of tools. Future research should thus, on one hand, explore various hybrids with the aim of developing the best hybrid that can achieve this fit. On the other hand, future studies should consider use of more sophisticated tools like Bart machine, extremely randomized trees, gradient boosting machine and extreme Gradient Boosting among others, as they might be the answer. As tools that can handle qualitative variables have been exposed in this study, future studies should consider the inclusion of qualitative variables in their BPM development as suggested in many studies (e.g. Abidali & Harris, 1995; Alaka et al., 2017; Argenti, 1980; among others). Further, future studies should look to combine qualitative and qualitative variables using these tools, with a view of developing better performing BPMs. Finally, future research should attempt to confirm/falsify the SVM tool's excellent capability to make predictions on small data size since it is generally known that the larger the data, the better.

References

Abellán, J., & Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 41*(8), 3825–3830.

Abidali, A. F., & Harris, F. (1995). A methodology for predicting failure in the construction industry. *Construction Management and Economics, 13*(3), 189–196.

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.

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.

Ahn, H., & Kim, K. J. (2009). Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing, 9*(2), 599–607.

Alaka, H., Oyedele, L., Toriola-Coker, O., Owolabi, H., Akinade, O., Bilal, M., et al. (2015). Methodological approach of construction businesses failure prediction studies: A review. In A. B. Raidén, & E. Aboagye-Nimo (Eds.), Proceedings 31st annual ARCOM conference, *7-9 September 2015, Lincoln, UK, Association of Researchers in Construction Management* (pp. 1291–1300).

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Ajayi, S. O., Bilal, M., & Akinade, O. O. (2016). Methodological approach of construction business failure prediction studies: A review. *Construction Management and Economics, 34*(11), 808–842.

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Oyedele, A. A., Akinade, O. O., Bilal, M., et al. (2017). Critical factors for insolvency prediction: Towards a theoretical model for the construction industry. *International Journal of Construction Management, 17*(1), 25–49.

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

Altman, E. (1993). *Corporate financial distress and bankruptcy* (2nd ed.). New York: John Wiley and Sons.

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of banking and finance, 18*(3), 505–529.

Appiah, K. O., Chizema, A., & Arthur, J. (2015). Predicting corporate failure: A systematic literature review of methodological issues. *International Journal of Law and Management, 57*(5), 461–485.

Argenti, J. (1980). *Practical corporate planning*. London: Allen and Unwin.

Arieshanti, I., Purwananto, Y., Ramadhani, A., Nuha, M. U., & Ulinnuha, N. (2013). Comparative study of bankruptcy prediction models. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 11*(3), 591–596.

Aziz, M. A., & Dar, H. A. (2006). Predicting corporate bankruptcy: Where we stand. *Corporate Governance, 6*(1), 18–33.

Back, B., Laitinen, T., & Sere, K. (1996). Neural networks and genetic algorithms for bankruptcy predictions. *Expert Systems with Applications, 11*(4), 407–413.

Bal, J., Cheung, Y., & Wu, H. C. (2013). Entropy for business failure prediction: An improved prediction model for the construction industry. *Advances in Decision Sciences, 2013*, 1–14.

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

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

Beaver, W. H., McNichols, M. F., & Rhie, J. W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies, 10*(1), 93–122.

Bemš, J., Starý, O., Macaš, M., Žegklitz, J., & Pošík, P. (2015). Innovative default prediction approach. *Expert Systems with Applications, 42*(17), 6277–6285.

Boritz, J. E., & Kennedy, D. B. (1995). Effectiveness of neural network types for prediction of business failure. *Expert Systems with Applications, 9*(4), 503–512.

Boritz, J. E., Kennedy, D. B., & Albuquerque, A. M. (1995). Predicting corporate failure using a neural network approach. *Intelligent Systems in Accounting, Finance and Management, 4*(2), 95–111.

Brodley, C. E., & Utgoff, P. E. (1995). Multivariate decision trees. *Machine Learning, 19*(1), 45–77.

