INTELLIGENT FINANCIAL DISTRESS PPREDICTION – RECENT CONTRIBUTIONS AND THEIR RESPONSE TO THE PROBLEMS OF CLASSIC PREDICTION METHODOLOGIES

Lappeenranta–Lahti University of Technology LUT
Master's Programme in Strategic Finance and Analytics
2021
Lauri Jaatinen
Examiners: Professor Eero Pätäri
Professor Sheraz Ahmed

ABSTRACT

Lappeenranta–Lahti University of Technology LUT
LUT School of Business and Management
Strategic Finance and Analytics

Lauri Jaatinen

Intelligent financial distress prediction – Recent contributions and their response to the problems of classic prediction methodologies

Master's thesis
2021
103 pages, 5 figures, 5 tables and 5 appendices
Examiners: Professor Eero Pätäri and Professor Sheraz Ahmed

Keywords: Financial distress, Prediction, Intelligent methods, Classification learning

This thesis examines current trends in the intelligent financial distress prediction field, and how recently published studies respond to the problems in classic financial distress prediction methodologies. Financial distress prediction is an extensively studied subject, where the goal is to find a classifier that describes whether a firm belongs to a non-distressed or distressed state in the future. Traditional statistical methods, like discriminant analysis and logistic regression, are commonly used to derive the optimal classification function. However, more and more models with intelligent methods are developed due to their higher predictive performance and flexibility. Intelligent methods have introduced novel and high-performing models, but they cannot solve all the problems in classic prediction methodologies by themselves. Indeed, most of the difficulties are not strictly dependent on the prediction model but on common practices and assumptions of the process. Therefore, the thesis analyses 36 peer-reviewed, intelligent financial distress prediction studies to see current trends in the field, and investigates whether the studies respond to the fundamental problems. In addition, improvements and suggestions for future research are given.

The results showed that the studies had various research areas and dispersed objectives, which indicates a certain level of complexity and multidimensionality around the subject. Multiple classifier systems, or ensemble methods, were the most popular and the highest performing methods. Imbalanced datasets were applied more often than balanced, but only a minority of the studies addressed the problems that arise due to this practice. Responses to the problems in classic prediction methodologies were somewhat insufficient. Although studies implemented dynamic properties and sophisticated techniques in the modelling phase, most of the problems remained unsolved. For the future, in-depth theoretical analysis is needed to find meaningful features and eliminate unsuitable practices in the process. Also, studies are encouraged to continue with imbalanced datasets and to implement dynamic modelling practices for a more realistic modelling framework. Comparative studies of ensemble methods and alternative features and feature types should be investigated further.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT
LUT-kauppakorkeakoulu
Strategic Finance and Analytics

Lauri Jaatinen

Älykäs maksuvaikeuden ennustaminen – Viimeisimmät tutkimukset ja niiden vastaukset klassisten ennustemenetelmien ongelmiin

Kauppatieteiden pro gradu -tutkielma
103 sivua, 5 kuvaa, 5 taulukkoa ja 5 liitettä
Tarkastajat: professori Eero Pätäri ja professori Sheraz Ahmed

Avainsanat: Maksuvaikeus, Ennustaminen, Älykkäät menetelmät, Klassifiointi

Tämä tutkielma tarkastelee älykkään maksuvaikeuden ennustamisen nykyisiä trendejä ja viimeaikaisten akateemisten julkaisujen tarjoamia ratkaisumalleja klassisten ennustemenetelmien ongelmiin. Maksuvaikeuden ennustamisen tavoitteena on löytää luokittelukriteeristö, joka pystyisi mahdollisimman luotettavasti erottelemaan terveet yritykset lähitulevaisuudessa taloudellisiin vaikeuksiin ajautuvista kriisiyrityksistä. Perinteisiä tilastollisia menetelmiä, kuten erotteluanalyysia ja logistista regressiota, on tyypillisesti käytetty optimaalisen luokittelukriteeristön määrittämiseen. Luokittelukriteeristöjä kehitetään nykyisin yhä enemmän älykkäillä menetelmillä johtuen niiden paremmasta ennustetarkkuudesta ja joustavuudesta. Älykkäiden menetelmien soveltaminen on nostanut esiin uusia ja paremmin suoriutuvia ennustemalleja, jotka eivät kuitenkaan yksistään ratkaise klassisten ennustemenetelmien ongelmia, sillä monet näistä ongelmista eivät johdu suoraan käytettävästä ennustemallista, vaan yleisistä toimintatavoista ja oletuksista ennusteprosessissa.

Tämä tutkielma analysoi 36 vertaisarvioitua, älykkään maksuvaikeuden ennustamiseen keskittyvää tutkimusta. Tulokset osoittivat, että aihealueella on monia tutkimuskohteita ja -tavoitteita, mikä viittaa monimutkaiseen ja moniuloitteiseen ilmiöön. Meta-analyysin mukaan moniluokittelijajärjestelmiä on käytetty eniten ennustemalleina ja ne ovat myös yltäneet keskimäärin parhaaseen ennustetarkkuuteen. Tarkastelluissa tutkimuksissa on käytetty useammin epätasapainoista kuin tasapainoista dataa, mutta vain pieni osa tutkimuksista on pyrkinyt ratkaisemaan tästä metodivalinnasta johtuvia ongelmia. Klassisten ennustemenetelmien ongelmiin ratkaisut ovat osin puutteellisia. Vaikka tutkimuksissa on käytetty dynaamisia elementtejä ja monimutkaisia menetelmiä mallintamisen vaiheissa, ei suurinta osaa ongelmista ole otettu aktiivisesti huomioon. Merkityksellisten muuttujien identifioimiseksi tarvitaan tulevaisuudessa laajempaa teoreettista analyysiä. Myös moniluokittelijajärjestelmiä vertaileva analyysi tarjoaa potentiaalisen jatkotutkimuskohteen.

Table of contents

List of appendices

List of figures

List of tables

List of abbreviations

ACA	Ant colony algorithm
ADASVM-TW	Adaboost SVM integrated time weighting
ANN	Artificial neural network
ANS-REA	Adaptive neighbor SMOTE-Recursive ensemble approach
BE-LWS	Batch-based ensemble with local weighted scheme
BGEV	Generalized extreme value model
BPNN	Back-propagation neural network
BSM-SAE	Borderline Synthetic Minority-Stacked AutoEncoder
CART	Classification and regression tree
CHAID	Chi-square automatic interaction detection
CNN	Convolutional neural network
DBN	Deep belief network
DEVE-AT	Double expert voting ensemble with Adaboost-SVM
DFDP	Dynamic financial distress prediction
DNN	Deep neural network
DT	Decision tree
EBW-VSTW-SVM	SVM with Entropy-based weighting and vertical sliding time window
ELM	Extreme learning machine
ES	Example selection
EW	Example weighing
FD	Financial distress
FDP	Financial distress prediction
FS	Feature selection
FSCGACA	Fitness-scaling chaotic genetic ant colony algorithm
GA	Genetic algorithm
GACA	Genetic ant colony algorithm
GAM	Generalized additive model
GAMSEL	Generalized additive model selection
GBDT	Gradient boosting decision tree
GBM	Gradient boosting machine
GLM	Generalized linear model
GRNN	Generalized regression neural network

GSKELM	Grid-search optimized KELM
GSPCA-SVM	Grouping sparse PCA-SVM
GWO	Grey wolf optimization
HACT	Hybrid associative memory with translation
IB	Incremental bagging
IF	Isolation forest
IST-RS	Incorporating sentiment and textual information into RS
KDA	Kernel LDA
KELM	Kernel extreme learning machine
KNN	K-Nearest Neighbors
KPCA	Kernel PCA
KRR	Kernel ridge regression
Lasso	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
LR	Logistic regression
LSAD	Least-Squares Anomaly Detection
LSTM	Long-short term memory
MDA	Multiple discriminant analysis
MARS	Multivariate adaptive regression spline
ML	Machine learning
MLP	Multi-Layer Perceptron
MWMOTE	Majority weighted minority oversampling technique
NN	Neural network
obRF	Oblique RF
OCSVM	One class SVM
OF-SVM	Original features-SVM
PCA	Principal component analysis
PLS	Partial least squares
PNN	Probabilistic neural network
PSOFKNN	Particle swarm optimization enhanced fuzzy KNN
PSOKELM	Particle swarm optimized KELM
QDA	Quadratic discriminant analysis

RACOG	Rapidly converging Gibbs sampling technique
RBFNN	Radial basis function neural network
RF	Random forest
RFE	Recursive feature elimination
RNN	Recurrent neural network
RS	Random subspace
RWO	Random walk oversampling approach
SAE	Stacked auto encoder
SaE-ELM	Self-adaptive evolutionary extreme learning machine
SBE	Selective bagging ensemble
SDFP	Static financial distress prediction
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority oversampling technique
SPCA	Sparse PCA
ST	Special treatment company
SVM	Support vector machine
WRACOG	Wrapper-based RACOG

1. Introduction

Financial distress (FD) has received a great deal of attention in academic literature, starting from the 1930s (Bellovary et al., 2007). Specifically, studies have focused on finding the main factors that contribute to deterioration of financial health of a firm. Indeed, predicting whether a firm faces financial distress in the future, can be highly beneficial for different stakeholders in various industries (e.g., credit risk assessment in banking, suppliers' default risk). Definitions of financial distress ranges from "early-warning" signals (i.e., the first indicators of financial deterioration) to corporate failure or a bankruptcy announcement. The definition is not standardized, and terms like, "corporate failure prediction", bankruptcy prediction", "business failure prediction", "solvency prediction", "high credit risk" and "default prediction", are often used in research papers to describe the same phenomenon.

Traditionally, financial distress prediction (FDP) is treated as a classification task, where two distinct groups of companies, non-distressed and distressed, are defined and the goal is to find a function (i.e., combination of features) that is the best at describing both groups. In general, status of companies (distressed or non-distressed) and input features (e.g., financial ratios) are collected and the classification function is found by exposing collected data to a certain statistical or intelligent method. Since the aim is to predict status of a company in the future, an output feature describes company's financial health at time *t* and input features at time *t*-1 or *t*-2 etc. Other modelling frameworks, like contingent claim models and survival analysis, have also contributed to FDP scheme. However, these are beyond the scope of this thesis, but an interested reader can find more details, for instance, in Shumway (2001), Bauer & Agarwal (2014), Hillegeist et al. (2004), and Agarwal & Taffler (2008).

The first classification models were based on traditional statistical methods, e.g., univariate analysis, discriminant analysis (DA), logit and probit analysis. In the 1980s, more sophisticated models emerged, for instance, recursive partition algorithm (RPA) and neural networks (NN), to introduce flexibility and higher prediction performance. These intelligent methods proved to be reasonably powerful in many comparative studies. Intelligent methods

established their role in FDP scheme and are now implemented in most of the studies (Veganzones & Severin, 2020). Generally, intelligent methods consist of non-parametric models that do not have restrictive assumptions of distribution of data and linearity. The most popular intelligent methods are, for instance, neural network models, machine learning models (e.g., support vector machine (SVM) and decision trees (DT)), evolutionary approaches (e.g., genetic algorithms (GA)). More and more, novel intelligent applications are built, which seems to be the current trend in the field. Specifically, many of the recent studies have focused on multiple classifier systems, or ensemble methods (Veganzones & Severin, 2020).

Although many successful financial distress prediction models have been developed, there are still open questions to be answered. Even though intelligent FDP models have yielded high accuracy rates, criticism around classic financial distress prediction methodologies is not disappeared. Balcaen & Ooghe (2006) described thoroughly the problems in current FDP methodologies. They categorised them into four dimensions: 1) the classic paradigm, 2) neglect of time dimension, 3) application focus, and 4) other problems. The classic paradigm consists of problems of "arbitrary" choice of output feature and performance metric, non-stationarity and data instability and non-random sampling. Neglect of time dimension criticises common practices of using cross-sectional data and treating company failure as uniform and steady phenomenon. Application focus demonstrates how, in general, a prediction model and features are both somewhat arbitrarily selected without proper theoretical background. "Other problems"-category outlines the problems of linearity and statistical models and extensive use of financial features. In addition, financial distress prediction has properties, like class imbalance and cost-sensitivity, which require extra attention in the modelling phase.

Obviously, the main problems in FDP implementations are not simply solved by introducing more sophisticated and flexible methods, but to seriously address the fundamental issues that restrict the framework. Indeed, the vast majority of intelligent FDP studies have one particular objective that tackles only one component of the complex FDP modelling framework, e.g., designing the best performing model, introducing a new hyperparameter tuning algorithm, and improving a feature selection process. Although there are overwhelming number of research papers around the subject, there is still room for analysing

the “bigger picture”. There exist many extensive literature reviews, for instance Ravi Kumar & Ravi (2007), Kirkos (2015) and Veganzones & Severin (2020), but very few of them are interested in analysing how the intelligent FDP studies truly address the main problems in classic prediction methodologies.

Therefore, the thesis concentrates on reviewing recent contributions in intelligent FDP domain, by introducing first current trends in the field and, secondly, assessing how the studies respond to the fundamental problems. Total of 36 peer-reviewed studies, published within the last six years, were collected and analysed. All the studies applied at least one intelligent method. In case a study presented some novel model, the primary method had to be an intelligent one. Introduction of the studies focuses on five perspectives: 1) main objectives, 2) data, 3) methods, 4) results, and 5) conclusions.

The studies showed that intelligent FDP is a complex, multidimensional subject with various research areas. Prediction process is still heavily dependent on the usage of financial features, since only one-third of the studies applied other feature types, like management factors and textual features. Imbalanced datasets were commonly used, although problems occurring due to this practice were not actively addressed. Ensemble and hybrid methods were the most popular and yielded the highest performance scores. However, there was no clear consensus of the superior model.

Analysis of the studies showed that the main problems in classic financial distress prediction methodologies remain still unsolved. In the future, studies of intelligent FDP should focus more on fundamental properties of financial distress, e.g., dynamic process, path-dependence, imbalance, cost-sensitivity. In-depth theoretical analysis is needed to find meaningful relationships between input-output pairs in the prediction task. Besides the studies answering to the problems related strictly to traditional statistical models, they also introduced dynamic FDP models and solutions to imbalance problem. However, further investigation is required to improve the quality in intelligent financial distress prediction. Indeed, research community is encouraged to cooperate and collectively challenge common practices in the field. Only 36 studies were analysed, thereby covering only a small proportion of the whole set of published papers. This is, evidently, a limitation of the thesis and more extensive research with different set of studies should be examined.

The thesis is structured as follows: Section 2 describes the main characteristics of financial distress prediction. Firstly, financial distress is defined followed by a description of prediction scheme, or more precisely, classification scheme. Next, the history of FDP is introduced to highlight the most significant work around the subject. Then, two main prediction methods are presented, namely, statistical and intelligent methods. Particularly, the subsection concentrates more on intelligent methods, where machine learning approach is also described in detail, since it is a major part of the intelligent techniques. Finally, the section ends with introducing common steps in the intelligent financial distress prediction process. Section 3 focuses on the problems related to classic financial distress prediction methodologies. The section follows closely Balcaen & Ooghe (2006), where each problem category is described extensively. In Section 4, a collection of intelligent financial distress prediction studies is presented, in terms of main objectives, datasets, methods, results, and conclusions, to see the current trends and practices around the subject. Then, the studies are analysed through each problem category to find out if and how problems are answered in the intelligent FDP domain. In addition, further suggestions of the solutions are given which hopefully offers something to grasp on in the future. The final section is left for concluding remarks where main findings and limitations of the thesis are highlighted, and recommendations for the future research are given.

2. Financial distress prediction

In this section, descriptions of financial distress and financial distress prediction are presented. Firstly, financial distress and its main characteristics are defined. Next, the history of FDP and commonly used methods are described in detail. A brief review of traditional statistical methods is given, but the main focus is on intelligent methods and their properties. Finally, common practices in intelligent financial distress prediction are presented. As noted in the previous section, only classification prediction scheme is considered here to assure in-depth description of the subject.

Financial distress is a general term of a status of a company having difficulties to meet its financial obligations due to insufficient cash flows. It is described as a process of deterioration of a company's financial capabilities. FD is dynamic, ongoing, and evolutionary in nature, where the critical point of a company's status transforming from healthy to distressed is ambiguous. (Nwogugu, 2007)

There is no single, universally agreed metric to describe a financially distressed company, which has led to various definitions in academic literature. Definitions are derived from all types of financial deterioration signs, ranging from less severe "early warning signals" (e.g., temporary cash flow difficulties) to final corporate failure or bankruptcy. In practice, simple criteria, like bankruptcy or regulatory announcements, are preferred in research papers to clearly make a distinction of a financially distressed and a non-distressed company. (Sun et al., 2014) One widely used regulatory announcement is "special treatment" status of listed companies on Shanghai Stock Exchange (SHSE) and Shenzhen Stock Exchange (SZSE), which obviously can be applied only companies that are listed on the respective stock exchanges (Zhou, 2013).

Financial distress is described as path-dependent and time-dependent process (Nwogugu, 2007). A company has its unique route from the founding to present, with its financial health constantly changing through time. However, studies have recognized certain failure trajectories which can be used to identify "early warning" signals (Flores-Jimeno & Jimeno-Garcia, 2017; Ooghe & De Prijcker, 2008). Indeed, extensive literature exists around

financial distress modelling, which would not be possible without any generalization capabilities.

In general, financial distress prediction studies assume a dichotomous view of the world, i.e., only two distinct groups of non-distressed and distressed firms exist where various levels of financial distress are ignored (Tsai, 2013). However, some studies have suggested to use more than one criterion to describe financial distress (Sun et al., 2014). For instance, Farooq et al. (2018) described financial distress in a three-stage dynamic model. They argued that financially healthy firms, first, experience profitability problems which then leads to more severe liquidity issues and ultimately bankruptcy. Therefore, their model described financial distress as a three-stage process of: 1) profit reduction or mild liquidity problem, 2) severe liquidity problem, and 3) a legal bankruptcy. At any stage a firm could recover from distress, but it gets more difficult the more severe the distress is. Many recovery strategies in different stages are suggested, e.g., retrenchment and efficiency-approvement procedures (Arogyaswamy et al., 1995). The literature around recovery strategies (more closely, the term "turnaround" is used consistently) is extensive and purposely left out of the scope of this thesis. More details can be found, for instance, Schweizer & Nienhaus (2017) and Schoenberg et al. (2013). Similarly, Tsai (2013) categorized financial distress into different levels based on severity: "slight financial distress", reorganization, and bankruptcy.