Bryant, S. M. (1997). A case-based reasoning approach to bankruptcy prediction modeling. *Intelligent Systems in Accounting, Finance and Management, 6*(3), 195–214.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Buta, P. (1994). Mining for financial knowledge with CBR. *Ai Expert, 9*(2), 34–41.

Callejón, A. M., Casado, A. M., Fernández, M. A., & Peláez, J. I. (2013). A system of insolvency prediction for industrial companies using a financial alternative model with neural networks. *International Journal of Computational Intelligence Systems, 6*(1), 29–37.

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

Chen, H. L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S. J., et al. (2011). A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method. *Knowledge-Based Systems, 24*(8), 1348–1359.

Chen, J. H. (2012). Developing SFNN models to predict financial distress of construction companies. *Expert Systems with Applications, 39*(1), 823–827.

Chen, M. Y. (2011). A hybrid model for business failure prediction-utilization of particle swarm optimization and support vector machines. *Neural Network World, 21*(2), 129–152.

Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J., & Neves, J. C. (2011). A genetic algorithm-based approach to cost-sensitive bankruptcy prediction. *Expert Systems with Applications, 38*(10), 12939–12945.

Cheng, C. B., Chen, C. L., & Fu, C. J. (2006). Financial distress prediction by a radial basis function network with logit analysis learning. *Computers and Mathematics with Applications, 51*(3–4), 579–588.

Cho, S., Hong, H., & Ha, B. C. (2010). A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3482–3488.

Chuang, C. L. (2013). Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction. *Information Sciences, 236*, 174–185.

Chung, H. M. M., & Tam, K. Y. (1993). A comparative analysis of inductive-learning algorithms. *Intelligent Systems in Accounting, Finance and Management, 2*(1), 3–18.

Chung, K. C., Tan, S. S., & Holdsworth, D. K. (2008). Insolvency prediction model using multivariate discriminant analysis and artificial neural network for the finance industry in New Zealand. *International Journal of Business and Management, 39*(1), 19–28.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning, 3*(4), 261–283.

Coats, P. K., & Fant, L. F. (1993). Recognizing financial distress patterns using a neural network tool. *Financial Management, 22*(3), 142–155.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

De Andrés, J., Lorca, P., de Cos Juez, F. J., & Sánchez-Lasheras, F. (2011). Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). *Expert Systems with Applications, 38*(3), 1866–1875.

De Andrés, J., Landajo, M., & Lorca, P. (2012). Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios. *Knowledge-Based Systems, 30*, 67–77.

Dimitras, A. I., Slowinski, R., Susmaga, R., & Zopounidis, C. (1999). Business failure prediction using rough sets. *European Journal of Operational Research, 114*(2), 263–280.

Divsalar, M., Firouzabadi, A. K., Sadeghi, M., Behrooz, A. H., & Alavi, A. H. (2011). Towards the prediction of business failure via computational intelligence techniques. *Expert Systems, 28*(3), 209–226.

Divsalar, M., Roodsaz, H., Vahdatinia, F., Norouzzadeh, G., & Behrooz, A. H. (2012). A robust data-mining approach to bankruptcy prediction. *Journal of Forecasting, 31*(6), 504–523.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*(5), 352–359.

Du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing, 73*(10), 2047–2060.

Du Jardin, P. (2015). Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research, 242*(1), 286–303.

Du Jardin, P., & Séverin, E. (2011). Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model. *Decision Support Systems, 51*(3), 701–711.

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.

Edmister, R. O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis, 7*(02), 1477–1493.

Edum-Fotwe, F., Price, A., & Thorpe, A. (1996). A review of financial ratio tools for predicting contractor insolvency. *Construction Management and Economics, 14*, 189–198.

Fletcher, D., & Goss, E. (1993). Forecasting with neural networks: An application using bankruptcy data. *Information and Management, 24*(3), 159–167.

Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting, 29*(6), 536–555.

Gordini, N. (2014). A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. *Expert Systems with Applications, 41*(14), 6433–6445.

Greco, S., Matarazzo, B., & Slowinski, R. (2001). Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research, 129*(1), 1–47.