According to Laitinen (2005), financial distress process can be separated to different stages and their respective financial indicators. Each stage consists of typical factors (primary covariates) that have an influence on the FDP process. Also, some additional factors were described (secondary covariates), for instance, size, industry and age of a firm, which may affect the process. Stages and financial indicators of financial distress process is depicted in Figure 1:

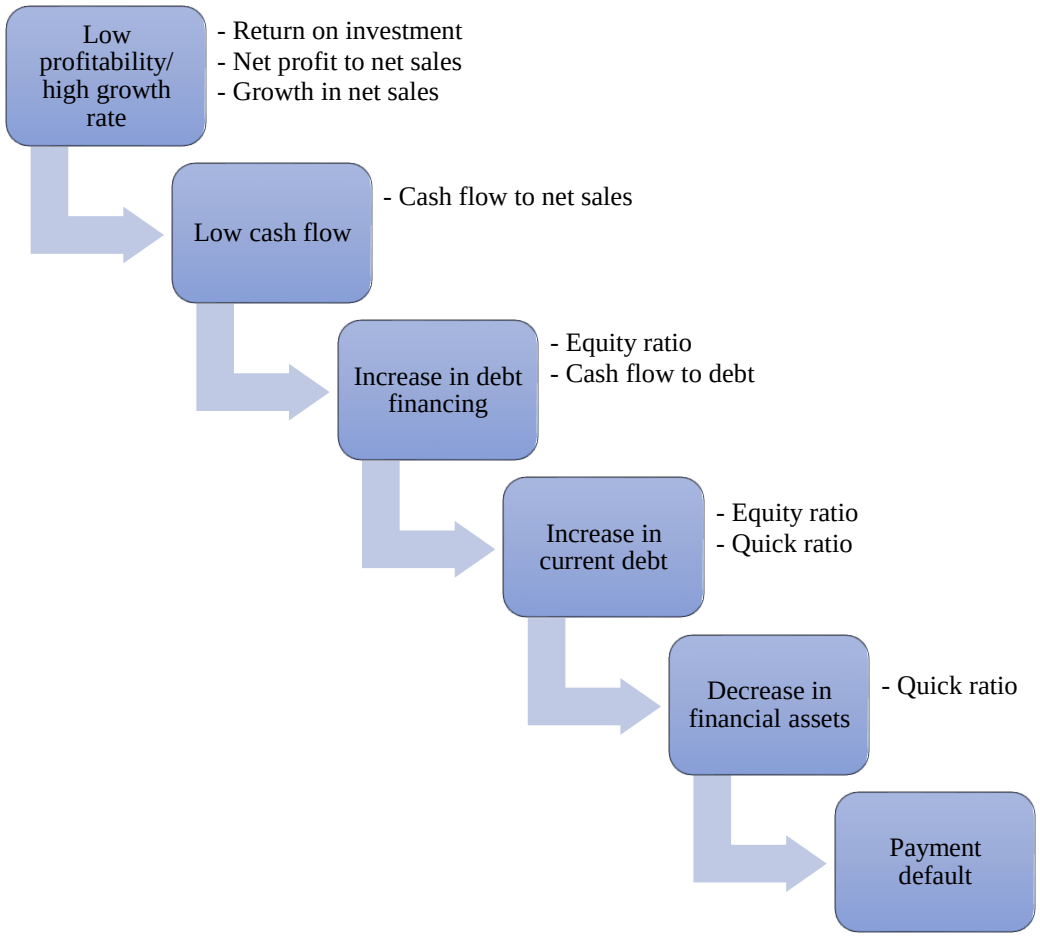

Figure 1. Financial distress process (Adapted from Laitinen, 2005)

Often, the very first signal of financial distress is high deviation between profitability and growth rate, that associates either with overly ambitious growth strategy or diminishing profitability or some combination of the two. Financial ratios such as return on investment or net profit to net sales and growth in net sales are commonly used indicators for profitability rate and growth rate, respectively. Diminished returns and unstable growth lead to poor operating cash flows and a company is forced to rely on debt financing to pay its obligations. The level of debt will keep rising, and in case cash flows remain low, the firm needs to use financial assets or worse to sell current assets to cover its debts. Consequently, the firm defaults on its payments. Notably, indicators depicted for each stage may stay effective through the whole process, or they can deteriorate as the process evolves. (Laitinen, 2005) This means that, for instance, profitability and growth ratios may not be suitable features to use in FDP if collected distressed companies are in last stages of the process.

Evidently, the ultimate stage of financial distress is bankruptcy. The Ministry of Justice Finland (2020) defines it as "a procedure where the assets of the debtor are used all at once in order to cover his or her debts, in proportion to the amounts of the individual debts". It is a legal process, imposed by a court order. The process varies, depending on under which jurisdiction bankruptcy is filed. The Ministry of Justice Finland (2020) described common steps in bankruptcy process in Finland: 1) filing bankruptcy application, 2) appointing an administrator by the district court to manage the bankruptcy application, 3) taking over debt and estate of the debtor, 4) composing an estate inventory and written account of the debtor's business prior the bankruptcy and the causes for the bankruptcy, 5) estimating sufficiency of the assets to cover the debt owned to creditors, 6) setting a date by which creditors have to file their claims if the debtor has sufficient assets to cover all financial obligations. If there are no sufficient assets, the district court can lapse the bankruptcy process, and the remaining assets are surrendered to the enforcement authority. However, the debtor is not released from liability, and it is also obliged to pay debts with assets received after the bankruptcy. In most cases, however, company ceases to exist.

Ooghe & De Prijcker (2008) listed five factors that have the greatest impact on probability of a bankruptcy: 1) immediate environment, 2) general environment, 3) management, 4) corporate policy, and 5) company's characteristics. Immediate environment describes the interactions between a company and its stakeholders, e.g., customers, suppliers, and competitors. General environment consists of external factors, such as economic and technological changes, and political and social factors. The third group relates to characteristics of corporate management, their motivation, qualities, and skills. The fourth factor, corporate policy, includes factors of, for instance, corporate governance, strategic and operational decisions. The final group consists of characteristics of a company, such as age, size, and industry. Particularly, younger, and smaller companies tend to have a higher probability of failure (Kücher et al., 2020). Also, relationship between industry and corporate failure has proved to be significant (Platt & Platt, 1990).

It is highly beneficial for corporate management to understand the fundamental reasons for financial distress and being able to predict future financial health of a company. Consider, for example, a bank that measures credit worthiness of a company for lending purposes or a manufacturing company assessing potential suppliers and their credit risk. Both situations

demand an estimation of how likely it is a debtor to default in a given timeframe and what are the consequences resulting from it. Most companies would find significant value in estimating more accurately their debtors' financial condition. Platt & Platt (2002) highlighted the benefits of financial distress prediction: "In early warning system model that anticipates financial distress of supplier firms provides management of purchasing companies with a powerful tool to help identify and, it is hoped, rectify problems before they reach a crisis."

The fundamental idea in financial distress prediction is to build a classification model that estimates companies' financial health in the future, given the available current information. The outcome of the process is a model that is capable of predicting a new instance accurately (either distressed or non-distressed) as farther in the future as possible. Mostly, the prediction task is conducted in the binary classification context, i.e., prediction output is in a binary form (distressed vs. non-distressed, bankrupt vs. non-bankrupt, healthy vs. unhealthy etc.). As detailed in Chen et al. (2016), financial distress prediction problem can be expressed as: "given a number of companies labelled as bankrupt/healthy, and a set of financial variables that describe the situation of a company over a given period, predict that the company become bankrupt during the following years." More closely, the definition concerns bankruptcy prediction problem, but the same idea holds if instead of a bankruptcy, a prediction output is distressed vs. non-distressed. In addition, input variables are not restricted to only financial features. The prediction model can include non-financial features, such as corporate management metrics, or market features that describe current business cycle. However, financial features are overwhelmingly the most common feature type.

2.1 History

Financial distress prediction is an intensively studied subject. The first attempts to understand corporate failure prediction can be traced back to the 1930s. The Bureau of Business Research, in 1930, published a study which explored univariate model of individual financial ratios of failing industrial firms. The study found total of eight ratios (e.g., sales to total assets, cash to total assets and working capital to total assets) that are potentially good indicators to describe a failing firm. Smith & Winakor (1935) did a follow-up study where they analyzed ratios of nearly two hundred failed firms. The results indicated that working

capital to total assets was significantly better for prediction purposes than cash to total assets or current ratio. Also, a study of "A comparison of ratios of successful industrial enterprises with those of failed companies." by FitzPatrick in 1932, found two significant ratios, that are net worth to debt and net profits to net worth. The study comprised total of 13 ratios of 19 failed and successful firms. In the period between 1940 and 1966, the use of univariate analysis continued to grow, and significant results were found, for example, "Financing small corporations in five manufacturing industries, 1926-1936." by Merwin in 1942 and "A Study of Published Industry Financial and Operating Ratios." by Jackendoff in 1962. (Bellovary et al., 2007)

Beaver (1966) is the first who shifted from the simple univariate comparison into statistical discriminant analysis. In his study, 79 failed and 79 non-failed firms were compared in terms of average values of 30 financial ratios. The study tested individual ratios' prediction performance in a classification task (bankrupt vs. non-bankrupt firm). The results indicated that net income to total debt has the highest explanation power (92% accuracy in one-year prediction horizon).

The groundbreaking study of Altman (1968) conducted the first multivariate discriminant analysis (MDA) to predict bankruptcy in manufacturing industry. In his study, the famous five-factor "Z-score" model was built, and the results were promising for one-year prior bankruptcy predictions (95% accuracy). However, the prediction performance decreased significantly when prediction horizon increased (72%, 48%, and 29% accuracy for two-year, three-year, four-year prior to bankruptcy, respectively). Initial five-factors in the prediction model were: 1) working capital to total assets, 2) retained earnings to total assets, 3) EBIT to total assets, 4) market value equity to book value of total debt, and 5) sales to total assets. After Altman's study, the number of research papers and new prediction models have increased significantly. According to Bellovary et al. (2007), the number of studies climbed from 28 in the 1970s to 70 in the 1990s.

In the 1980s, second-generation models, or binary response models, were developed, which estimate probability of corporate failure using logistic or probit function (Kim et al., 2020). The study of Ohlson (1980) is a cornerstone paper in implementing logistic function in

bankruptcy prediction. One of the first studies that applied probit estimation was done by Zmijewski (1984).

Discriminant analysis and its various forms (univariate, multivariate, linear, quadratic etc.) and probit and logit models are traditional statistical techniques that are generally accepted, standard methods to develop corporate failure prediction models. However, these methods have many disadvantages, which is why in the late 1980s and in the early 1990s, more advanced methods with non-parametric characteristics were introduced to the subject. Frydman et al. (1985) is one of the first studies to present intelligent methods for bankruptcy prediction. They introduced a recursive partitioning algorithm, which is non-parametric, classification tree algorithm, and compared its prediction performance to DA models. Messier & Hansen (1988) presented an attributable algorithm, inductive dichotomizer, and compared it to DA models, individual judgements, and group judgements with two different datasets.

Neural network models were introduced in the 1990s to financial distress framework, and they became the most popular method (Bellovary et al., 2007). Studies, such as Koster et al. (1991), Tam (1991), Salchenberger et al. (1992), Fletcher & Goss (1993), and Yang et al. (1999), are good examples of utilizing these models. Around the same time, studies began to implement dynamic properties into the prediction task by introducing hazard models. To name a few, Lane et al. (1986) introduced Cox proportional hazards model to predict bank failures, Lee & Urrutia (1996) compared logit and hazard models to predict insolvency, and Shumway (2001) compared DA and simple hazard model for bankruptcy prediction.

Veganzones & Severin (2020) reviewed corporate failure studies in the 21st century. They assessed total of 106 papers published between years 2000 and 2017. According to their study, single statistical methods, especially discriminant analysis and logistic regression, are still widely used. Artificial intelligence methods are utilized even more, particularly neural network models, case-based reasoning, decision trees, and support vector machines. After the year 2007, ensemble methods became the most prominent approach to predict corporate failure.

2.2 Statistical methods

Two broad categories of methods in FDP are traditional statistical methods and intelligent methods (Ravi Kumar & Ravi, 2007; Chen, 2011). The traditional statistical methods are fully parametric, i.e., a model's structure is specified a priori, and all parameters are determined in finite dimensional parameter space (Chen et al., 2016). In addition, they are inherently limited due to their strict assumptions like linearity, normality, and independence of predictor variables (Hua et al., 2007). However, statistical models provide, mostly, clear interpretation of the model and they are still widely used in the corporate failure context (Chen et al., 2016). The traditional statistical models, that are used in financial distress prediction, comprise univariate analysis, risk index model, discriminant analysis, logit and probit analysis (Balcaen & Ooghe, 2006). Particularly, discriminant analysis and logit models are most popular methods and are briefly detailed in below (Balcaen & Ooghe, 2006).

Discriminant analysis is the first-generation method in corporate failure prediction, where linear combination of predictor features that separate two or more output classes is selected (Kim et al., 2020). The most popular DA method, multiple discriminant analysis (Balcaen & Ooghe, 2006), is based on a discriminant function of the following form, according to Dimitras et al. (1996):

$$D_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + \cdots + \beta_m * X_{im} \quad (1)$$

where D_i is the discriminant score for firm *i*, X_{im} is the value of attribute X_m for firm *i*, β_0 is intercept term, and β_{im} is the linear discriminant coefficient of firm *i* of attribute *m*. The objective of the method is to provide the linear combination of predictor variables in a way that the variance between the populations relative to within class variances is maximized (Dimitras et al., 1996).

After the discriminant scores are calculated for each sample, an optimal cut-off score is determined. Given that a company's discriminant score is lower (higher) than the optimal cut-off score, the company is classified as distressed (non-distressed). Discriminant analysis allows to rank companies, i.e., in most studies, the lower the discriminant score the poorer the financial health of a company. (Balcaen & Ooghe, 2006)

MDA holds the following assumptions: 1) independent, multivariate normally distributed predictor variables, 2) variance-covariance structure for each class is equivalent, 3) specified prior probabilities of failure and misclassification costs, 4) absence of multicollinearity, and 5) discrete and identifiable groups. (Charitou et al., 2004; Dimitras et al., 1996; Balcaen & Ooghe, 2006; Eisenbeis, 1977)

The first two assumptions barely ever hold, especially if predictor variables only consist of financial ratios (Richardson & Davidson, 1983; Mcleay & Omar, 2000). Studies have suggested various techniques, such as transforming predictor variables and excluding outliers, to solve the issue of multivariate normal distribution (Balcaen & Ooghe, 2006). However, transformation can only guarantee univariate normality, which is not a sufficient condition to multivariate normality (Balcaen & Ooghe, 2006). In addition, transformation may lead to distorted interrelationships among the predictor variables (Eisenbeis, 1977). Outlier deletion should also be considered with a great care since there is a possibility of losing key information (Ezzamel & Mar-Molinero, 1990). Alternative models, like quadratic discriminant analysis (QDA), are suggested to mitigate the problem of dispersed variance-covariance structure (Eisenbeis, 1977). Violation of either one of the assumptions leads to biased significance tests (Balcaen & Ooghe, 2006).

In practice, MDA implicitly assumes equal misclassification cost for each group and similar class proportionality between sample and population, leading to wrongly specified prior probabilities of failure and misclassification costs. This will generate misleading accuracy rates, since corporate failure is infrequently occurring event, much rarer than a healthy company. Similar problems may occur if the assumption of multicollinearity is violated. (Balcaen & Ooghe, 2006) However, multicollinearity is mostly considered irrelevant in MDA models. MDA also assumes that groups are properly defined, i.e., identifiable and discrete. (Eisenbeis, 1977) Arbitrarily defined groups, like forming groups from a discretized continuous variable, have several troubling properties that results inappropriate use of the method and misleading outcomes (Eisenbeis, 1977; Balcaen & Ooghe, 2006).

Both probit and logit models are conditional probability models that utilize cumulative probability function to derive likelihood of a sample belonging to a certain class (e.g., financial distress vs. non-distress), given the sample's characteristics (i.e., predictor

variables). Classification is conducted by choosing a cut-off point for probability measure that separates, in a binary case, the two classes. For instance, given a cut-off point of 0.5, a sample is classified as financially distressed if its probability of belonging to class "financial distress" is higher than 0.5. The optimal cut-off point should be based on minimization of Type I and Type II errors. (Dimitras et al., 1996)

Probit and logit analysis have similar structures, except that logit assumes cumulative logistic function and probit cumulative standard normal distribution. Logit model is far more common in corporate failure prediction, hence only details of logit is included here. (Balcaen & Ooghe, 2006)

Following Foreman (2003) and Charitou et al. (2004), logit model is depicted by:

$$P(y_i = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \cdots + \beta_{mi}X_{mi})}} \quad (2)$$

where $P(y_i = 1)$ is the probability of failure of company i, $\beta_1, \beta_2, \ldots, \beta_m$ are coefficients for predictor variables of $X_1, X_2, \ldots, X_m$. Coefficients can be estimated by maximizing log-likelihood, where the likelihood function is given by:

$$L = \prod_{i=1}^{N} F(\beta' X_i)^{y_i} * \left(1 - F(\beta' X_i)\right)^{1-y_i} \quad (3)$$

where Π (*) is product symbol over a set of all companies i=1,…,N, and

$$F(\beta' X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \cdots + \beta_{mi}X_{mi})}}$$

The main advantage in logit models is that the method does not make any assumptions about the prior probability of failure or the distribution of input features, thereby being less demanding compared to discriminant analysis (Ohlson, 1980). However, logit models are sensitive to multicollinearity which is a concern in the corporate failure context (Doumpos & Zopoundis, 1999). Indeed, FDP models extensively utilize financial ratios, which tend to be highly correlated. In addition, logit model assumes, like DA method, discrete and identifiable groups and specified misclassification costs. (Balcaen & Ooghe, 2006)

2.3 Intelligent methods

Intelligent methods were introduced to financial distress prediction already in the late 1980s, in the form of non-parametric classification trees, to remedy the problems of statistical methods. Indeed, Frydman et al. (1985) introduced recursive partitioning algorithm, a nonparametric classification technique based on pattern recognition. They summarized the main benefits of RPA as follows: 1) superior accuracy results compared to DA, and 2) nonparametric properties, i.e., free of restrictive assumptions of DA. Messier & Hansen (1988) presented a similar method, concept learning algorithm (inductive dichotomizer), to loan default and bankruptcy data. The algorithm is a data-driven model that can produce IF-THEN rules based on inductive learning. The model outperformed other benchmark models, such as discriminant analysis, and the authors assessed it to be promising alternative method for developing expert systems.

In the 1990s, the number of corporate failure studies with neural network models started to grow. Neural networks are inspired by the neural architecture of the brain, where input-hidden layer(s)-output structure learns meaningful relationships from the data (Salchenberger et al., 1992). Some studies have found them to produce outperforming accuracy results (Tam, 1991; Salchenberger et al., 1992). However, Yang et al. (1999) showed mixed results, based on which commonly used back-propagation NN models were inferior to discriminant analysis. Tam (1991) outlined the main advantages of NN models: 1) robustness and nonparametric properties, 2) continuous scoring system, 3) allowing incremental adjustment when new data is fed to the system, and 4) possibility of reducing adverse effects of within-group clusters on prediction accuracy. In the same study, main disadvantages were also highlighted: 1) difficulty in model interpretation, 2) lack of tools for dimension reduction, 3) computational burden, and 4) necessity of choosing certain topology for the model, i.e., configuration.

Since then, the number of various intelligent techniques has grown intensively, some of them being used more frequently in financial distress prediction. Ravi Kumar & Ravi (2007) conducted an extensive review on statistical and intelligent techniques applied in bankruptcy prediction. They studied the most popular categories of intelligent methods: 1) neural network models, 2) decision trees, 3) case-based reasoning, 4) evolutionary approaches, 5)

rough sets, 6) soft computing techniques (hybrid and ensemble methods), 7) operational research techniques, and 8) other techniques (support vector machine and fuzzy logic). Various architectures of neural network models were included, such as multi-layer perception, self-organizing map and learning vector quantization. They showed that statistical methods are no longer the most popular method for bankruptcy prediction, and that intelligent methods have replaced them due to their higher prediction performance. Neural network models were the most popular category, followed by rough sets, CBR, operational research techniques, evolutionary approaches and other techniques. Also, the study highlighted the rising trend of hybrid and ensemble systems, which have shown to outperform individual intelligent techniques.

Kirkos (2015) agrees on the importance of hybrid and ensemble systems, emphasizing that novel approaches for bankruptcy prediction are often conducted by using a composite system. Similarly, Veganzones & Severin (2020) realized the trend of ensemble learning. Before 2007, 31% of the studies used statistical methods, 56% artificial intelligence methods, and only 13% used ensemble methods. After 2007, statistical methods have dropped to 13%, artificial intelligence to 36%, but ensemble methods have grown to 51%. However, in review by Duarte & Barboza (2020), traditional statistical methods (logistic regression, linear discriminant analysis, and multiple discriminant analysis) were found to be still relevant techniques, especially logistic regression which was the most popular followed by support vector machines, artificial neural network, and decision tree.