Hafiz, A., Lukumon, O., Muhammad, B., Olugbenga, A., Hakeem, O., & Saheed, A. (2015). Bankruptcy prediction of construction businesses: Towards a big data analytics approach. In *Big data computing service and applications (BigDataService), 2015 IEEE first international conference on, March* (pp. 347–352). IEEE.

Haykin, S. (1994). *Neural networks: A comprehensive foundation*. New York: Mc Millan.

Heo, J., & Yang, J. Y. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing, 24*, 494–499.

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computing*. New York: Addison Wesley.

Higgins, J. P. (Ed.) (2008). *Cochrane handbook for systematic reviews of interventions*: vol. 5. Chichester, England: Wiley-Blackwell.

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.

Ho, C. Y., McCarthy, P., Yang, Y., & Ye, X. (2013). Bankruptcy in the pulp and paper industry: Market's reaction and prediction. *Empirical Economics, 45*(3), 1205–1232.

Huang, S. C., Tang, Y. C., Lee, C. W., & Chang, M. J. (2012). Kernel local Fisher discriminant analysis based manifold-regularized SVM model for financial distress predictions. *Expert Systems with Applications, 39*(3), 3855–3861.
Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems, 37*(4), 543–558.
Iturriaga, F. J. L., & Sanz, I. P. (2015). Bankruptcy visualization and prediction using neural networks: A study of US commercial banks. *Expert Systems with Applications, 42*(6), 2857–2869.
Jackson, R. H., & Wood, A. (2013). The performance of insolvency prediction and credit risk models in the UK: A comparative study. *The British Accounting Review, 45*(3), 183–202.
Jeng, B., Jeng, Y. M., & Liang, T. P. (1997). FILM: A fuzzy inductive learning method for automated knowledge acquisition. *Decision Support Systems, 21*(2), 61–73.
Jeong, C., Min, J. H., & Kim, M. S. (2012). A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction. *Expert Systems with Applications, 39*(3), 3650–3658.
Jo, H., & Han, I. (1996). Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. *Expert Systems with applications, 11*(4), 415–422.
Jo, H., Han, I., & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications, 13*(2), 97–108.
Joy, O. M., & Tollefson, J. O. (1975). On the financial applications of discriminant analysis. *Journal of Financial and Quantitative Analysis, 10*(5), 723–739.
Kasgari, A. A., Divsalar, M., Javid, M. R., & Ebrahimian, S. J. (2013). Prediction of bankruptcy Iranian corporations through artificial neural network and Probit-based analyses. *Neural Computing and Applications, 23*(3-4), 927–936.
Keasey, K., & Watson, R. (1987). Non-financial symptoms and the prediction of small company failure: A test of Argenti's hypotheses. *Journal of Business Finance and Accounting, 14*(3), 335–354.
Keasey, K., & Watson, R. (1991). Financial distress prediction models: A review of their usefulness[1]. *British Journal of Management, 2*(2), 89–102.
Khademolqorani, S., Zeinal Hamadani, A., & Mokhatab Rafiei, F. (2015). A hybrid analysis approach to improve financial distress forecasting: Empirical evidence from Iran. *Mathematical Problems in Engineering, 2015*, 1–9.
Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine, 96*(3), 118–121.
Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1), 307–319.
Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications, 37*(4), 3373–3379.
Kim, S. Y. (2011). Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. *The Service Industries Journal, 31*(3), 441–468.
Kolodner, J. (1993). *Case-based reasoning*. San Francisco, California: Morgan Kaufmann Publishers Inc.
Koyuncugil, A. S., & Ozgulbas, N. (2012). Financial early warning system model and data mining application for risk detection. *Expert Systems with Applications, 39*(6), 6238–6253.
Kristóf, T., & Virág, M. (2012). Data reduction and univariate splitting—Do they together provide better corporate bankruptcy prediction? *Acta Oeconomica, 62*(2), 205–228.
Langford, D., Iyagba, R., & Komba, D. M. (1993). Prediction of solvency in construction companies. *Construction Management and Economics, 11*, 317–325.
Lee, K., Booth, D., & Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications, 29*(1), 1–16.
Lee, S., & Choi, W. S. (2013). A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. *Expert Systems with Applications, 40*(8), 2941–2946.
Li, H., Lee, Y. C., Zhou, Y. C., & Sun, J. (2011). The random subspace binary logit (RSBL) model for bankruptcy prediction. *Knowledge-Based Systems, 24*(8), 1380–1388.
Liang, D., Tsai, C. F., & Wu, H. T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73*, 289–297.
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.
Lin, F. Y., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems, 14*(3), 189–195.
Lin, L., & Piesse, J. (2004). Identification of corporate distress in UK industrials: A conditional probability analysis approach. *Applied Financial Economics, 14*(2), 73–82.
Lin, T. H. (2009). A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models. *Neurocomputing, 72*(16), 3507–3516.