2.3.1 Machine learning approach

In general, intelligent methods consist of techniques that are non-parametric, i.e., no assumptions of data distributions and no fixed set of parameters are defined a priori (Chen et al., 2016). Also, they extract information strictly from the data (Chen et al., 2016). This machine learning (ML) approach is a subset of artificial intelligence which has become popular modelling technique due to the rise of big data and increase of computation power and efficiency of machines (Mehta et al., 2019). The term “machine learning” was invented in 1959 by Arthur Samuel in his study of utilizing machine learning in the game of checkers (Samuel, 1959). Mitchell (1997) provided a formal definition of machine learning in his book ‘Machine Learning’: “A computer program is said to learn from experience E with

respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.". In the context of financial distress prediction, T represents the task of predicting a company distressed vs. non-distressed, P certain performance metric (accuracy, Type I error, Type II error etc.), and E data samples. The prediction task is categorized into supervised learning, more closely, classification learning since the output variable is in a binary form.

According to Mehta et al. (2019) machine learning approach in supervised learning requires three fundamental components: 1) a dataset, 2) a mapping function, and 3) a cost function. A dataset, D, consists of input-output pairs (x, y) where x is a vector of predictor variables and y represents an output variable. If y can take only discrete values, a prediction task is called classification learning. In case y is continuous, it is called a regression task. A mapping function, f, is a model $f(x; \theta)$ that produces output value, given input value and parameters of the model: $f : x \rightarrow y$. A cost function, $c(y, f(x; \theta))$, is a function to determine how well the mapping function performs. The mapping function is chosen by finding parameter values of θ such that the cost function is minimized.

A typical cost function is squared error, and method for minimizing it is called least square method. It involves searching for optimal combination of weights in a way that loss function of squared errors is minimized:

$$\hat{\beta} = \arg\min \|y - X\beta\| \tag{4}$$

where $\hat{\beta}$ is the optimal vector of β weights, y is a vector of true output values, $X\beta$ is multiplication of input data and β weights (i.e., prediction of output values). The operator, $\|*\|$, indicates Euclidean norm. Optimal solution in least squares method is found analytically or using first-order optimization method like gradient descent. Once the weights are determined, the mapping function can be evaluated in some performance scheme. (Mehta et al., 2019)

Supervised learning system follows statistical learning theory which contains some basic assumptions. Firstly, there exists an unknown target function, $y = f(x)$, that describes perfectly observed input-output pairs (x, y). The target function is hypothetical which cannot be observed directly. Secondly, a hypothesis set H consists of all functions that are

considered to represent the target function. This is a choice made by a system designer, for instance, H may include only linear combinations of input variables. Since the target function is unknown, the choice of hypothesis space H introduces bias to the system. The goal is to find function h from hypothesis set $h \in H$ that is as close as possible to the target function $y = f(x)$. Indeed, function h approximates the target function. To get an intuition, if chosen hypothesis approximates the target function, it should be performing well also against unseen data, given that the unseen data is drawn from the same input-output distribution. Therefore, in a standard machine learning approach, chosen hypothesis, h, is evaluated against unseen samples. (Mitchell, 1997)

A typical ML procedure is, first, randomly partition dataset D into a training set and a test set. A training set is used to fit the model, i.e., to find the mapping function. The final evaluation of the model is done against a test set where the input data from the test set is fed into the fitted model and its prediction output is compared against true output of the test set by using a certain cost function. This evaluation procedure gives some intuition of the model's generalization capabilities. Value that a cost function generates from a training set is called in-sample error, E_i, and value from a test set is called out-of-sample error, E_o. The objective of supervised learning is to minimize the out-of-sample error. (Mehta et al., 2019)

In general, out-of-sample error can be broken down to two components, bias, and variance terms. Bias describes the assumption of hypothesis space that is made by a system designer. Increasing model complexity will, in general, decrease bias term. The more hypotheses are in the hypothesis space the more likely there exists one which is close to target function. However, as model complexity increases, so does the variance error term. Higher model complexity requires more data points, and in some situations the training set is too small for a complex model being generalized in the test set. There is a balance between bias and variance term, a concept known as bias-variance tradeoff. A simple (complex) model introduces higher (lower) bias and lower (higher) variance, and the optimum level, i.e., minimum of out-of-sample error, is found somewhere between. (Mehta et al., 2019)

Random partition of dataset D and evaluating the model against a test set is a model validation technique called cross-validation. Indeed, it is a technique used for assessing how well the model generalizes to independent datasets (Arlot & Celisse, 2010). Various cross-

validation methods have been developed, of which holdout method is the simplest one where random partition and evaluation is done only once (Arlot & Celisse, 2010). This single-run technique provides only one performance estimation and may be misleading (Nakatsu, 2021). More common multiple-run methods, like v-fold and leave-one-out methods, are used in practice to produce multiple performance measurements which can provide a better estimation of the predictive accuracy (Nakatsu, 2021). To give some intuition, for instance, v-fold method splits a dataset into *v* equal subsamples and for each subsample, *v*-1 subsamples are used as a training data to fit a model, and v^{th} subsample is treated as a test set to measure its performance. Then, *v* performance results are averaged to get one estimation of the model performance. (Arlot & Celisse, 2010)

For other than model validation purposes, cross validation is commonly used for hyperparameter tuning, a critical step in every machine learning system (Duarte & Wainer, 2017). Hyperparameter tuning refers to procedure in which algorithm's optimal hyperparameters are defined. Hyperparameters are parameters that control the configuration of a model before actual parameters are derived. (Yang et al., 2017) Each algorithm has a certain set of hyperparameters that can be tuned, for instance, number of neighbors in KNN algorithm (Antal-Vaida, 2021).

2.3.2 Structures of intelligent systems

Intelligent learning systems are categorized into single, hybrid, and multiple classifier systems. In a single system, only one individual classifier technique is used. In a hybrid system, two or more heterogeneous techniques are utilized to produce one classification output. A typical procedure is to first use one technique for data preprocessing (i.e., feature selection or dimension reduction) and then apply actual classifier to the preprocessed dataset. (Chen et al., 2016)

Multiple classifier systems, or ensemble methods, utilize a combination of many classifiers (heterogeneous or homogeneous), to produce the final output (Opitz & Maclin, 1999). The idea is to fit *n* weak classifiers and then use certain combination technique, for instance, voting method or averaging, to aggregate *n* outputs into one (Opitz & Maclin, 1999). In designing of ensemble system, various strategies are implemented, which are generally

divided into four categories: 1) method diversity, 2) fusion diversity, 3) topology diversity, and 4) output diversity (Chen et al., 2016). The most prominent ensemble methods are bootstrap aggregating, boosting, and stacking (Hall et al., 2011).

In bootstrap aggregating, or bagging, *n* datasets are generated by randomly drawing with replacement *n* times from the initial dataset. For each random subsample *n*, a classifier is trained which yields a total of *n* different classifiers. A voting method (in a classification task) is then used to aggregate *n* outputs into one. Bagging belongs to method diversity strategy where variance of training data used to build multiple classifiers. (Breiman, 1996; Chen et al., 2016)

Boosting focuses on producing series of classifiers. Multiple classifiers are constructed sequentially, in a way that each classifier tries to compensate for the weaknesses of its predecessor. The idea is to combine a set of weak learners to build a strong learner. In boosting, random training set is drawn (with replacement) from the original dataset and the first classifier is built. For the next iteration, random draw is not uniform, but the samples that the first classifier misclassified are given more weight so that the second classifier would possibly achieve higher performance in these samples. The iteration process continues until a certain stopping criterion is achieved. Boosting has many variations, for instance AdaBoost, XGBoost, and GradientBoost, but the main idea of sequentially proceeding and allocating more weight to misclassified examples remains the same in all boosting methods. (Opitz & Maclin, 1999)

Stacking, or stacked generalization, is another ensemble method to combine multiple classifiers. Stacking generally uses heterogeneous models, where the main goal is to build a system with 0-1 level structure. The first level, level zero, consists of multiple base classifiers that each produce an output to the learning problem. In the last level, level one, a single classifier is used to derive the final, single output, based on the predictions of the base classifiers. The structure can be extended to a multi-level system, where extra base-classifier-levels are added. (Wolpert, 1992; Czarnowski & Jedrzejowicz, 2017)

2.3.3 Prediction process with intelligent methods

In general, intelligent FDP process includes the following steps that are described in Figure 2:

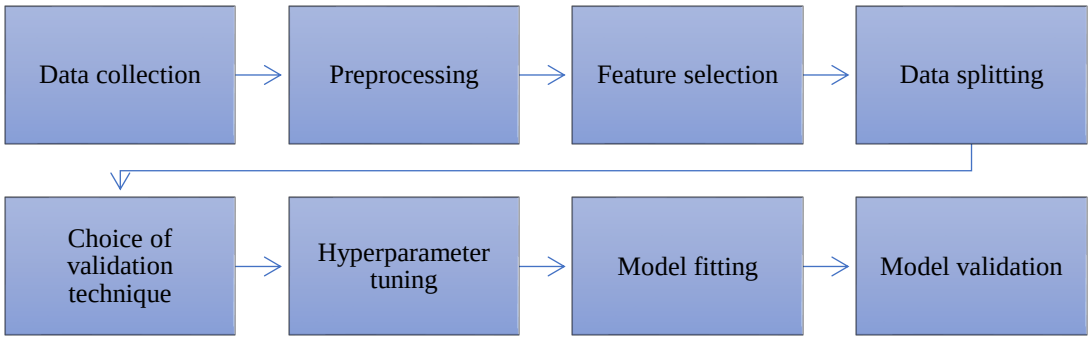

Figure 2. Intelligent FDP process (Adapted from Chen et al., 2016)

Figure 2 is only one example of the process and may vary across different implementations. For instance, feature selection method can be embedded to the prediction algorithm in model fitting phase. Note, the prediction process differs when a traditional statistical method is applied. As an example, learning phase with traditional statistical methods does not usually involve data splitting and hyperparameter tuning.

The process starts by collecting sample of companies and their respective financial distress statuses. There are two different ways to construct dataset, namely, imbalanced and balanced approach (Sun et al., 2014). In balanced approach, even or close to even number of distressed and non-distressed companies are collected to form a dataset. In the real world, however, financial distress or corporate failure is a rare event and artificially balancing class distribution may introduce choice-based sample bias creating misleading prediction accuracy and biased parameter estimates (Platt & Platt, 2002). The argument for using balanced datasets is to get higher representation of minority samples which may enhance prediction accuracy (Leevy et al., 2018). Also, commonly used paired-sampling technique, i.e., sampling that is based on collecting first distressed firms and then matching each firm with a non-distressed firm of similar characteristic, results balanced datasets (Balcaen & Ooghe, 2006). Paired-sampling is a popular technique since it enables to control for some

variables that are assumed to have at least some impact on prediction performance but are excluded in the set of predictor variables (Keasey & Watson, 1991).

Oversampling failing firms avoids certain difficulties that imbalanced datasets impose. Primarily, if the class distribution is severely skewed, prediction models tend to undermine minority samples, yielding high misclassification rate in the minority class (Leevy et al., 2018). The problem is even worse in FDP, which is particularly interested in predicting accurately firms that are financially distressed in the future (Chen et al., 2016). This characteristic of cost-sensitivity (i.e., non-uniform cost of misclassification between classes) is a common problem in classification learning where samples of minority class are more important to identify (Krawczyk, 2016). In addition, addressing the imbalance problem introduces complexity to the prediction system since mitigation of the problem requires additional procedures.

Various methods, both data and algorithm approaches, can be used to alleviate the imbalance problem (Sun et al., 2009). The main approach, at the data level, is to use some resample technique which rebalances the class distribution. Many resample techniques have been developed, including randomly over- and undersampling, informative resampling, synthesizing, and some combination of them. (Elreedy & Atiya, 2019) At the algorithm level, an existing classifier is modified in a way that it focuses more on the minority class (Krawczyk, 2016). The modifications are usually algorithm-specific, for instance, decision trees with novel pruning techniques and support vector machine with different penalty costs for each class (Sun et al., 2009). In addition, cost-sensitive learning and ensemble methods, especially boosting, are common practices with imbalanced datasets (Sun et al., 2009). Cost-sensitive learning incorporates both data and algorithm approaches, where algorithm assumes higher misclassification cost in the minority class and optimize the total misclassification cost (Krawczyk, 2016). As demonstrated, there exist many solutions for the imbalance problem, but it is not obvious to choose the most useful one. All the approaches contain specific advantages and disadvantages, and not a single approach is proved to be superior one (Kaur et al., 2019).

After defining financial status for each instance in a dataset, predictor variables are determined and gathered. Typically, predictor variables consist of financial features derived

from annual reports and other financial reports that are publicly available (Veganzones & Severin, 2020). Financial features are common in practice because they are relatively easy to collect and studies have shown them to have at least some explanation power (Veganzones & Severin, 2020). However, using only financial features has several problems which are highlighted in the next subsection. Indeed, many studies emphasise to explore alternative predictor variables, such as economic variables, operational, sentiment, textual, and management variables (Korol, 2019; Mselmi et al., 2017; Figini et al., 2017; Tang et al., 2020). The final collection of predictor variables and financial statuses represents input-output feature pairs of firms in the dataset.

Data, in its raw format, is almost never usable for a prediction task and may require several actions before implementation is possible. Thus, data preprocessing is a crucial step in the process. Depending on quality of data and classification method, a preprocessing phase may include: 1) discretization and normalization, 2) feature selection, 3) noise reduction, 4) outlier handling, 5) instance selection, and 6) missing value imputation (Alexandropoulos et al., 2019).

Discretization of a variable, i.e., a continuous variable transformed to discrete space, may be useful in certain learning algorithms, like decision trees, and it can save computation time while maintaining or even improving prediction accuracy (Tsai & Chen, 2019). Feature normalization should be considered when a dataset consists of variables with significantly different scales since this leads to the dominance of some variables over the others (Xie et al., 2016).

Many learning algorithms are sensitive to noise and outliers, hence various preprocessing techniques, e.g., noise filters, statistics-based methods, distance-based methods, are developed to detect these instances (Alexandropoulos et al., 2019). However, there is no standardized way of handling them, especially, in the case of outliers where a simple deletion technique may accidentally remove important information of the underlying process (Aguinis et al., 2013).

High-dimensional data can have adverse effects on classifiers (e.g., sensitivity and overfitting), and curse of dimensionality phenomenon becomes more and more critical

(Alexandropoulos et al., 2019; Li et al., 2018). In addition, high-dimension data may significantly increase required resources, for instance, computation time, money and data storage (Li et al., 2018). Therefore, effective feature selection (FS) methods, that attempt to return only a subset of significant variables from the original dataset, are developed extensively (Alexandropoulos et al., 2019). FS methods are categorized into filter, wrapper, and embedded methods, where wrapper methods rely on prediction performance of a predefined algorithm (Li et al., 2018). In contrast, filter methods are independent of any learning algorithms, and they are often computationally more efficient (Li et al., 2018). Embedded methods, like the name suggests, adds feature selection component strictly into the model, for instance, embedding a regularization term (Li et al., 2018). In addition to FS methods, feature extraction (FE) is used to dimensional reduction purposes. Unlike FS methods, feature extraction transforms current high-dimensional space into lower-dimensional space, thus creating a new set of features (Li et al., 2018).

After the final dataset is complete, a prediction model is designed. Learning phase follows common machine learning process, which involves: 1) random data partition, 2) choosing validation procedure (e.g., v-fold cross-validation), 3) hyperparameter tuning, 4) model fitting, and 5) out-of-sample evaluation (see Section 2.3.1).

When the intelligent model is fitted to training data, i.e., certain loss function is minimised and the model's parameters are defined, out-of-sample performance is measured by some performance metric(s). The chosen metrics should indicate ordering preference of alternative classifiers (Branco et al., 2016). Classification learning is based on a wide pool of different metrics, of which, the accuracy rate is the most common one (Lopez et al., 2013). However, when evaluating models in imbalanced domains, many traditional metrics, for instance accuracy, are not suitable and may lead to suboptimal solutions (Branco et al., 2016; He & Garcia, 2009). To illustrate, given that a majority class dominates 99% of the whole sample space, then simply predicting each new instance to the majority class would produce 99% accuracy rate. Obviously, this does not offer any meaningful details of the model's true prediction performance. Alternative metrics are preferred to overcome problems of imbalanced datasets, for example, 1) recall, 2) precision, 3) F-measure, 4) G-mean, 5) ROC-curve, and 6) AUC (Branco et al., 2016; He & Garcia, 2009).

3. Problems in classic financial distress prediction methodologies

Balcaen & Ooghe (2006) outlined the main problems related to classic statistical methodologies in financial distress prediction. They categorized them into four dimensions: 1) the classical paradigm, 2) the neglect of the time dimension of failure, 3) the application focus, and 4) other problems. The problems are described below, and later in the "Discussion"-section, studies of intelligent financial distress prediction (Section 4) are assessed based on their responses to these issues.

3.1 The classic paradigm

The first dimension points out the problems of the nature of failure prediction modelling, namely, supervised classification. The authors highlighted four main issues: 1) lack of standard, proper definition of output variable, 2) non-stationary and instable data, 3) non-random sampling, and 4) arbitrary choice of optimization criteria.

In most of corporate failure studies, a juridical definition, often declaration of bankruptcy, is used as an output variable in the classification task (Charitou et al., 2004). Juridical approach is considered more objective than arbitrary financial metrics where the exact date of failure can be easily specified (Charitou et al., 2004; Balcaen & Ooghe, 2006). However, bankruptcy is not universally consistent term due to various parameters (e.g., bankruptcy laws, disclosure laws) and constraints (e.g., high influence of the bankruptcy judge and the creditors' committee on the decision process, partys' relative bargaining power) (Nwogugu, 2007). In addition, bankruptcy definitions are usually based on liquidity or solvency figures (Balcaen & Ooghe, 2006). A sample of bankrupted companies may be contaminated by firms that are declared bankrupt even though they do not have any real signs of failure as well as by firms that went bankrupt due to unexpected event, like natural disasters (Balcaen & Ooghe, 2006).

Besides bankruptcy term, "financial distress" is commonly used as an output variable. Financial distress is an arbitrary metric and defined various ways in the past corporate failure literature. To name a few, the following metrics are used as an indication of financial

distress: 1) many years of negative net operating income, 2) stopping of dividend payments, 3) significant restructuring or layoffs, 4) loss-making public firms whose shares have been sold to private investors, 5) a capital restructuring or a reorganization, 6) accumulated losses (Platt & Platt, 2002; McLeay & Omar, 2000). The arbitrary definition of an output variable makes comparative analysis extremely difficult, i.e., studies are non-generalizable and high accuracy results appear only in a certain context. Like outlined in the beginning of Section 2, financial distress is a process with many stages, so it's only natural that various definitions emerge in the field.

The classic paradigm of failure prediction modelling assumes that the relationship between input and output variables remain stable over time, i.e., variables are stationary (Balcaen & Ooghe, 2006). However, there is a vast amount of evidence that this is not the case (Zmijewski, 1984; Kahya et al., 2001; Mensah, 1984). Changes in market environment, like inflation and interest rates, business cycle, and serial correlation among variables can lead to data instability and non-stationary behaviour (Mensah, 1984; Kahya et al., 2001). Commonly used pooling technique, (i.e., gathering sample of failed companies across different time periods to achieve larger set of examples), will make this particular problem even a greater concern (Mensah, 1984; Balcaen & Ooghe, 2006). Data instability results in predictive performance to deteriorate as time progress and new samples are predicted (Balcaen & Ooghe, 2006). This will, inevitably, require estimating a prediction model's parameters repeatedly, which might result contradictive outcomes among consecutive models and making prediction process unstable (Mensah, 1984).

Random sampling is a basic assumption of the classical paradigm of corporate failure prediction modelling, where a set of examples are collected by randomly drawing from a population that represents the whole population of companies. However, many studies use non-random datasets due to common practices in the field. (Balcaen & Ooghe, 2006) For instance, researchers often use choice-based sampling techniques that results non-random samples. One of the techniques is to first identify two distinct groups of distressed and non-distressed companies and then a final dataset is randomly drawn from both populations separately. (Zmijewski, 1984) In addition, failing companies tend to be younger, and hence, more prone to have incomplete data (Balcaen & Ooghe, 2006). Nevertheless, many studies apply "complete data" criterion which might lead to a sample selection bias and non-random

samples. Lastly, commonly used paired-sampling technique induce similar problems (see Section 2.3.3). A dataset endowed with non-random samples, may result in biased parameter and probability estimates and unreliable predictions. (Zmijewski, 1984)

The last problem in the classic paradigm is the choice of optimization criteria. In fact, "performance metric" would be a better term in this context than "optimization criteria" since the latter term may be misleading. Indeed, in machine learning approach, optimization relates to optimization of a loss function which is not meant in this particular problem. Nevertheless, Balcaen & Ooghe (2006) described how the choice of performance metric is often based on purely arbitrary decision without further theoretical reasoning. There are numerous performance metrics to choose from but none of them is generally accepted to be superior one. Normally, different optimization criteria lead to a different prediction model which, evidently, means that prediction models depend on the choice of performance metrics. Since the choice is done somewhat arbitrarily, it's difficult to compare performance results among academic studies and get a true sense of which of the modelling practices produce the most significant outcomes. (Balcaen & Ooghe, 2006) In addition, requirements for a performance metric differ in imbalanced domain where traditional metrics, like accuracy, are not recommended due to their inherent properties (see Section 2.3.3).

3.2 Time dimension

In a typical failure prediction modelling, only one single observation for each firm is used to build a prediction model. This static perspective conflicts with the real-world failure process, that is dynamic by nature. The use of one observation relies on the assumption that consecutive financial variables are independent and repeated measurements (Balcaen & Ooghe, 2006). This assumption has been proven to be false due to serial correlation of financial features (Theodossiou, 1993). This "one observation" practice is more likely to result a selection bias, since observation is based on one "snapshot" of a firm's financial situation which probably does not represent its true financial health (Shumway, 2001).

It is argued that failure prediction modelling should be based on the process of financial health, but this time-series behaviour is often ignored in the studies (Kahya & Theodossiou, 1999). As described in Luoma & Laitinen (1991), the classic failure prediction modelling

assumes that the failure process is steady and uniform. The steady state refers to the idea that failure process is seen as stable throughout time and the process does not hold any distinct phases (Balcaen & Ooghe, 2006). Uniform failure process indicates that the failure process is not unique, but every firm experience same dynamic path. Clearly, this contradicts with reality. Different failure paths are discussed, for example, in Laitinen (1991), and Ooghe & De Prijcker (2008).

To include time dimension to the modelling phase, several suggestions are made. One possible solution is to include trend variables into the model (Edmister, 1972). Also, the prediction model could be updated on regular basis, so that its parameter estimates are up to date (Balcaen & Ooghe, 2006). However, it will not resolve the initial problem of static perspective, and studies have shown that repeated model building may lead to instability where the signs of consecutive parameter estimates are opposite (see, for instance, Zavgren (1985) and Keasey & McGuinness (1990)). In addition, other modelling approaches have been suggested, for instance, survival analysis. Briefly, survival analysis is a statistical technique that estimates the probability of failure at each point in time by using longitudinal data (Laitinen, 2005). An interested reader can find more details, for instance, in Luoma & Laitinen (1991), Shumway (2001) and Laitinen (2005).

3.3 Application focus

The third dimension underlies the fact that many financial distress prediction models are application driven, i.e., models are built to support third-party's point-of-view or designed for commercial purposes. The models' forecasting horizon is short, and rather than investigating core reasons for financial distress, many studies simply conduct a statistical search by, first, choosing a vast number of financial variables and then apply some statistical method to find the most significant ones. (Balcaen & Ooghe, 2006) The lack of proper, in-depth theoretical analysis may result a well performing model but only in its own context. Empirical findings are not, most likely, generalizable and may generate counter-intuitive outcomes.

Also, no consensus has been reached on the superior predictor variables or on the superior prediction models. Many studies have yielded high accuracy results with different feature

sets and prediction models but concluding remarks of specific features and models are hard to make. In general, financial features related to profitability, solvency, and growth have shown their significance but there exist a lot of variability between and within different categories. Studies have argued that relatively simple modelling methods should be chosen, since more sophisticated ones often yield only marginal approvements. (Balcaen & Ooghe, 2006)

3.4 Miscellaneous

In the final dimension, several issues are pointed out. Firstly, in the classical prediction scheme, linear classification rules are used most often, even though the relationship between predictor variables and financial health could be best described through nonlinear modelling. Moreover, traditional linear models produce fixed score outcomes which contradicts with reality of dynamic financial distress process. (Balcaen & Ooghe, 2006)

Secondly, many studies rely heavily on publicly available account information. The primary problem is that preparing and publishing financial ratios is restricted which already excludes certain companies from the pool of prospective samples. Restriction is typically based on firm size thus the whole sample space is often populated only by larger companies. (Balcaen & Ooghe, 2006) Also, researchers, almost naively, assume that annual financial reports generated by a firm show a completely true view of its financial situation. Studies have shown that firms, especially distressed ones, are incentivized to manipulate financial figures in their favour (DeFond & Jiambalvo, 1994; Sweeney, 1994; Rosner, 2003). Moreover, using only financial ratios implicitly assumes that all relevant information regarding corporate failure is rooted in these ratios. However, this is not most likely the case, and many studies have suggested using non-financial and qualitative variables in the prediction models (Becchetti & Sierra, 2003; Sheppard, 1994; Doumpos & Zopounidis, 1999). Lastly, the final dimension highlights the importance of having a model with multidimensional capabilities since univariate models represent corporate failure unrealistically (Balcaen & Ooghe, 2006).

To sum up, classic financial distress prediction methodologies are somewhat filled with biased practices and hold assumptions that are usually overtaken in research process. Some of the problems are (partially) answered in the past literature but many of these issues are

still relevant and should be addressed in the following studies. To see the current trend, the next section introduces recent studies of intelligent financial distress prediction, and, in Section 5, the studies are investigated to see if and how they respond the issues presented here.

4. Literature review – Intelligent financial distress prediction

In this section, recent contributions in intelligent financial distress prediction are described. A full list of academic studies and their main objectives are depicted in Appendix 1. Studies were chosen based on the following criteria: 1) peer-reviewed article, 2) published within the last six years, 3) subject of the study relates to financial distress prediction in a binary classification framework, and 4) at least one intelligent method was utilized in a prediction task. If a primary method and secondary methods were separable in a study, the primary would have to be an intelligent one. Note, studies related to "corporate failure prediction", "bankruptcy prediction", "credit risk prediction" etc. are considered also due to arbitrary definition of financial distress. To give an example, a study's title including "financial distress prediction" may still use bankruptcy status as an indication of financial distress (see for instance, Liang et al. (2018) and Uthayakumar et al. (2020)). All in all, 36 different studies are presented here.

The analysis of the studies is categorized into five different parts: 1) objectives, 2) data, 3) methods, 4) results, and 5) conclusions. Main goals of each study are presented in the objectives part. Data part describes the datasets that were used in prediction tasks. Particularly, data source, time period, feature types and forecast horizon are reported here. Methods category summarizes prediction framework which consists of feature selection methods, dataset type, prediction models and benchmark models. Main findings and performance metrics of each study are given in results part. The last category presents final conclusions and future research topics suggested in the studies.

4.1 Objectives

Appendix 1 shows the main objectives for each study. 21 out of 36 studies focused on one method or framework and estimated its performance in prediction task. Zoricak et al. (2020) investigated three different one-class methods to solve class imbalance problem in bankruptcy prediction. Uthayakumar et al. (2020) proposed hybrid classification model for financial distress prediction based on k-means clustering and a fitness-scaling chaotic genetic ant colony algorithm. Valencia et al. (2019) proposed generalized additive model,

which is a nonparametric, statistical model. In addition, they embedded GAMSEL method for the feature selection process. Antunes et al. (2017) utilized probabilistic approach to bankruptcy prediction, namely, Gaussian process model.

Jan (2021), Smiti & Soui (2020), Sreedharan et al. (2020), Vochozka et al. (2020) and Lahmiri & Bekiros (2019) used deep learning algorithms in the prediction task. Jan (2021) focused on comparing convolutional and deep neural networks, while Vochozka et al. (2020) compared neural networks with a long short-term memory layer. Lahmiri & Bekiros (2019) investigated four different neural network topologies, BPNN, PNN, RBFNN, and GRNN. Sreedharan et al. (2020) proposed a novel hybrid approach based on multi-layer perceptron and genetic algorithm for FDP. Smiti & Soui (2020) addressed the class imbalance problem by introducing hybrid model based on Borderline-SMOTE and stacked autoencoder.

Zeng et al. (2020), Shrivastav & Ramudu (2020), and Sun et al. (2016) focused on support vector machines and their predictive performance in FDP. Zeng et al. (2020) proposed sparse PCA with SVM to improve feature selection process. Shrivastav & Ramudu (2020) utilized relief algorithm in feature selection and compared prediction power of SVMs with two different kernels, linear and RBF. Sun et al. (2016) proposed a novel hybrid approach (EBW-VSTW-SVM) for dynamic financial distress prediction. They used entropy-based weighting in a feature selection phase and final predictions were made with SVM algorithm.

Zhao et al. (2017) and Wang et al. (2017) implemented kernel extreme machine (KELM) algorithm for bankruptcy prediction. Zhao et al. (2017) used two-grid search strategy to optimize hyperparameters for kernel extreme machine algorithm and compared the optimized model to other benchmark models, such as SVM and extreme learning machine (ELM) algorithms. Wang et al. (2017) introduced grey-wolf method to optimize hyperparameters in kernel extreme machine learning algorithm. They compared grey-wolf KELM bankruptcy prediction model to other optimization techniques, for example, genetic algorithm-KELM and particle swarm optimization-KELM.

Bussmann & Giudici (2021), Shen et al. (2020), Sun et al. (2019), Wang et al. (2018), Liang et al. (2018), Sun et al. (2017), and Liu & Wu (2017) introduced a multiple classifier system approach for FDP. Bussmann & Giudici (2021) proposed XGBoost model for a company

default prediction and suggested similarity networks of Shapley values to improve model's interpretability. Shen et al. (2020) used an oversampling technique to address class imbalance problem and they introduced recursive ensemble approach to predict financial distress. Sun et al. (2019) presented dynamic prediction of relative financial distress model of SMOTE-Adaboost. Wang et al. (2018) proposed a random subspace method with SVM base classifier to increase financial distress prediction performance. It was the only study to incorporate sentiment and textual features in the prediction model. Liang et al. (2018) studied unanimous voting implementation instead of, more commonly used, majority voting approach. They compared unanimous voting to simple single models and majority voting ensemble models. Sun et al. (2017) proposed two novel ensemble approaches for dynamic FDP: DEVE-AT and ADASVM-TW. Both are based on Adaboost-SVM model. DEVE-AT uses double expert voting which consists of two models, Adaboost-SVM and Timeboost-SVM. ADASVM-TW uses Adaboost-SVM in which time-weighting of features is internally integrated into the model. Liu & Wu (2017) proposed hybrid ensemble model of incremental bagging (neural network as base classifier) with genetic algorithm to improve a model selection process.

10 out of 36 studies did not have specific model under investigation. Instead, their main objective was to compare different models and estimate the best performing method in bankruptcy or financial distress prediction. Tang et al. (2020) included 12 different learning models to study how well textual and management features can predict financial distress. Gregova et al. (2020) compared random forest, neural network model, and logistic regression in the FDP context with Slovak enterprise dataset.

Son et al. (2019) compared five ML techniques in bankruptcy prediction domain. However, their main target was to tackle financial data skewness problem via Box-Cox transformation and investigate feature importance with XGboost and lightGBM models. Huang & Yen (2019) compared several different models (supervised, unsupervised, and hybrid supervised-unsupervised learning) for FDP. Korol (2019) investigated how the performance of bankruptcy prediction changes when forecast horizon increases. The study compared models that do not require threshold manipulation (recurrent neural network, MLP, fuzzy sets, and decision trees). Zhu et al. (2017) compared single, ensemble and integrated ensemble methods for credit risk prediction. Mselmi et al. (2017) concentrated on financial distress

prediction for French small-and medium sized firms. They included statistical models, single and hybrid machine learning models in their comparative study.

Barboza et al. (2017) tested machine learning bankruptcy prediction models, such as support vector machine and random forest, against traditional statistical ones (discriminant analysis, logistic regression). Figini et al. (2017) tested whether a multivariate outlier detection technique (local outlier factor) can increase out-of-sample performance of credit risk prediction models. They compared many different models (statistical, single, and ensemble machine learning) with original dataset and original dataset with two additional features extracted from local outlier factor. Jones et al. (2017) compared total of 16 different bankruptcy prediction models of which five were based on machine learning.

The objectives of the last five studies were miscellaneously related to FDP. Perboli & Arabnezhad (2021) built a decision support system for bankruptcy prediction that utilized machine learning models. The decision support system's structure included data extraction from several different sources (public databases, AIDA etc.), data cleaning module, feature selection module, machine learning model training module, and report module. Nyitrai & Virag (2019) investigated winsorization method to handle outliers in bankruptcy dataset. They compared several machine learning models with original dataset and datasets with different winsorization methods applied to them. Lin et al. (2018) compared two feature selection methods (filter and wrapper) in the bankruptcy prediction context. They deployed single, bagging, and boosting machine learning models in the prediction task. Le et al. (2018) compared five different oversampling techniques to address class imbalance problem in bankruptcy prediction domain. Lastly, Du Jardin (2017) proposed a framework that deploys financial evolution of a firm into bankruptcy prediction models. The purpose was to improve forecast horizon up to five years by quantizing changes in a firm's financial condition.

4.2 Data

In Appendix 2, datasets employed in the studies, are shortly described in terms of data source, number of samples, time period, input feature type, output feature type, and forecast horizon. In all datasets, each sample represents information of a company. Some datasets were longitudinal, but the vast majority applied cross-sectional datasets. The number of

samples are separated by a comma, in case multiple datasets were used from the same source. Any missing information is noted as “N/A”.

The most used data sources were UCI (in seven studies), CSMAR (in six studies), and TEJ (in four studies). UCI is a free machine learning repository maintained by University of California and containing more than 550 datasets. CSMAR is a free database developed by Shenzhen CSMAR Data Technology CO., focusing on finance and economy datasets. TEJ, short for Taiwan Economic Journal, provides data on Asian companies. Du Jardin (2017) used the largest dataset (nearly 200 thousand rows), and Sun et al. (2016) used the smallest (29 rows per dataset).

Time period specifies the minimum and maximum years of which samples occurred. The longest time period was 29 years (Barboza et al. (2017)), ranging from 1985 to 2013. The narrowest was two years, Bussmann & Giudici (2021) between 2015 to 2016 and Le et al. (2018) between 2016-2017.

“Input features”-column represents type of features that were used to build prediction models in the studies. They are divided into five different categories: 1) FF (financial features), 2) FFM (financial features, including equity features), 3) MF (management features), 4) TF (textual features), and 5) OF (other features). FF represents features that are derived from a company’s balance sheet and income statement, for example, total assets and net income. FFM means that a study also used stock market features in addition to financial features, for example, shareholders’ equity and earnings per share. Management features describe structure of a company and internal control management, for example, number of directors and board members and auditor’s opinion of the company. Textual features represent information about a company’s reports, mainly annual reports. Tang et al. (2020) derived features from annual reports of Chinese listed firms, such as number of words, number of sentences, and number of positive vocabularies based on L&M dictionary. “Other features”-category includes miscellaneous features that could not be allocated to any of the previously mentioned categories. These features are, for example, qualitative risk metrics (Uthayakumar et al. (2020); Lahmiri & Bekiros (2019)) and Herfindahl index (Shen et al. (2020)).

Financial features were used in almost all studies (33 out of 36), of which ten also included stock features in the prediction models. Many of the financial features were related to company's key performance metrics, such as profitability, solvency, growth rates, and operation efficiency. Management features, textual features and other features were used in four, two, and eight studies, respectively.

"Output"-column refers to the target classes that the studies were trying to predict. Most of the studies had bankruptcy status as an output feature (21 out of 36 studies). Financial distress feature was used in 13 studies. However, financial distress was determined various ways. In Jan (2021), a company was defined financially distressed, for example, if a company was declared bankrupt, had a negative corporate net worth, or CPA had going concern doubt about the company. Gregova et al. (2020) defined a company as financially distressed if the following three criteria are met: 1) equity-to-liability ratio does not exceed 0.4, 2) the current ratio is less than 1, and 3) earnings after taxes are not positive. In many studies, that utilized Chinese datasets, a company was labelled financially distressed if China Securities Regulatory Commission (CSRM) had declared it as "Special Treatment" (ST) company (for instance, Tang et al. (2020); Zeng et al. (2020); Shen et al. (2020); Wang et al. (2018)). The ST status is assigned to a company if: a) a company has negative earnings in two consecutive years, b) net assets per share are less than the face value per share, or c) CSRM or Chinese stock exchange has identified abnormal financial behaviour (Jiang & Jones, 2018). Other than bankruptcy and financial distress, "default" and "high credit risk" were used as an output feature.

"Forecast horizon"-column refers to the time difference between input features and an output feature. For example, in Gregova et al. (2020), value of *t*-1 indicates that financial distress occurred at year *t* was predicted with input features occurred at year *t*-1. The longest forecast horizon was reported in Korol (2019), where prediction models used input features of up to ten years before bankruptcy event of a firm. The study constructed ten different models for each year, whereas Sun et al. (2016), which also had *t*-10 features, used all of the input features of different years to train one model. The shortest forecast horizon was in Huang & Yen (2019) where input features of one fiscal quarter before financial distress status were applied.

4.3 Methods

Appendix 3 depicts basic information of the methods applied in each study. The second column covers the feature selection process where type and name of the feature selection algorithm is described. A study with no FS method is indicated as "N/A". "Multiple" specifies that several FS methods were included. A type of datasets, either imbalanced or balanced, is described in the third column. In case some sampling method was used to solve imbalance problem, the method is depicted in the brackets. "Type of model(s)"-column categorized prediction models either as ensemble or single/hybrid models. Prediction models and benchmark models are presented in the last two columns. Models are abbreviated to save space and explanations can be found in Abbreviations- section. In some studies, all of the models did not fit in the table, which is indicated as an ellipsis ("…").

Feature selection methods were used in 20 studies, of which twelve, six and three utilized wrapper method, filter method and embedded method, respectively. Multiple feature selection methods were applied in six studies, and 16 studies did not use any. Lin et al. (2018) was the only study whose main objective was to compare different feature selection methods (filter and wrapper) in the bankruptcy prediction context.

Twenty-one (17) studies used imbalanced (balanced) dataset in the prediction task. Out of the 21 studies, six tried to solve the imbalance problem by using some sampling method. Other remedies for imbalance problem, such as ensemble and one class classification methods, are not depicted here because they are strictly related to prediction models. Synthetic Minority Oversampling Technique (SMOTE), or some variation of SMOTE, was the most popular sampling technique. Only Lin et al. (2018) used several different oversampling techniques. Their main objective was to compare different oversampling methods and show the effectiveness of oversampling in the bankruptcy prediction context. Balanced datasets were usually created by choosing available distressed companies and then including the same number of healthy companies. Healthy companies were often chosen by using some matching criteria so that bankrupt-healthy company pairs are comparable. For example, Perboli & Arabnezhad (2021) chose healthy companies based on similarity in revenues, and Liu & Wu (2017) used two pair matching criteria: 1) the pairwise firms should

be in the similar industry, and 2) the firms should have approximately equal total assets, within the range from -20% to +20%.

29 out of 36 studies used single or hybrid models, and 19 out of 36 utilized ensemble models in the prediction task. All the different models per study were calculated and categorized according to the type of a model. The categories' frequency bar chart is shown below (Figure 3). Benchmark models are not included.