Lovell, M. C. (1983). Data mining. *Review of Economic Statistics, 65*(1), 1–12.
Martens, D., Van Gestel, T., De Backer, M., Haesen, R., Vanthienen, J., & Baesens, B. (2010). Credit rating prediction using ant colony optimization. *Journal of the Operational Research Society, 61*(4), 561–573.
Martin, A., Manjula, M., & Venkatesan, D. V. P. (2011). A business intelligence model to predict bankruptcy using financial domain ontology with association rule mining algorithm. *International Journal of Computer Science Issues, 8*(3) 1694-0814.
Martin, A., Balaji, S., & Venkatesan, V. P. (2012). Effective prediction of bankruptcy based on the qualitative factors using FID3 algorithm. *International Journal of Computer Applications (0975 – 8887), 43*(21), 28–32.
Mason, R. J., & Harris, F. C. (1979). Predicting company failure in the construction industry. In *Proceedings of institution of civil engineers, 01 May 1979, 66 (2)* (pp. 301–307). London: Thomas Telford.
Mensah, Y. M. (1984). An examination of the stationarity of multivariate bankruptcy prediction models: A methodological study. *Journal of Accounting Research, 22*(1), 380–395.
McKee, T. E. (2000). Developing a bankruptcy prediction model via rough sets theory. *Intelligent Systems in Accounting, Finance and Management, 9*(3), 159–173.
McKee, T. E. (2003). Rough sets bankruptcy prediction models versus auditor signalling rates. *Journal of Forecasting, 22*(8), 569–586.
Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), 603–614.
Min, S. H., Lee, J., & Han, I. (2006). Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert Systems with Applications, 31*(3), 652–660.
Nasir, M. L., John, R. I., Bennett, S. C., Russell, D. M., & Patel, A. (2000). Predicting corporate bankruptcy using artificial neural networks. *Journal of Applied Accounting Research, 5*(3), 30–52.
Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. In *IJCNN international joint conference on neural networks. San Diego, California, 17-21 June 1990* (pp. 163–168). IEEE.
Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109–131.
Olmeda, I., & Fernández, E. (1997). Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction. *Computational Economics, 10*(4), 317–335.
Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences, 11*(5), 341–356.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.
Quinlan, J. R. (1993). *Programs for machine learning*. San Mateo, CA: Morgan Kaufmann Series in Machine Learning.
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.
Ren, D. (2012). Application attributes reduction of rough set in power system for fault diagnosis. *Journal of Convergence Information Technology, 7*(13), 300–308.
Richardson, F. M., & Davidson, L. F. (1984). On linear discrimination with accounting ratios. *Journal of Business Finance and Accounting, 11*(4), 511–525.
Schlosser, R. W. (2007). Appraising the quality of systematic reviews. *Focus: Technical Briefs*, (17), pp. 1–8.
Shie, F. S., Chen, M. Y., & Liu, Y. S. (2012). Prediction of corporate financial distress: An application of the America banking industry. *Neural Computing and Applications, 21*(7), 1687–1696.
Shin, K. S., & Lee, Y. J. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications, 23*(3), 321–328.
Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications, 28*(1), 127–135.
Smith, V., Devane, D., Begley, C. M., & Clarke, M. (2011). Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Medical Research Methodology, 11*(1), 15.
StatSoft (2014). Support Vector Machines (SVM) [online]. Available from: http://www.statsoft.com/Textbook/Support-Vector-Machines#index [Accessed 26 September 2016]
Stenbacka, C. (2001). Qualitative research requires quality concepts of its own. *Management Decision, 39*(7), 551–556.
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.
Taffler, R. J. (1982). Forecasting company failure in the UK using discriminant analysis and financial ratio data. *Journal of the Royal Statistical Society. Series A (General), 145*, 342–358.
Taffler, R. J. (1983). The assessment of company solvency and performance using a statistical model. *Accounting and Business Research, 13*(52), 295–308.
Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science, 38*(7), 926–947.
Tranfield, D. R., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*, 207–222.
Trochim, W. M., & Donnelly, J. P. (2006). *The research methods knowledge base* (3rd ed.). Ohio: Atomic Dog Publishing.
Tsai, C. F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion, 16*, 46–58.
Tsai, C. F., & Cheng, K. C. (2012). Simple instance selection for bankruptcy prediction. *Knowledge-Based Systems, 27*, 333–342.
Tsai, C. F., & Hsu, Y. F. (2013). A meta-learning framework for bankruptcy prediction. *Journal of Forecasting, 32*(2), 167–179.
Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing, 24*, 977–984.
Tseng, F. M., & Hu, Y. C. (2010). Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks. *Expert Systems with Applications, 37*(3), 1846–1853.