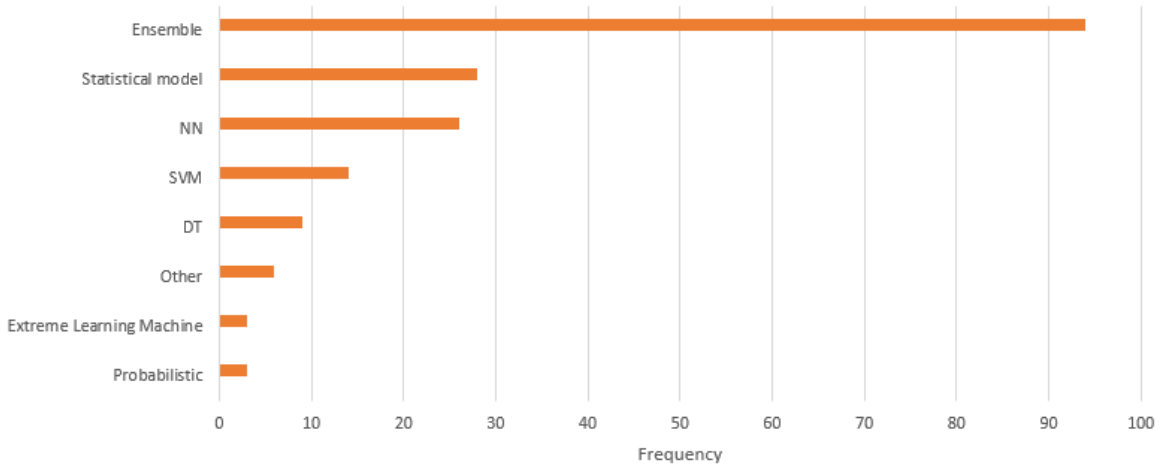

Figure 3. Categories of prediction models

Models that utilized some ensemble technique belong to ensemble category. Each ensemble technique with specific base classifier is calculated as one model. Statistical model category consists of statistical models, such as logistic regression, LDA and MDA. The importance of statistical models is somewhat misleading since only one study, Jones et al. (2017), contributed a great proportion of the total frequency (11/28). However, there were also many comparative studies that included statistical methods.

Neural network models and stacked autoencoder models belong to NN category. Each type of neural network is calculated as one model. For example, Jan (2021) utilized CNN and DNN models which counts as two models, whereas Vochozka et al. (2020) built several LSTM neural network models which only counts as one. Support vector machine models were not counted separately if the models differ only in terms of hyperparameter tuning. For example, Shrivastav & Ramudu (2020) built several SVM models with different kernel functions, but those were aggregated to one. Choosing different kernel function is really a

hyperparameter tuning, not a model selection. In contrast, SVM models in Huang & Yen (2019) were counted as two since they used normal SVM and hybrid DBN-SVM model.

Models that used decision tree technique, are allocated to DT category. These are, for example, decision tree, CART and CHAID. Probabilistic category includes models that are probabilistic in nature, for example, Gaussian process and Naïve Bayes. Extreme learning machine category includes all different ELM models, such as standard extreme learning machine and kernel extreme learning machine (KELM). Other category depicts rest of the models that were used less than three times in the studies. The category includes the following models: KNN, genetic programming, clustering, fuzzy set, and Cox's survival model.

Ensemble models with different base classifiers was the most popular category by far, with over 90 different models. Figure 4 shows the frequencies of different ensemble methods. Bagging and boosting were the most frequently used methods, followed by random subspace and REA. RS-boosting, and Multiboosting were the least popular, both utilized only once.

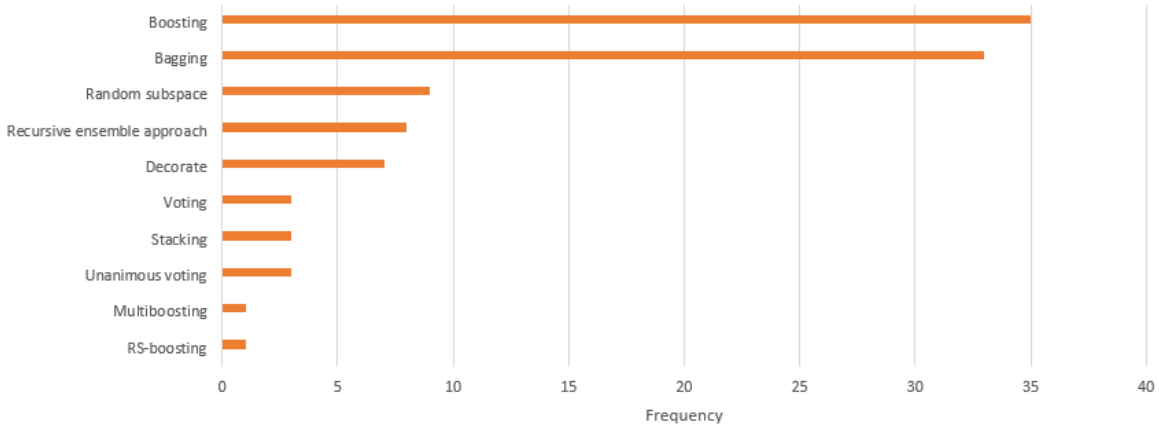

Figure 4. Bar chart of ensemble methods

Statistical model is the top-2 category. Statistical models have a strong history in financial distress prediction, and many studies utilized them together with machine learning methods. The most popular statistical models were logistic regression and discriminant analysis. Neural network models were the third most frequent category, with 26 different models.

Some studies did not specify NN model, but from those who did, multi-layer perceptron was the most common. Stacked autoencoder was utilized only in Smiti & Soui (2020).

Total of 14 different SVM models were built of which nine were single SVM models. Often, hybrid SVM models utilized common statistical methods in a feature selection phase. For instance, Mselmi et al. (2017) used partial least squares method together with SVM, and Zeng et al. (2020) used principal component analysis to design GSPCA-SVM hybrid model. Nine models utilized decision tree technique. Most of them were named as "Decision tree" or "DT". Other decision tree models were CHAID, classification tree and CART.

Three different probabilistic and extreme learning machine models were used in the studies. Probabilistic category includes Gaussian process, Naïve Bayes and least-squares anomaly detection models. Various extreme learning machine models were implemented, for instance, standard ELM, kernel extreme learning machine, and hybrid model of grey wolf optimization and KELM (GWO-KELM). Other category consists of survival model, fuzzy set, clustering, one of each, and KNN model that was used in two studies. Clustering model utilized fuzzy c-means clustering algorithm which is close to standard k-means algorithm.

Various benchmark models were designed to analyse if the proposed model outperforms in the classification task. In case benchmark models were not specifically mentioned in a study, or a study was only comparing different models, "Benchmark(s)"-column was marked as "N/A". Logistic regression was the most popular statistical benchmark model. Some studies used models that are similar to the proposed model. For instance, Wang et al. (2017) compared GWO-KELM model to other implementations of KELM (e.g., GSKELM, GA-KELM) and Uthayakumar et al. (2020) compared their k-means FSCGACA model to different variations of ACA models (e.g., standard ACA, GACA). Otherwise, benchmark models were somewhat arbitrarily chosen.

4.4 Results

In Appendix 4, main statistics of the studies' results are given. PM-column lists all performance metrics that were used to compare and analyse models. Some performance metrics are left out of the table to save space. This is indicated as an ellipsis ("…") in the

column. TM-column depicts the top performed model(s) within the study. PMO-column indicates whether a proposed model outperformed (yes/no). In case “N/A” is marked, the study did not propose any specific model or method. ST-column describes whether any statistical tests were used to find statistically significant outperformance. The final column, “Robust”, describes whether the top model had any contradictive results. “Yes” indicates that the top model outperformed in terms of all measures. A brief description is given in case of contradictive results (“No”-value in the column). To be determined as robust, a proposed model had to outperform in most test cases and researchers had to clearly determine the top model.

In the bar chart below (Figure 5), the most common performance metrics are depicted. A bar represents the number of studies. Metrics that were used only in one study are not included in the chart. Accuracy and AUC were the most common ones (20/36 and 19/36 studies, respectively). Interestingly, accuracy metric was the most popular one, given that it is not suggested to be used in imbalanced domains. Indeed, nine out of 21 studies that utilized imbalanced datasets were also using accuracy rate in a model evaluation phase. Most studies, 23 out of 36 studies, used more than one performance metric. However, only one study, Bussmann & Giudici (2021), used effective summarization technique to enhance interpretability of the proposed model.

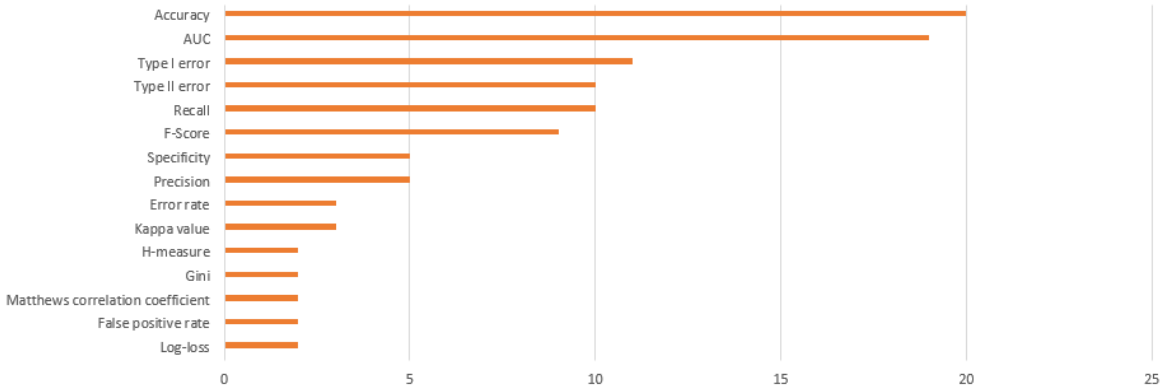

Figure 5. Performance metrics

The top performed models were, in the majority of cases, either ensemble or hybrid ones. Some studies did not specify the top model, for example, Tang et al. (2020) described only that ensemble and deep learning models produced the highest prediction scores. In Du Jardin

(2017), the top model was not an actual machine learning model, but a new framework of data segmentation combined with ensemble method. In only one study that introduced a novel method, the proposed model did not outperform. Valencia et al. (2019) introduced statistical GAM model with GAMSEL feature selection method, but the best performance was achieved with random forest. On the other hand, the proposed model yielded comparable results and had higher interpretability. Half of the studies did not propose a new model or method.

Only one third of the studies compared the results with some statistical test. The most common one was paired t-test, followed by Wilcoxon and Friedman tests. Nemenyi, DeLong and variance tests were utilized only once. 20 out of 36 studies did not report contradictive results. Of course, they were only superior in their own framework and contradictive results may be found in future research. 16 out of 36 studies reported ambiguous results, in which the "top model" was not clearly stated, or it was hard to determine.

4.5 Conclusions

In Appendix 5, conclusions and future research suggestions are described for each study. Bussmann & Giudici (2021) showed that ensemble models applied with Shapley values and similarity networks can increase interpretability and understanding of the determinants of credit risk of a company. Their model can be employed in regulated banking sector, where black box AI models are not yet suitable methods due to regulatory requirements. They suggested extending their model to other datasets, especially to imbalanced ones.

Perboli & Arabnezhad's (2021) decision support system with a machine learning module was able to make accurate forecasts up to 60 months. Their two-phase training procedure improved prediction performance. Next, they plan to expose the system to other data sources and include dynamic evolution into the machine learning module. In addition, they attend to incorporate extreme value theory to present level of uncertainty in the module.

All deep learning models in Jan (2021), achieved a high accuracy rate, convolutional networks performing the best. CHAID algorithm in a feature selection phase improved the accuracy of the models. Similarly, Vochozka et al. (2020) found out their neural network

model being capable of predicting financial distress. Their NN model was based on long short-term memory layer. Jan (2021) did not mention any specific future developments, but Vochozka et al. (2020) aims to simplify presentation of the model and ease its applicability. Lahmiri & Bekiros (2019) also compared neural network models. The generalized regression neural network outperformed in accuracy and specificity metrics, but radial basis function neural network yielded the best results in sensitivity metric. No further development was mentioned in the study.

Sreedharan et al.'s (2020) hybrid model of genetic algorithm and neural network outperformed other classic machine learning models. Smiti & Soui (2020) came to similar conclusions with their hybrid model which consisted of stacked autoencoder (SAE) and Borderline SMOTE (BSM). BSM was used to solve imbalanced data problem while genetic algorithm in Sreedharan et al. (2020) was used to tune hyperparameters of NN model. BSM was able to improve all classifiers, although the proposed model had the worst computation time. Neither study reported any future research areas.

Zoricak et al. (2020) tackled imbalanced data problem by building one-class classifiers. Here, LSAD model outperformed others in nearly all datasets and authors proposed it to further investigation. They suggested to investigate other imbalanced data approaches, like resampling techniques and cost-sensitive learning at more detailed level. In addition, they concluded that feature selection methods for imbalanced data have not received enough attention and that new features for bankruptcy prediction should be explored further.

Tang et al. (2020) found out that wrapper-based feature selection method with ensemble or deep learning model improves prediction performance of financial distress compared to single ML models. In addition, they noted that textual features can be useful in the prediction task and should be investigated further. For the future development, they suggested to apply the model to imbalanced datasets and collect financial distressed companies over a longer time horizon. Also, they highlighted the importance of exploring ways to identify financial deterioration at an earlier time.

Uthayakumar et al. (2020) proposed a novel hybrid model of k-means clustering with FCSGACA algorithm. The purpose of k-means clustering was to eliminate wrongly

clustered data. Their rule-based model was optimized using FCSGACA. The model was superior compared to other meta-heuristics like ACA and GACA. Their next target is to improve the model by introducing feature selection methods to the training process.

Gregova et al. (2020) concluded that machine learning models, especially neural network models, yielded higher performance results in almost all metrics compared to logistic regression. Their next step is to test other bankruptcy prediction methods and extend the study to longer time horizon. Similarly, Barboza et al. (2017) showed that ML models outperformed traditional statistical models. Especially, ensemble methods like bagging, boosting, and RF methods proved to outperform other models in all tests. The authors aim to extend the study to incorporate growth rates and time effect of variables into the model.

Zeng et al. (2020) introduced hybrid model of principal component analysis and support vector machine. They found out that feature selection method of sparse PCA can improve prediction performance of SVM model. The next step for them is to extend the study to imbalanced datasets, explore fuzzy theory methods in the FDP context, and to develop effective prediction model by combining financial management theory and machine learning principles. Shrivastav & Ramudu (2020) also applied SVM but only as single models. They concluded that SVM with linear kernel had better accuracy in bankruptcy prediction of Indian banks than SVM with radial basis kernel.

Figini et al. (2017) applied solvency prediction models to the same industry. Their multivariate outlier detection techniques enhanced the prediction results. In addition, they found that BGEV model, which is similar to logistic regression, obtained similar performance results as sophisticated ensemble methods. The authors suggested to analyse nonparametric techniques by applying Bayesian approaches, expert opinions, and business knowledge. Also, they mentioned that combining existing data with qualitative information, news and textual features could enhance the prediction performance.

Adaptive Neighbor SMOTE-Recursive Ensemble Approach (ANS-REA), in Shen et al. (2020), proved to be superior in predicting financial distress compared to other ensemble and hybrid models. ANS-REA with random forest as a base classifier yielded the best prediction results. The model still needs further optimization, which is left for the future

development. In addition, the study mentioned of including additional features and using data with a longer time horizon.

Skewness of financial data is a common problem, and it affects negatively to bankruptcy prediction. According to Son et al. (2019), it can be alleviated by various methods. They solved the skewness problem via Box-Cox transformation, which improved AUC 17% on average. The authors assessed that the proposed model (XGBoost) is applicable to the industry since it yielded high accuracy results.

Valencia et al. (2019) demonstrated that general additive models are comparable to machine learning models in bankruptcy prediction. The proposed GAM model with embedded variable selection not only produced comparable prediction results, but also the model had a higher level of interpretability compared to ML models. In the future, the authors will inspect the model's performance in higher dimensional setting with different country data. In addition, they are interested in exploring how the model performs with highly correlated data.

Nyitrai & Virag (2019) concluded that decision tree is a robust method for bankruptcy prediction in the presence of outliers. They introduced different winsorization methods to handle outliers and compared them to CHAID based categorization and deletion method. It turned out that CHAID based categorization achieved the best results in all datasets. Their next step is to extend the study to other data binning methodologies, to other classifiers, and to data of public companies.

Huang & Yen (2019) compared four different supervised learning models in the FDP context. The results showed that XGBoost model produced the most accurate prediction performance. Also, hybrid model of DBN-SVM yielded satisfactory results. The authors intend to implement DBN to other supervised learning approaches and extend the study to different datasets.

Korol (2019) introduced dynamic bankruptcy prediction models. The stability and the effectiveness of the forecasts were positively affected when dynamic elements were included in the model. In addition, the study concluded that fuzzy set was a superior method compared

to neural network and decision trees. The author will continue the research by including macroeconomic variables in prediction models.

Sun et al. (2019) also studied dynamic models, concentrating on hybrid ensemble approach of SMOTE-Adaboost. SMOTE-Adaboost yielded satisfactory results in several base classifiers, which indicates that the proposed FDP model can effectively be used in financial risk management. The study's empirical results showed that the dynamic industry's relative FDP model significantly improved prediction accuracy compared to stationary models. Operating capacity and solvency features had the greatest effect on the predictions. Sun et al. (2017) introduced two novel approaches for dynamic FDP, DEVE-AT and ADASVM-TW. Both models are based on Adaboost and SVM methods. ADASVM-TW yielded the best prediction results. Both models outperformed single SVM and BE-LWS methods. The authors suggested to explore applications of dynamic SVM, for instance, the incremental SVM presented by Laskov et al. (2006) and SVM for nonlinear timeseries by Mukherjee et al. (1997), and evaluate their performance relative to dynamic SVM ensemble approaches.

Liu & Wu (2017) introduced a hybrid ensemble model of incremental bagging and genetic algorithm for dynamic FDP. Genetic algorithm was used to optimize a set of base classifiers for the next iteration round in the prediction process. Neural network was used as a base classifier. The proposed model, GA-SBE, enhanced dynamic prediction performance. Sun et al. (2016) presented their novel dynamic FDP model which consists of entropy-based weighting, vertical slide time window, and SVM classifier. They showed that the proposed model is feasible and effective tool for vertical relative financial distress prediction. However, for each company data must be accumulated from a sufficiently long time period in order the prediction system being useful since relativeness measure is based on firm's own historical records.

Wang et al. (2018) introduced a novel FDP model which utilized textual and sentiment features and random subspace method. They found out that incorporating these different feature types can contribute significantly to the predictive performance. The model successfully managed high-dimensional features and the class imbalance problem. Their next step is to extend the study to different datasets and conduct a theoretical analysis of the

proposed method. They also emphasized that ensemble models are computationally intensive which is why parallel computing techniques should be investigated further.

Liang et al. (2018) presented unanimous voting (UV) for bankruptcy prediction. The results demonstrated that UV ensemble approach outperformed other classifiers, single models and majority voting ensemble approach, in all the datasets. The authors did not suggest any specific future development ideas.

Lin et al. (2018) studied feature selection methods in the bankruptcy prediction context. They concluded that, in general, implementing a feature selection phase to the prediction process leads to better performance. Specifically, they showed that wrapper-based genetic algorithm outperformed information gain method (filter-based). Their best model consisted of combination of GA and either Naïve Bayes or support vector machines.

Le et al. (2018) tested five different oversampling techniques for the bankruptcy prediction task. Their results showed that oversampling techniques may enhance the prediction performance. Particularly, SMOTE method with edited nearest neighbour (ENN) proved to be the best oversampling technique. To be more precise, SMOTE is the oversampling technique and ENN is applied afterwards to remove overlapping samples. SMOTE + ENN with random forest yielded the greatest AUC value. The authors aim to extend the study to other approaches that handle imbalanced data.