Tserng, H. P., Chen, P. C., Huang, W. H., Lei, M. C., & Tran, Q. H. (2014). Prediction of default probability for construction firms using the logit model. *Journal of Civil Engineering and Management, 20*(2), 247–255.

Vapnik, V. (1998). *Statistical learning theory*: vol. 2. New York: Springer.

Virág, M., & Nyitrai, T. (2014). Is there a trade-off between the predictive power and the interpretability of bankruptcy models? The case of the first Hungarian bankruptcy prediction model. *Acta Oeconomica, 64*(4), 419–440.

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.

Wallrafen, J., Protzel, P., & Popp, H. (1996). Genetically optimized neural network classifiers for bankruptcy prediction-an empirical study. In *Proceedings of the twenty-ninth hawaii international conference on system sciences. Hawaii, USA, 3-6 Jan. 1996. Washington: vol. 2* (pp. 419–426). IEEE Computer Society Press, cop..

Watson, I. (1997). *Applying case-based reasoning: Techniques for enterprise systems*. San Francisco, California: Morgan Kaufmann Publishers Inc.

Won, C., Kim, J., & Bae, J. K. (2012). Using genetic algorithm based knowledge refinement model for dividend policy forecasting. *Expert Systems with Applications, 39*(18), 13472–13479.

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems, 11*(5), 545–557.

Xiong, T., Wang, S., Mayers, A., & Monga, E. (2013). Personal bankruptcy prediction by mining credit card data. *Expert Systems with Applications, 40*(2), 665–676.

Xu, M., & Zhang, C. (2009). Bankruptcy prediction: The case of Japanese listed companies. *Review of Accounting Studies, 14*(4), 534–558.

Yang, Z., You, W., & Ji, G. (2011). Using partial least squares and support vector machines for bankruptcy prediction. *Expert Systems with Applications, 38*(7), 8336–8342.

Yeh, C. C., Chi, D. J., & Lin, Y. R. (2014). Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences, 254*, 98–110.

Yoon, J. S., & Kwon, Y. S. (2010). A practical approach to bankruptcy prediction for small businesses: Substituting the unavailable financial data for credit card sales information. *Expert Systems with Applications, 37*(5), 3624–3629.

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

Zavgren, C. V. (1985). Assessing the vulnerability to failure of American industrial firms: A logistic analysis. *Journal of Business Finance and Accounting, 12*(1), 19–45.

Zhang, G., Hu, M. Y., Patuwo, E. B., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research, 116*(1), 16–32.

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.

Zhou, L., Lai, K. K., & Yen, J. (2014). Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation. *International Journal of Systems Science, 45*(3), 241–253.

Zurada, J. M., Malinowski, A., & Cloete, I. (1994). Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *IEEE International Symposium on Circuits and Systems, ISCAS'94, London, United Kingdom, 30 May – 2 June, 1994: vol. 6* (pp. 447–450).