Zhu et al. (2017) investigated single, ensemble and integrated ensemble models for credit risk prediction of Chinese SME companies. They concluded that integrated ensemble models produced the highest accuracy scores in credit risk prediction. In particular, RS-boosting yielded promising results.

Zhao et al. (2017) proposed kernel extreme learning machine model with two-step grid search for bankruptcy prediction. The model outperformed other benchmark models, like SVM, ELM and PSOFKNN. It can be potentially used as an alternative early-warning system. Similarly, Wang et al. (2017) introduced a novel KELM model with grey wolf optimizer. GWO technique was used for hyperparameter tuning. The study found out that the proposed model was superior compared to six other models. In the future, Zhao et al.

(2017) aim to improve the model and explore the most important features for the prediction task. Wang et al. (2017) plan to extend to larger datasets. In addition, they are going to implement feature selection to their framework and test other meta-heuristics, such as fruit fly optimizer and multi-verse optimizer.

Mselmi et al. (2017) tested different FDP models for French SME firms. Their study came up with many contradictive results. Firstly, no optimal set of financial features were found for prediction task. For one-year prior to financial distress, a combination of six ratios had the best classifier performance. But, for two-year prior to financial distress, a combination of ten ratios had the best classifier performance, of which only two remained the same as in the one-year prediction. In addition, none of the tested models showed its absolute outperformance. The authors' next task is to add macroeconomic variables to the model and study corporate governance variables and their contribution to the prediction performance.

Antunes et al. (2017) used probabilistic approach for bankruptcy prediction. They showed that the proposed Gaussian Process model outperformed both SVM and logistic regression methods in many different scenarios and datasets. In the future, the authors try different kernel functions in their model and extend the study to other datasets with different class balance ratios.

Jones et al. (2017) showed that "new age" classifiers, like ensemble approaches (Adaboost, generalised boosting and random forest), outperformed other traditional classifiers. In addition, they concluded that many classifiers can be improved through Box-Cox power transformations. SVD approach, (for missing value imputation), showed little or no improvement to overall prediction performance. The authors aim to compare different classifiers in multinominal settings and in other business-related contexts, in the future.

Du Jardin (2017) introduced a new framework for bankruptcy prediction in which a firm's financial health changes can be quantized over time and implement into the prediction model. The author found out that the new framework did improve prediction performance, but the improvement occurs mostly when the forecast horizon exceeds three years. For future research, the author suggested to improve the output of the proposed model by changing

from a score value prediction to probability measure and to explore other types of segmentation criteria.

After an extensive list of studies, there are at least six points to highlight. Firstly, as demonstrated in the objective part, there are various research areas under financial distress prediction, from proposing new sophisticated prediction models to tuning of feature selection process or hyperparameters of intelligent meta-heuristics. It's an indication that the field is complex, multidimensional, and intensively studied.

Secondly, financial features, which are derived from companies' financial statements, are the most frequently used predictor variables. However, many studies concluded that other various data sources and feature types should be investigated even further. Interestingly, not a single study utilized macroeconomic variables in prediction models. Clearly, market conditions effect companies' performance hence market features are most likely useful in prediction of bankruptcy or financial distress. However, market changes may have very different impact on different companies, so it is not obvious which market features to choose from and what kind of companies to include in a prediction dataset. Adding market features to prediction task was encouraged by several studies. Also, like mentioned in Section 2.0, a financial indicator's prediction power varies depending on the stage of financial distress. Nevertheless, this issue was not addressed in any of the studies.

Thirdly, there were numerous methods to define bankruptcy and financial distress status of a company, for example, judicial announcements and ratios from financial reports that measure financial deterioration. It is difficult to compare classifiers from different studies and their performance if the prediction tasks have, ultimately, different objectives. Fourthly, prediction tasks were mainly for one or two years ahead, and only few studies included dynamic elements to the models. Financial distress is an on-going process, hence dynamic evolution of a company's performance should be included in the models. Also, imbalanced datasets were used more often, although only a minority of the studies addressed the imbalance problem.

Fifthly, ensemble and hybrid models were the most popular ones, and in most cases, they received the best prediction results compared to either traditional statistical models or single

machine learning models. However, the ultimate model still remains a mystery. In addition, there was no consensus on the most important predictor variables. Many studies did not report the full list of the most significant variables, and those who did, the variables differed from study to study. Typically, the variables belonged to common financial categories like profitability, solvency, and liquidation. Analysing new features and feature types was encouraged by many studies. Moreover, various performance metrics were utilized, accuracy rate being the most popular one. Accuracy metric was used somewhat extensively with imbalanced datasets, which should be avoided.

Finally, future development of FDP and bankruptcy prediction seems to take several paths. Studies that proposed a new model or framework, continue to develop them even further. They suggested, for instance, implementing algorithms for feature selection and hyperparameter tuning, and test the models with new data. Especially, imbalanced datasets were emphasized. Thorough feature investigation was also suggested several times, since there were no clear answers to which features have the greatest prediction power. It seems that there is a lack of fundamental knowledge of the financial deterioration process. However, only Zeng et al. (2020) proposed to combine financial management theory with machine learning, and Wang et al. (2018) suggested to conduct in-depth theoretical analysis of their model. In addition, dynamic properties of the financial distress process should be investigated even more. Almost all the studies derived prediction models and features from the past literature which can lead to academic stagnation. There is a room for exploring qualitative analysis that are conducted since the 1970s (e.g., Argenti (1976)), and also, for searching profound reasons of financial distress that are applicable to the prediction task.

5. Discussion

This section analyses how the problems in classic financial distress prediction methodologies are addressed in the intelligent FDP studies (Section 4). Tables 1-4 present the problems under discussion, status of the problems, and suggested solutions.

5.1 The classic paradigm

Problems of the classic paradigm are depicted below:

Table 1. Problems of the classic paradigm

Dimension	Problem	Status/respond	Solutions
The classic paradigm	Lack of standard output feature	No consensus of the definition of "financial distress". Many alternative output features are still used. "ST" status was the most popular one.	1) Collaboration of research community 2) Definition of different levels of financial distress
	Non-stationarity & data instability	Studies did not take stance on non-stationarity or data instability. Cross-sectional data, static models, and extensive use of financial features are the main practices in the field.	1) Dynamic FDP modelling 2) Search for alternative features
	Non-random sampling	Non-random sampling is still relevant in FDP. Balanced datasets are used, although the majority of the studies applied models with imbalanced data. "Complete data" criterion and "drawing from two distinct populations" were used often.	1) Imbalanced data 2) Ensemble methods 3) Cost-sensitive learning models 4) Missing value management 5) Emphasis on SME data
	Arbitrary choice of performance metric	Various performance metrics were introduced. Interestingly, accuracy was the most popular metric, which is not ideal in imbalanced domain.	1) Effective summarization techniques 2) Robust metrics for imbalanced datasets

Problems related to the classic paradigm are: 1) lack of standard output feature, 2) non-stationarity and data instability, 3) non-random sampling, and 4) arbitrary choice of performance metric. According to the studies, there was no consensus of the definition of

financial distress. Twenty-one studies used bankruptcy status, and 13 used “financial distress” as an output feature. Financial distress was defined various ways, of which “ST” status was the most popular one. However, “ST” can be used only with companies operating under China Securities Regulatory Commission (CSRM), and it seems to be standard way in that context. One interesting method of defining FD is using a relative financial performance, instead of an absolute metric. Sun et al. (2019) defined a relative financial distress feature by comparing a company’s financial performance to other companies within the same industry. In Sun et al. (2016), FD status is based on comparing a firm’s financial performance to its previous period performance. Similar trend seems to continue, where bankruptcy status, or judicial approach generally, is still the most common indicator for financial distress. As outlined in Section 3.1, bankruptcy status is not a universal metric and may mislead the interpretation of a prediction model’s performance. For instance, collecting data of bankrupted companies from different countries is problematic, since judicial practices in each country may have a great impact on the output feature.

The very definition of FD considers all the levels of financial health deteriorations (from “early warning” signals to bankruptcy), thus it’s only natural that a variety of FD definitions have been developed. In the future, studies are encouraged to define more closely the level of severity of financial distress. Indeed, multi-stage studies, like Farooq et al. (2018), are important for distinguishing different stages of FD to set up the basis for each stage. From more broad definition of FD to specific stage-definitions would be beneficial, as long as the definitions within-stage remain somewhat standard. No matter what, closer communication in research community is needed. Like shown in the previous section, some datasets are used more frequently than others, (e.g., UCI bankruptcy), demonstrating certain level of standardization in the field.

Non-stationarity and data instability problem was not properly addressed in most of the studies. The vast majority used cross-sectional data with financial features, which are usually serially correlated and non-stationary. Especially, static cross-sectional data is highly affected by non-stationarity and data instability. For instance, Jan’s (2021) sample consists of cross-sectional data from the period between 2000 and 2019, which is obviously embedded with various business cycles and different economic environments. By using a longer time period in the sampling phase, the instances are then, most likely, drawn from a

different population distribution and possibly decreasing model's prediction performance. However, sampling from a shorter time period may deliver great performance results but using the same static model in different time period will most likely produce poorer prediction outcomes. Only few studies alleviated the problems, by suggesting a dynamic FDP model. More details of the dynamic FDP modelling are given in the next subsection which concentrates on the problems of time dimension. Besides dynamic FDP modelling, searching for alternative, robust features would be beneficial to the field.

Non-random sampling is still relevant issue in FDP domain. Non-random sampling occurred, most often due to common practices, like use of balanced datasets, "complete data" criterion, and "sampling from two distinct groups". Balanced dataset with a pair-matching strategy is still a widely accepted method for data collection (for instance, Tang et al. (2020), Zeng et al. (2020), and Huang & Yen (2019)), although the majority of the studies used imbalanced datasets. However, problems of the use of imbalanced datasets should be emphasized more often in the future research (see Section 2.3). Only six out of twenty-one studies that used imbalanced datasets, actively tried to solve the issue by introducing sampling techniques (mostly SMOTE variations) to the prediction process. Although synthetic sampling leads to a non-random sampling, class imbalance problem is considered more serious threat in the prediction process. An ensemble method also mitigates class imbalance problem, and it was the most popular modelling technique in the studies. However, justification of using ensemble methods was usually based on higher prediction performance rather than solving class imbalance problem. In the future, development of multiple classifier systems or other cost-sensitive learning models is encouraged.

The "complete data" criterion was actively used in the sampling phase. Nearly half of the studies (15/36) described some technique to handle missing values. Most studies deleted instances with missing information, and only three of them used some imputation technique to fill empty data cells (Perboli & Arabnezhad, 2021; Son et al., 2019; Jones et al., 2017). The rest of the studies did not report any method to manage missing values. The number of intelligent FDP studies that investigate different imputation techniques is limited and requires more attention in the future. Also, it would be useful to focus on small- and medium-sized companies since they are more prone to have missing values and outliers (Zoricak et al, 2020). Indeed, only four studies used specifically SME data in the prediction task

(Zoricak et al., 2020; Zhu et al., 2017; Mselmi et al., 2017; Figini et al., 2017). Similarly, practice of "drawing samples from two distinct populations" was used frequently. For instance, studies with a pair-matching strategy collected first distressed samples and then identified similar companies from a non-distressed group and included them to the final set.

The problem of the choice of performance metric remains unsolved. Total of 27 different metrics were used in the studies and 15 were applied more than once. The most common metric was accuracy, followed by AUC, Type I and Type II errors. Interestingly, accuracy was the most popular one, even though it is not considered suitable in imbalanced domain. The use of various performance metrics is an indication that there is still no consensus of the optimal measurement technique. Chen et al. (2016) suggested that studies should use multiple metrics to gain more information of the model performance. The metrics can then be summarized through effective aggregation, clustering and visualization techniques. For the summarization, the study introduced multi-dimensional scaling technique which projects multi-dimensional performance data into 2D space and eases a comparison of different classifiers. In the future, studies are encouraged to use multiple metrics and investigate proper summarization techniques. Also, when applying classifiers in imbalanced domain, only suitable performance metrics should be chosen in model evaluation phase (see Section 2.3).

5.2 Time dimension

Two main problems related to time dimension are presented in table below:

Table 2. Problems of time dimension

Dimension	Problem	Summary	Solutions
Time dimension	Cross-sectional data	Practice of "one observation" is still used in the majority of cases.	1) Dynamic FDP modelling
	"Failure is uniform and steady state"	Static intelligent FDP modelling is more popular than dynamic. There was no serious effort to evaluate different failure paths and differences in the process of financial distress.	1) Dynamic FDP modelling 2) In-depth theoretical analysis

The neglect of time dimension in intelligent FDP is a concern. The vast majority of the intelligent FDP studies were still using cross-sectional data where only one observation for each company is used. This “snapshot” characteristic may result selection bias. In addition, most of the studies considered corporate failure as steady and uniform. In general, FDP models did not include any information about the dynamics of the failure process and used only static features to describe financial distress status. Finally, uniformity of failure process underlay almost all intelligent FDP models. Differences in failure processes were not truly considered. In-depth theoretical analysis of various failure paths was not included in the studies.

However, there were seven studies that introduced dynamic properties into the modelling phase. Four of the studies rationalized the need of dynamic modelling by introducing concept of drift, the main problem in dynamic data classification (Shen et al., 2020; Sun et al., 2019; Sun et al., 2017; Sun et al., 2016). Concept of drift is a phenomenon of which distribution of dataset changes as the time progress and the newest samples are included into the dataset (Schlimmer & Granger, 1986). Older samples are not representative of the new concept environment hence classification performance decreases through time (Sun et al., 2017). Three different methodologies are suggested to handle concept of drift: 1) instance selection with time window, 2) sample weighting (time-weighting), and 3) dynamic ensemble methods (Sun et al. 2017).

All three methods belong to dynamic FDP modelling which is a continuous process of collecting new samples and fitting prediction model over again. The simplest of the three methods is instance selection with fixed time window, where older data is just excluded from the training set and prediction model is fitted again with newer samples (Shen et al., 2020). Sample weighting or time-weighting simply refers to an approach in which newer samples are treated as more important than older ones (Sun et al. 2017). Dynamic ensemble methods either add new base classifiers or update weights of the existing ones when new data is introduced to the system (Sun et al., 2017).

Two of the studies that introduced concept of drift (Sun et al., 2016; Sun et al., 2019) used instance selection with fixed time window and the other two (Sun et al., 2017, Shen et al., 2020) used dynamic ensemble approach in their dynamic models. Also, Liu & Wu (2017)

used dynamic ensemble approach in their FDP modelling, by implementing bagging procedure with fixed time window and genetic algorithm.

Korol (2019) implemented dynamic properties by adding dynamic variables into a dataset. Dynamic variables were based on relative change of features' past and current values. Du Jardin (2017) embodied factors to the prediction model that may affect the change of firm's financial evolution over time. The model was based on quantizing changes in financial health of a firm, and then characterize the changes and design models that fit for each type.

Besides re-estimating models and including relative variables into the model, alternative dynamic modelling techniques have been introduced in the past literature. Bao et al. (2015) and Zhuang & Chen (2014) presented dynamic financial distress prediction based on Kalman filtering. The studies incorporated the cumulative characteristics of financial distress by using a state-space model, and Kalman filtering was used to derive models' parameter estimates. Both models yielded high prediction scores. Another popular dynamic modelling approach is survival analysis, which is a statistical method that uses longitudinal data to model the duration of time prior financial distress (Laitinen, 2005). Studies, like Laitinen (2005), Kim et al. (2016) and Kim & Partington (2015), have reported superior results of survival analysis over traditional statistical methods. Also, Cumulative Sums (CUSUM), a model that incorporates multi-period information and stationary input features, has been presented, for instance in Kahya & Theodossiou (1999).

The problems related to time dimension are relevant topics to further discussion. Studies are encouraged to implement dynamic FDP modelling to alleviate problems of cross-sectional data and non-stationarity. Like demonstrated, many interesting dynamic modelling approaches have been introduced in the past literature, but most of them are applicable with certain statistical method. Dynamic models that apply intelligent techniques are rare to some extent and require more attention in the future. Intelligent FDP domain demands more evidence of different failure paths that are exploitable in a modelling phase. Indeed, a shift from static and steady approach to dynamic modelling is needed where financial distress is treated more as a process rather than a discrete event.

5.3 Application focus

The problems related to application focus are depicted in Table 3:

Table 3. Problems of application focus

Dimension	Problem	Summary	Solutions
Application focus	Empirically selected features	Studies utilized mostly financial features. There was no consensus of the most significant features. Choice of features was based on, in the majority of cases, past research papers. There was no in-depth theoretical analysis of the financial distress.	1) In-depth theoretical analysis
	"Arbitrary" model selection	Various models were suggested, and ensemble methods seems to be the most promising and the most popular method.	1) Comparative studies

In classic financial distress prediction, common practices are often driven by application focus, where prediction models are built without deeper understanding of the fundamentals of financial distress. Similar trend seems to continue in the intelligent FDP studies. In general, numerous financial features were applied in the prediction models, and there was no common agreement on which of the variables are the most significant ones. Intelligent FDP studies focused more on efficient modelling practices, and no serious effort was made to build theoretical framework of the determinants of financial distress process.

Feature selection was, mainly, based on contributions in the past literature. Generally, the intelligent FDP studies practiced "statistical search" approach, where a pool of features were collected and a subset of features was chosen by some statistical method. The subset represents the most significant indicators of predicting a company to become financially distressed. Among the studies, there was no agreement on the feature set that provides generally the greatest results. The studies concentrated more on building efficient feature selection methods rather than using theoretical analysis. For instance, Huang & Yen (2019) used genetic algorithm in the feature selection phase, and Valencia et al. (2019) embedded GAMSEL algorithm into the GAM-model. The majority of studies (20/36) reported some feature selection method.

Besides feature selection problem, various intelligent methods were used in the prediction tasks somewhat arbitrarily. There was no consensus on the superior modelling technique and further investigations are still required. However, multiple classifier systems seem to be the current trend in FDP modelling since they were actively applied, and many studies reported them to produce the most promising performance results. A superior intelligent method is yet to be discovered and may never be. On the other hand, finding the best intelligent method should not be the top priority in the field since many of the methods are already performing reasonably well and the definition of the top model is somewhat subjective. For instance, the model can provide the best prediction power, but what about its level of interpretability and computation cost? There are many factors to consider, and different applications may result different preference order among these factors. Also, comparison of intelligent methods is demanding since performance of FDP models is case-specific where many factors, other than the method itself, influence the outcome (choice of performance metrics, datasets, features etc.).

The following suggestions are given to alleviate the problems of application focus. Since empiricism governs the current intelligent FDP domain, more research effort should be allocated to theoretical work. By understanding the fundamental reasons in financial distress process, datasets with relevant indicators or proxies of relevant indicators would be somewhat easier to design. Randomly pooling vast number of features is likely to yield misleading results since there is a higher probability that unrelated features “dominate” in a prediction task. Ensemble methods seem to show constantly higher prediction results than other modelling techniques, but more comparative studies of different techniques are still needed. Numerous different ensemble methods were implemented but none of them showed superior results. Therefore, it would be beneficial to have studies that focus only on different ensemble methods and their prediction abilities in FDP domain. However, finding the best model should not be the main priority here, but more importantly, studies should consider more carefully applicability of the models in different context. For instance, in imbalanced domain, ensemble or cost-sensitive learning algorithms are more likely to result higher prediction accuracy.

5.4 Miscellaneous

Table 4 presents two main miscellaneous problems in classic FDP methodologies:

Table 4. Miscellaneous problems

Dimension	Problem	Summary	Solution
Miscellaneous	Linearity & statistical models	Intelligent methods address this issue directly. Strict assumptions that statistical models hold, are no longer problem when intelligent methods are applied. The most popular system was multiple classifier system. Statistical models are still widely used in comparative studies and as benchmark models.	1) Apply intelligent, non-parametric techniques 2) Increase interpretability of black-box models
	Extensive use of financial features	Financial features were the most common predictor variables. Only in the minority of cases, qualitative and non-financial features (e.g., textual, sentiment) were applied.	1) Theoretical framework

Linear models, or generally statistical models, restrict modelling of financial distress due to their inherent, strict assumptions (see Section 2.2.1). The intelligent FDP studies responded this problem directly, since intelligent techniques are non-parametric, and information extraction is based on inductive learning. Intelligent techniques are capable of building models with richer and more complex structure thus giving an edge over statistical models in a prediction task. Indeed, the studies almost always reported higher prediction scores with an intelligent method than with statistical methods. However, the main problem in intelligent methods is their relatively low level of interpretability, which is one of the reasons why statistical models are still widely used in comparative studies. Intelligent models in regulated industry, like banking, are required to have at least some level of interpretability to preserve transparency in the system. Besides judicial reasons, interpretability is necessary to have in case a prediction model is part of crucial decision-making process, like in financial distress predictions. Decision-makers that apply FDP models need to understand the reasons of why a firm is heading towards financial distress to respond accordingly.

Only one study focused on enhancing model's interpretability: Bussmann & Giudici (2021) implemented SHAP framework (Shapley Additive Explanations) by computing the

importance values for each feature and then used minimal spanning tree to employ similarity networks. In the intelligent FDP domain, model's interpretability enhancement should be addressed more detailly in the future.

Linardatos et al. (2021) introduced several interpretability techniques that could be utilized with black-box models: 1) local interpretable model-agnostic explanations, 2) SHAP, 3) anchors, 4) contrastive explanations method, 5) counterfactual explanations, 6) protodash, 7) permutation importance, 8) L2X, 9) PDP, 10) ICE plots, 11) Accumulated Local Effect plots, 12) LIVE, 13) breakDown, and 14) ProfWeight. In addition, they presented interpretability methods for deep learning models: 1) gradient explanation techniques, 2) DeepLIFT, 3) Guided BackPropagation, 4) Deconvolution, 5) Class Activation Maps, 6) Layer-wise Relevance Propagation, 7) RISE, 8) Concept Activation Vectors, and 9) Deep Taylor decomposition. Finally, they depicted several sensitivity analysis methods that can enhance interpretability of the models. As demonstrated, various methods exist to overcome interpretability issues and hopefully further investigation is done in the intelligent FDP domain.

Problems related to extensive use of financial features were depicted in Section 3.3.4. The trend seems to continue, since the vast majority of predictor variables in the intelligent FDP studies were indeed financial features. On the other hand, one third of the studies used other features as well. Full list of studies and examples of other features are depicted below (Table 5):

Table 5. List of studies and examples of other than financial features.

Study	Features
Jan (2021)	Audited Big 4 (binary)
Tang et al. (2020)	Board structure, internal control information, number of text words in annual report, number of positive vocabularies
Uthayakumar et al. (2020)	Industrial risk, management risk
Zeng et al. (2020)	Corporate governance
Shen et al. (2020)	Board size, Number of independent directors
Lahmiri & Bekiros (2019)	Industrial risk, management risk
Son et al. (2019)	External auditor (binary)
Wang et al. (2018)	Sentiment
Zhu et al. (2017)	Price rigidity, liquidation and vulnerable degree of trade goods
Antunes et al. (2017)	Number of employees last year
Figini et al. (2017)	Control features: dimension, legal form
Jones et al. (2017)	Control features: firm size, age

As shown in Table 5, management factors, for instance features that describe corporate structure, were somewhat popular. Only two studies (Tang et al., 2020; Wang et al., 2018) used textual or sentiment features. Although, relatively great proportion of the studies used also other than financial features, more emphasis should still put on investigation of significant predictor variables. Intelligent FDP studies have used various feature types and absolute number of different features can be measured in hundreds. There is no consensus of the superior predictor variables and usually the shortest path is taken, i.e., selecting features based on the past literature rather than building a theoretical framework.

All in all, current trend in intelligent FDP literature points to building novel classification models with high prediction performance capabilities. High performing models are usually built by using either hybrid or ensemble approach. Imbalanced datasets are more common than balanced ones, but the problems emerging from imbalanced domain are not intensively addressed. More effort should be put on investigating proper methods to handle classification process with imbalanced datasets, for instance applying cost-sensitive learning and sampling techniques. Despite high performing models, more evidence is needed to discover truly

superior models. The problem with the introduced models is that they are applied to specific environment and context. The models should be exposed to various FDP domains to find out their true generalization capabilities. Indeed, many of the studies plan to assess their prediction model in different domains, e.g., different datasets and time periods, in the future.

Regarding the problems outlined in the Section 3, intelligent FDP studies gave some answers, but most of the issues remained unsolved. The intelligent FDP is multidimensional and complex process which demands knowledge from various domains, computer science, organizational theory etc. This has fragmented the research around the subject, where each component of the intelligent FDP process is investigated separately, and the bigger picture is left in the background.

The intelligent FDP process is exposed to subjectivity. For instance, an output feature depends on the main objective of the study where some of the applications aim to determine "early-warning" signals and others allow only for corporate default cases. If the financial distress process consists of all the levels of financial deterioration, from the first signals to bankruptcy, then studies are well justified of using whatever definition that suits their purposes. To advance in this issue, closer communication of the research community and clearly defined levels of financial distress are needed. Multi-stage studies would be interesting to assess in greater detail.

Obviously, the intelligent FDP studies contributed most to the problems of statistical models. Non-parametric techniques allow richer and more complex modelling structure that unambiguously enhances prediction accuracy. In addition, dynamic properties were introduced in some FDP models. However, most studies still used static cross-sectional data which do not consider the very nature of financial distress from the viewpoint of time and path dependence. Indeed, more emphasis should be put on the modelling phase and on new ways of including dynamic attributes into the prediction system.

6. Conclusions

Financial distress prediction is a major practice in companies that are concerned counterparties' financial health over time. Companies, whose business relies on the practice, have adapted their systems to incorporate sophisticated models that are capable of producing accurate results. The prediction task is commonly based on dichotomous approach, where current, available information (input features) is used to predict whether a company will be either distressed or non-distressed (an output feature) in the future.

FDP domain has been extensively studied subject, starting from the 1930s. Studies like Beaver (1966), Altman (1968) and Ohlson (1980) were among the first that utilized traditional statistical methods, namely, discriminant analysis and logit models. In the 1990s, models with a higher complexity emerged (neural network and classifier trees) offering more flexibility and, usually, higher accuracy. Inevitably, this led to ever increasing number of studies applying intelligent techniques which are now more common than the first-generation models. Specifically, the use of multiple classifier systems, i.e., systems based on multiple base classifiers, is current trend in the field.

Despite their high accuracy rates, fundamental problems in classic financial distress prediction methodologies are not simply solved with these intelligent techniques. Indeed, most of the problems are independent of the modelling phase. They are often rooted in common practices and assumptions applied in the prediction process. Balcaen & Ooghe (2006) categorized the problems into four dimensions: 1) the classic paradigm, 2) neglect of time dimension, 3) application focus, and 4) other problems. The classic paradigm concentrates on the problems in the very nature of classic FDP, namely in the supervised context. Four main problems were identified: 1) lack of standard output feature, 2) non-stationarity & data instability, 3) non-random sampling, and 4) arbitrary choice of performance metric. Neglect of time dimension is realized, mainly, through two separate problems: 1) use of static cross-sectional data, and 2) assumptions of uniformity and steady state of failure process. Application focus criticises the common practices of empiricism in feature and model selection. Other problems are related to linearity and statistical models in general, and extensive use of financial features. In addition, FDP is endowed with imbalance

and cost-sensitivity, referring to infrequency of corporate failure event and the preference of prediction accuracy in the minority class, respectively. Bypassing imbalance and cost-sensitivity properties in classification learning may result in poor prediction performance in the minority class.

Evidently, FDP domain holds various difficulties which are, in many studies, taken for granted. Therefore, the thesis concentrated on evaluating recent contributions in the intelligent financial distress prediction studies by, first, introducing current trends around the subject, and then assessing how the studies responded to the fundamental methodological problems. Total of 36 peer-reviewed studies were analysed. The studies were chosen based on following criteria: 1) peer-reviewed, 2) published within the last six years, and 3) an intelligent technique is used in the financial distress classification. In case separation of primary and secondary methods is possible, the primary method had to be an intelligent one.

Literature review showed that the FDP scheme is complex and multidimensional. Main objectives of the studies concentrated on introducing novel intelligent methods, but many other issues were also targeted, for instance, enhanced hyperparameter tuning and feature selection algorithms. Research field is dispersed to various sub-research questions where local optimal solutions may be found but global generalisations are still unclear.

Imbalanced datasets were more popular than balanced, but only few studies addressed the problems that arise due to this particular practice. The most common intelligent methods were multiple classifier systems and hybrid methods, which also resulted the highest prediction scores.

The response to the problems in classic financial distress prediction methodologies was ambiguous. The problems related to classic paradigm remained unsolved. The studies used various definitions for an output feature, bankruptcy status being the most common one. Non-stationarity and data instability were not actively addressed, although some studies used dynamic FDP modelling techniques to alleviate the problems. Non-random sampling is still relevant issue in the field since practices like, balanced datasets, complete data criterion and “drawing sample from two distinct populations” were often used. Also, model evaluation

was measured various ways, accuracy rate being the most popular choice despite its weaknesses in imbalanced domains.

Similarly, problems related to neglect of time dimension were not addressed properly. The practice of “one observation” and static modelling are still common in the majority of studies. No serious effort was made to assess different financial distress processes. In addition, models and features were selected somewhat arbitrarily. A great number of different features were utilized, of which financial features had clearly a dominant position. There was no consensus of the most significant predictor variables, and usually the choice of features was based on past research papers, not on a theoretical framework. Also, various intelligent models were suggested. Ensemble methods were applied the most, followed by statistical models. In general, ensemble methods outperformed, but there was no consensus of the truly superior model.

Problems related to statistical models were, obviously, addressed. Strict assumptions of statistical models were no longer an issue when flexible intelligent techniques were used. However, statistical models are still widely applied in comparative studies and as benchmark models. Only one third of the studies used non-financial features, like management factors and control variables. Nearly all the studies reported using financial features, which is direct evidence of their extensive utilization.

Suggestions for future research have also been made: Firstly, the field requires more in-depth theoretical analysis of the determinants of different financial distress processes. Intelligent FDP studies are prone to select features based on a statistical search and practical approaches without theoretical background and understanding of the true nature of the process. Also, multi-stage financial distress studies should be developed further to separate different levels of financial distress. This would help to define output features more precisely and standardize the field.

Secondly, studies are encouraged to implement dynamic intelligent methods and reject the assumptions of static and steady financial distress status. It is well recognized that financial distress is a time- and path-dependent process, and these properties should be included in the modelling phase.

Thirdly, studies should continue to use imbalanced datasets to mimic the real-world situations and alleviate the occurrence of non-random sampling. Fourthly, since ensemble methods yielded, in general, the best prediction performance, it would be beneficial to further benchmark different multiple classifier systems. Also, intelligent techniques often suffer from lack of interpretability, which is a critical barrier in many industries (e.g., banking). In order to achieve a strong position in different business sectors, intelligent techniques should put more effort on investigation of efficient techniques for model interpretability.

Finally, to ease the choice of performance metric in imbalanced framework, studies should not use only one metric, especially traditional metrics like accuracy rate. Multiple metrics will give more information of a model's performance and comparison of several models would be more robust. However, this would require implementation of effective summarization techniques to find the best performing model in a multi-dimensional metric space. Indeed, effective summarization techniques are suggested for future research.

The analysis in the thesis was based on 36 different studies ranging from 2016 to 2021. Obviously, the set consists only a small proportion of the total number of published papers in intelligent FDP domain, thus further investigation is still needed. The thesis is unique in a sense that no other review has taken a systematic approach of studying how the field has responded to the problems in classic financial distress prediction methodologies and where researchers should focus in the future. Some of the same aspects have been overviewed in the past, but not as extensively as here. Hopefully, more innovative solutions are seen in the field to alleviate the problems above and inspire the whole research community to thrive.

References

Agarwal, V. & Taffler, R. 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. Journal of Banking and Finance, 32, 8, pp. 1541-1551.

Aguinis, H., Gottfredson, R. K. & Joo, H. 2013. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organizational Research Methods, 16, 2, pp. 270-301.

Alexandropoulos, S.-A. N., Kotsiantis, S. B. & Vrahantis, M. N. 2019. Data preprocessing in predictive data mining. Knowledge Engineering Review, 34, pp. 1-33.

Altman, E. I. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23, 4, pp. 589-609.

Antal-Vaida, C. 2021. Basic Hyperparameters Tuning Methods for Classification Algorithms. Informatica Economica, 25, 2, pp. 64-74.

Antunes, F., Ribeiro, B. & Pereira, F. 2017. Probabilistic modeling and visualization for bankruptcy prediction. Applied Soft Computing, 60, pp. 831-843.

Arlot, S. & Celisse, A. 2010. A survey of cross-validation procedures for model selection. Statistics Surveys, 4, pp. 40-79.

Arogyaswamy, K., Barker, V. L. & Yasai-Ardekani, M. 1995. Firm Turnarounds: An integrative two-stage model. Journal of Management Studies, 32, 4, pp. 493-525.

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, pp. 63-93.

Bao, X., Tao, Q. & Fu, H. 2015. Dynamic financial distress prediction based on Kalman filtering. Journal of Applied Statistics, 42, 2, pp. 292-308.

Barboza, F., Kimura, H. & Altman, E. 2017. Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 15, pp. 405-417.

Bauer, J. & Agarwal, V. 2014. Are hazard models superior to traditional bankruptcy approaches? A comprehensive test. Journal of Banking & Finance, 40, pp. 432-442.

Becchetti, L. & Sierra, J. 2003. Bankruptcy risk and productive efficiency in manufacturing firms. Journal of Banking & Finance, 27, pp. 2099-2120.

Beaver, W. H. 1966. Financial Ratios as Predictors of Failure. Journal of Accounting Research, 4, pp. 71-111.

Bellovary, J. L., Giacomino, D. E. & Akers, M. D. 2007. A Review of Bankruptcy Prediction Studies: 1930 to Present. Journal of Financial Education, 33, pp. 1-42.

Branco, P., Torgo, L. & Ribeiro, R. 2016. A Survey of Predictive Modeling on Imbalanced Domains. ACM Computing Surveys, 49, 2, pp. 1-50.

Breiman, L. 1996. Bagging predictors. Machine Learning, 24, pp. 123-140.

Bureau of Business Research 1930. A Test Analysis of Unsuccessful Industrial Companies. Bulletin No. 31, Urbana, University of Illinois Press.

Bussmann, N. & Giudici, P. 2021. Explainable Machine Learning in Credit Risk Management. Computational Economics, 57, pp. 204-216.

Charitou, A. Neophytou, E. & Charalambous, C. 2004. Predicting Corporate Failure: Empirical Evidence for the UK. European Accounting Review, 13, 3, pp. 465-497.

Chen, M.-Y. 2011. Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. Computers and Mathematics with Applications, 62, pp. 4514-4524.

Chen, N., Ribeiro, B. & Chen, A. 2016. A financial credit risk assessment: a recent review. Artificial Intelligence Review, 45, pp. 1-23.

Czarnowski, I. & Jedrzejowicz, P. 2017. Learning from examples with data reduction and stacked generalization. Journal of Intelligent & Fuzzy Systems, 32, pp. 1401-1411.

DeFond, M. L. & Jiambalvo, J. 1994. Debt covenant violation and manipulation of accruals. Journal of Accounting and Economics, 17, pp. 145-176.

Dimitras, A. I., Zanakis, S. H. & Zopounidis, C. 1996. A survey of business failures with an emphasis on prediction methods and industrial applications. European Journal of Operational Research, 90, pp. 487-513.

Doumpos, M. & Zopounidis, C. 1999. A Multicriteria Discrimination Method for the Prediction of Financial Distress: The Case of Greece. Multinational Finance Journal, 3, 2, pp. 71-101.

Duarte, D. L. & Barboza, F. 2020. Forecasting Financial Distress with Machine Learning – A Review. Future Studies Research Journal: Trends and Strategies, 12, 3, pp. 528-574.

Duarte, E. & Wainer, J. 2017. Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters. Pattern Recognition Letters, 88, pp. 6-11.

Du Jardin, P. 2017. Dynamics of firm financial evolution and bankruptcy prediction. Expert Systems with Applications, 75, 1, pp. 25-43.

Edmister, R. O. 1972. An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction. Journal of Financial and Quantitative Analysis, 7, 2, pp. 1477-1493.

Eisenbeis, R. A. 1977. Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics. The Journal of Finance, 32, 3, pp. 875-900.

Elreedy, D. & Atiya, A. F. 2019. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences, 505, pp. 32-64.

Ezzamel, M. & Mar-Molinero, C. 1990. The Distributional Properties of Financial Ratios in UK Manufacturing Companies. Journal of Business Finance & Accounting, 17, 1, pp. 1-29.

Farooq, U., Qamar, M. & Haque, A. 2018. A three-stage dynamic model of financial distress. Managerial Finance, 44, 9, pp. 1101-1116.

Figini, S., Bonelli, F. & Giovannini, E. 2017. Solvency prediction for small and medium enterprises in banking. Decision Support Systems, 102, pp. 91-97.

FitzPatrick, P. 1932. A comparison of ratios of successful industrial enterprises with those of failed companies. The Certified Public Accountant, 1932 vol. 12 July-December, pp. 598-605.

Fletcher, D. & Goss, E. 1993. Forecasting with neural networks: An application using bankruptcy data. Information & Management, 24, 3, pp. 159-167.

Foreman, R. D. 2003. A logistic analysis of bankruptcy within the US local telecommunications industry. Journal of Economics and Business, 55, pp. 135-166.

Frydman, H., Altman, E. I. & Kao, D.-L. 1985. Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress. The Journal of Finance, 40, 1, pp. 269-291.

Gregova, E., Valaskova, K., Adamko, P., Tumpach, M. & Jaros, J. 2020. Predicting Financial Distress of Slovak Enterprises: Comparison of Selected Traditional and Learning Algorithms Methods. Sustainability, 12, 10.

Hall, M. A., Frank, E. & Witten, I. H. 2011. Data mining: practical machine learning tools and techniques. 3rd ed., Burlington, Morgan Kaufmann.

He, H. & Garcia, E. A. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21, 9, pp. 1263-1284.

Hillegeist, S. A., Keating, E. K., Cram, D. P. & Lundstedt, K. G. 2004. Assessing the Probability of Bankruptcy. Review of Accounting Studies, 9, pp. 5-34.

Hua, Z., Wang, Y., Xu, X., Zhang, B. & Liang, L. 2007. Predicting corporate financial distress based on integration of support vector machine and logistic regression. Expert Systems with Applications, 33, pp. 434-440.

Huang, Y.-P. & Yen, M.-F. 2019. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. Applied Soft Computing, 83.

Jackendoff, N. 1962. A Study of Published Industry Financial and Operating Ratios. Philadelphia, Temple University, Bureau of Economic and Business Research.

Jan, C.-l. 2021. Financial Information Asymmetry: Using Deep Learning Algorithms to Predict Financial Distress. Symmetry, 13, 3.

Jiang, Y. & Jones, S. 2018. Corporate distress prediction in China: a machine learning approach. Accounting & Finance, 58, pp. 1063-1105.

Jones, S., Johnstone, D. & Wilson, R. 2017. Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Framework. Journal of Business Finance & Accounting, 44, 1, pp. 3-34.

Kahya, E., Ouandlous, A. S. & Theodossiou, P. 2001. Serial correlation, non-stationarity and dynamic performance of business failures prediction models. Managerial Finance, 27, 8, pp. 1-15.

Kahya, E. & Theodossiou, P. 1999. Predicting Corporate Financial Distress: A Time-Series CUSUM Methodology. Review of Quantative Finance and Accounting, 13, pp. 323-345.

Kaur, H., Pannu, H. & Mahli, A. 2019. A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Computing Surveys, 52, 4, pp. 1-36.

Keasey, K. & McGuinness, P. 1990. The Failure of UK Industrial Firms for the Period 1976-1984, Logistic Analysis and Entropy Measures. Journal of Business Finance & Accounting, 17, 1, pp. 119-135.

Keasey, K. & Watson, R. 1991. Financial Distress Prediction Models: A Review of Their Usefulness. British Journal of Management, 2, pp. 89-102.

Kim, H., Cho, H. & Ryu, D. 2020. Corporate Default Predictions Using Machine Learning: Literature Review. Sustainability, 12, 16.

Kim, M. H., Ma, S. & Zhou, Y. 2016. Survival prediction of distressed firms: evidence from the Chinese special treatment firms. Journal of the Asia Pacific Economy, 21, 3, pp. 418-443.

Kim, M. H. & Partington, G. 2015. Dynamic forecasts of financial distress of Australian firms. Australian Journal of Management, 40, 1, pp. 135-160.

Kirkos, E. 2015. Assessing methodologies for intelligent bankruptcy prediction. Artificial Intelligence Review, 43, pp. 83-123.

Korol, T. 2019. Dynamic Bankruptcy Prediction Models for European Enterprises. Journal of Risk and Financial Management, 12, 4.

Koster, A., Sondak, N. E. & Bourbia, W. 1991. A Business Application of Artificial Neural Network Systems. Journal of Computer Information Systems, 31, 2, pp. 3-9.

Krawczyk, B. 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5, 4, pp. 221-232.

Kücher, A., Mayr, S., Mitter, C., Duller, C. & Fieldbauer-Durstmüller, B. 2020. Firm age dynamics and causes of corporate bankruptcy: age dependent explanations for business failure. Review of Managerial Science, 14, pp. 633-661.

Lahmiri, S. & Bekiros, S. 2019. Can machine learning approaches predict corporate bankruptcy? Evidence from a qualitative experimental design. Quantitative Finance, 19, 9, pp. 1569-1577.

Laitinen, E. K. 1991. Financial Ratios and Different Failure Processes. Journal of Business Finance & Accounting, 18, 5, pp. 649-673.

Laitinen, E. K. 2005. Survival Analysis and Financial Distress Prediction: Finnish Evidence. Review of Accounting & Finance, 4, 4, pp. 76-90.

Lane, W. R., Looney, S. W. & Wansley, J. W. 1986. An application of the cox proportional hazards model to bank failure. Journal of Banking and Finance, 10, 4, pp. 511-531.

Laskov, P., Gehl, C., Krüger, S. & Müller, K.-R. 2006. Incremental support vector learning: analysis, implementation, and application. Journal of Machine Learning Research, 7, pp. 1909-1936.

Le, T., Lee, M. Y., Park, J. R. & Baik, S. W. 2018. Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset. Symmetry, 10, 4.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A. & Seliya, N. 2018. A survey on addressing high-class imbalance in big data. Journal of Big Data, 5, 1, pp. 1-30.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. & Liu, H. 2018. Feature Selection: A Data Perspective. ACM Computing Surveys, 50, 6, pp. 1-45.

Liang, D. Tsai, CF., Dai, AJ. & Eberle, W. 2018. A novel classifier ensemble approach for financial distress prediction. Knowledge and Information Systems, 54, pp. 437-462.

Lin, W. C., Lu, Y. H. & Tsai, C. F. 2018. Feature selection in single and ensemble learning-based bankruptcy prediction models. Expert Systems, 36, 2.

Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. 2021. Explainable AI: A review of Machine Learning Interpretability Methods, Entropy, 23, 1.

Liu, J. & Wu, C. 2017. Dynamic forecasting of financial distress: the hybrid use of incremental bagging and genetic algorithm – empirical study of Chinese listed corporations. Risk Management, 19, pp. 32-52.

Lopez, V., Fernandez, A., Garcia, S., Palade, V. & Herrera, F. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, pp. 113-141.

Luoma, M., Laitinen, E. K. 1991. Survival analysis as a tool for company failure prediction. Omega International Journal of Management Science, 19, 6, pp. 673–678.

McLeay, S. & Omar, A. 2000. The sensitivity of prediction models to the non-normality of bounded and unbounded financial ratios, British Accounting Review, 32, pp. 213-230.

Mehta, P., Bukov, M., Wang, C.-H., Day, A. G. R., Richardson, C., Fisher, C. K. & Schwab, D. J. 2019. A high-bias, low-variance introduction to Machine Learning for physicists. Physics Report, 810, pp. 1-124.

Mensah, Y. M. 1984. An examination of the Multivariate Bankruptcy Prediction Models: A Methodological Study. Journal of Accounting Research, 22, 1, pp. 380-395.

Merwin, C. 1942. Financing small corporations in five manufacturing industries, 1926-1936. New York, National Bureau of Economic Research.

Messier, W. F. & Hansen, J. V. 1988. Inducing rules for expert system development: an example using default and bankruptcy data. Management Science, 34, 12, pp. 1403-1415.

Mitchell, T. M. 1997. Machine Learning. 1st ed. McGraw-Hill.

Mselmi, N., Lahiani, A. & Hamza, T. 2017. Financial distress prediction: The case of French small and medium-sized firms. International Review of Financial Analysis, 50, pp. 67-80.

Mukherjee, S., Osuna, E. & Girosi, F. 1997. Nonlinear prediction of chaotic time series using support vector machines. Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop, [Online], pp. 511-520.

Nakatsu, R. T. 2021. An Evaluation of Four Resampling Methods Used in Machine Learning Classification. IEEE Intelligent Systems, 36, 3, pp. 51-57.

Nwogugu, M. 2007. Decision-making, risk and corporate governance: A critique of methodological issues in bankruptcy/recovery prediction models. Applied Mathematics and Computations, 185, pp. 178-196.

Nyitrai, T. & Virag, M. 2019. The effects of handling outliers on the performance of bankruptcy prediction models. Socio-Economic Planning Sciences, 67, pp. 34-42.

Ohlson, J. A. 1980. Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18, 1, pp. 109-131.

Ooghe, H. & De Prijcker, S. 2008. Failure processes and causes of company bankruptcy: a typology. Management Decision, 46, 2, pp. 223-242.

Opitz, D. & Maclin, R. 1999. Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research, 11, pp. 169-198.

Perboli, G. & Arabnezhad, E. 2021. A Machine Learning-based DSS for mid and long-term company crisis prediction. Expert Systems with Applications, 174.

Platt, H. D. & Platt, M. B. 1990. Development of a Class of Stable Predictive Variables: The Case of Bankruptcy Prediction. Journal of Business Finance & Accounting, 17, 1, pp. 31-51.

Platt, H. D. & Platt, M. B. 2002. Predicting Corporate Financial Distress: Reflections on Choice-Based Sample Bias. Journal of Economics and Finance, 26, 2, pp. 184-199.

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, pp. 1-28.

Richardson, F. M. & Davidson, L. F. 1983. An exploration into bankruptcy discriminant model sensitivity. Journal of Business Finance & Accounting, 10, 2, pp. 195-207.

Rosner, R. L. 2003. Earnings Manipulation in Failing Firms. Contemporary Accounting Research, 20, 2, pp. 361-408.

Salchenberger, L., Cinar, E. M. & Lash, N. A. 1992. Neural Networks: A New Tool for Predicting Thrift Failures. Decision Sciences, 23, 4, pp. 899-916.

Samuel, A. 1959. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, 3, 3, pp. 210-229.

Schlimmer, J. C. & Granger, R. H. 1986. Incremental Learning from Noisy Data. Machine Learning, 1, 3, pp. 317-354.

Schoenberg, R., Collier, N. & Bowman, C. 2013. Strategies for business turnaround and recovery: a review and synthesis. European Business Review, 25, 3, pp. 243-262.

Schweizer, L. & Nienhaus, A. 2017. Corporate distress and turnaround: integrating the literature and directing future research. Business Research, 10, pp. 3-47.

Shen, F., Liu, Y., Wang, R. & Zhou, W. 2020. A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment. Knowledge-Based Systems, 192.

Sheppard, J. 1994. Strategy and Bankruptcy: An Exploration into Organizational Death. Journal of Management, 20, 4, pp. 795-833.

Shrivastav, S. K. & Ramudu, P. J. 2020. Bankruptcy Prediction and Stress Quantification Using Support Vector Machine: Evidence from Indian Banks. Risks, 8, 2.

Shumway, T. 2001. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. The Journal of Business, 74, 1, pp. 101-124.

Smith, R. & Winakor, A. 1935. Changes in Financial Structure of Unsuccessful Industrial Corporations. Bureau of Business Research, Bulletin No. 51, Urbana, University of Illinois Press.

Smiti, S. & Soui, M. 2020. Bankruptcy Prediction Using Deep Learning Approach Based on Borderline SMOTE. Information Systems Frontiers, 22, pp. 1067-1083.

Son, H., Hyun, C., Phan, D. & Hwang, H. J. 2019. Data analytic approach for bankruptcy prediction. Expert Systems with Applications, 138.

Sreedharan, M., Khedr, A. M. & El Bannany, M. 2020. A Multi-Layer Perceptron Approach to Financial Distress Prediction with Genetic Algorithm. Automatic Control and Computer Sciences, 54, pp. 475-482.

Sun, J., Fujita, H., Chen, P & Li, H. 2017. Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. Knowledge-Based Systems, 120, 15, pp. 4-14.

Sun, J., Li, H., Chang, P.-C. & He, K.-Y. 2016. The dynamic financial distress prediction method of EBW-VSTW-SVM. Enterprise Information Systems, 10, 6, pp. 611-638.

Sun, J., Li, H., Chang, 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, pp. 41-56.

Sun, J., Zhou, M., Ai, W. & Li, H. 2019. Dynamic prediction of relative financial distress based on imbalanced data stream: from the view of one industry. Risk Management, 21, pp. 215-242.

Sun, Y., Wong, A. K. C. & Kamel, M. S. 2009. Classification of Imbalanced Data: A Review. International Journal of Pattern Recognition and Artificial Intelligence, 23, 4, pp. 687-719.

Sweeney, A. P. 1994. Debt-covenant violations and managers' accounting responses. Journal of Accounting and Economics, 17, pp. 281-308.

Tam, K. 1991. Neural Network Models and the Prediction of Bank Bankruptcy. Omega, 19, 5, pp. 429-445.

Tang, X., Li, S., Tan, M. & Shi, W. 2020. Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. Journal of Forecasting, 39, 5, pp. 769-787.

The Ministry of Justice Finland 2020. Bankruptcy, Accessed 4 September 2021, Available at the https://oikeus.fi/tuomioistuimet/en/index/asiat/velatkonkurssiyrityssaneeraus/konkurssi.html

Theodossiou, P. T. 1993. Predicting Shifts in the Mean of a Multivariate Time Series Process: An Application in Predicting Business Failures. Journal of American Statistical Association, 88, pp. 441-449.

Tsai, B.-H. 2013. An Early Warning System of Financial Distress Using Multinomial Logit Models and a Bootstrapping Approach. Emerging Markets Finance & Trade, 49, 2, pp. 43-69.

Tsai, C.-H. & Chen, Y.-C. 2019. The optimal combination of feature selection and data discretization: An empirical study. Information Sciences, 505, pp. 282-293.

Uthayakumar, J., Metawa, N., Shankar, K. & Lakshamanprabu, S. K. 2020. Intelligent hybrid model for financial crisis prediction using machine learning techniques. Information Systems and e-Business Management, 18, pp. 617-645.

Valencia, C., Cabrales, S., Garcia, L., Ramirez, J. & Calderona, D. 2019. Generalized additive model with embedded variable selection for bankruptcy prediction: Prediction versus interpretation. Cogent Economics and Finance, 7, 1.

Veganzones, D. & Severin, E. 2020. Corporate failure prediction in the twenty-first century: a review. European Business Review, 33, 2, pp. 204-226.

Vochozka, M., Vrbka, J. & Suler, P. 2020. Bankruptcy of Success? The Effective Prediction of a Company's Financial Development Using LSTM. Sustainability, 12, 18.

Wang, G., Chen, G. & Chu, Y. 2018. A new random subspace method incorporating sentiment and textual information for financial distress prediction. Electronic Commerce Research and Applications, 29, pp. 30-49.

Wang, M., Chen, H., Li, H., Cai, Z., Zhao, X., Tong, C., Li, J. & Xu, X. 2017. Grey wolf optimization evolving kernel extreme learning machine: Application to bankruptcy prediction. Engineering Applications of Artificial Intelligence, 63, pp. 54-68.

Wolpert, D. H. 1992. Stacked Generalization. Neural Networks, 5, pp. 241-259.

Yang, E.-S., Kim, J. D., Park, C.-Y., Song, H.-J. & Kim, Y.-S. 2017. Hyperparameter tuning for hidden unit conditional random fields. Engineering Computations, 34, 6, pp. 2054-2062.

Yang, Z. R., Platt, M. B. & Platt, H. D. 1999. Probabilistic Neural Networks in Bankruptcy Prediction. Journal of Business Research, 44, 2, pp. 67-74.

Zavgren, C. V. 1985. Assessing the Vulnerability to Failure of American Industrial Firms: A Logistic Analysis. Journal of Business Finance & Accounting, 12, 1, pp. 19-45.

Zeng, S., Li, Y., Yang, W. & Li, Y. 2020. A Financial Distress Prediction Model Based on Sparse Algorithm and Support Vector Machine. Mathematical Problems in Engineering.

Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M. & Chen, H. 2017. An Effective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach. Computational Economics, 49, 2, pp. 325-341.

Zhou, L. 2013. Predicting the removal of special treatment or delisting risk warning for listed company in China with Adaboost. Procedia Computer Science, 17, pp. 633-640.

Zhu, Y., Xie, C., Wang, G. & Yan, X. 2017. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. Neural Computing and Applications, 28, pp. 41-50.

Zhuang, Q. & Chen, L. 2014. Dynamic Prediction of Financial Distress Based on Kalman Filtering. Discrete Dynamics in Nature and Society, vol. 2014.

Zmijewski, M. E. 1984. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. Journal of Accounting Research, 22, pp. 59-82.

Zoricak, M., Gnip, P., Drotar, P & Gazda, V. 2020. Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. Economic Modelling, 84, pp. 165-176.

Appendices

Appendix 1. Main objectives of the studies

Reference	Year	Main Objective
Bussmann & Giudici (2021)	2021	To propose an explainable artificial intelligence model for company default prediction.
Perboli & Arabnezhad (2021)	2021	To build a machine-learning based DSS for bankruptcy prediction.
Jan (2021)	2021	To propose financial distress prediction models of two representative deep learning algorithms.
Sreedharan et al. (2020)	2020	To propose a hybrid approach with MLP and GA for financial distress prediction.
Zoricak et al. (2020)	2020	To compare three one-class methods for bankruptcy prediction with highly imbalanced dataset.
Smiti & Soui (2020)	2020	To propose a novel deep learning-based model for bankruptcy prediction.
Tang et al. (2020)	2020	To compare multiple machine learning models with multiple different factors in bankruptcy prediction framework.
Uthayakumar et al. (2020)	2020	To propose a cluster-based, hybrid classification model for bankruptcy prediction.
Gregova et al. (2020)	2020	To compare three different methods for financial distress prediction.
Zeng et al. (2020)	2020	To propose a financial distress prediction model based on feature selection algorithm (sparse PCA) and support vector machine.
Vochozka et al. (2020)	2020	To propose a classification model for bankruptcy prediction using artificial neural network with long short-term memory layer.
Shrivastav & Ramudu (2020)	2020	To propose a bankruptcy prediction model of relief algorithm (feature selection) and support vector machine.

Appendix 1. Main objectives of the studies (continued)

Reference	Year	Main Objective
Shen et al. (2020)	2020	To propose a novel dynamic financial distress forecasting approach using Adaptive Neighbor SMOTERecursive ensemble approach. (oversampling + recursive ensemble classifier)
Lahmiri & Bekiros (2019)	2019	To compare various neural network models for bankruptcy prediction.
Son et al. (2019)	2019	To compare various machine learning methods for bankruptcy prediction. Focuses on data skewness and feature importance.
Valencia et al. (2019)	2019	To propose a generalized additive model (a statistical method) for bankruptcy prediction and compare it to several machine learning models.
Nyitrai & Virag (2019)	2019	To study the effect of handling outliers in intelligent bankruptcy prediction models.
Huang & Yen (2019)	2019	To compare different machine learning models for financial distress prediction.
Korol (2019)	2019	To compare four dynamic intelligent bankruptcy models.
Sun et al. (2019)	2019	To propose a novel model for dynamic prediction of relative financial distress using feature selection module, PCA, and ensemble classifiers.
Wang et al. (2018)	2018	To propose IST-RS method for financial distress prediction.
Liang et al. (2018)	2018	To propose a novel classifier ensemble approach for financial distress prediction.
Lin et al. (2018)	2018	To compare two types of feature selection methods for bankruptcy prediction models.
Le et al. (2018)	2018	To compare five oversampling techniques for bankruptcy prediction.

Appendix 1. Main objectives of the studies (continued)

Reference	Year	Main Objective
Zhu et al. (2017)	2017	To compare six machine learning methods for financial distress prediction.
Zhao et al. (2017)	2017	To propose a kernel extreme learning machine model for bankruptcy prediction.
Mselmi et al. (2017)	2017	To compare different models for financial distress prediction.
Barboza et al. (2017)	2017	To compare different models for financial distress prediction.
Antunes et al. (2017)	2017	To apply Gaussian processes for bankruptcy prediction.
Figini et al. (2017)	2017	To apply multivariate outlier detection techniques for credit risk estimation.
Sun et al. (2017)	2017	To propose two new dynamic financial distress prediction approaches.
Wang et al. (2017)	2017	To propose new parameter tuning algorithm (grey wolf optimization) for bankruptcy prediction.
Jones et al. (2017)	2017	To compare 16 classifiers for bankruptcy prediction.
Liu & Wu (2017)	2017	To propose a hybrid use of incremental bagging and genetic algorithm for financial distress prediction.
Du Jardin (2017)	2017	To propose a method that apply financial evolution of a firm into bankruptcy prediction models.
Sun et al. (2016)	2016	To propose a dynamic approach for financial distress prediction based on entropy-based weighting, SVM and a firm's vertical sliding time window.

Appendix 2. Description of datasets used in the studies

Reference	Source: # samples	Period	Input	Output	Forecast horizon
Bussmann & Giudici (2021)	ECAI: 15045	2015-2016	FF	Default	*t*-1
Perboli & Arabnezhad (2021)	AIDA: 5 x 3000	2001-2018	FF	Bankruptcy	*t*-(1 to 5)
Jan (2021)	TEJ: 344	2000-2019	FFM, MF	Financial distress	N/A
Sreedharan et al. (2020)	Osiris: 613	2010-2019	FF	Financial distress	N/A
Zoricak et al. (2020)	N/A: Thousands	2010-2016	FF	Bankruptcy	*t*-(1 to 3)
Smiti & Soui (2020)	UCI: 7027, 10173, 10503, 9792, 5910	2002-2013	FF	Bankruptcy	*t*-(1 to 5)
Tang et al. (2020)	CSMAR: 424	2014-2018	FF, MF, TF	Financial distress	*t*-(3 to 5)
Uthayakumar et al. (2020)	UCI: 250, 43405 Pietruszkiewicz: 240	2000-2013[1a]	FF, OF, MF	Bankruptcy	*t*-(1 to 5)[1a]
Gregova et al. (2020)	registeruz.sk: 168914	2015-2018	FF	Financial distress	*t*-1
Zeng et al. (2020)	RESSET & CSMAR: 376	2015-2019	FFM, MF	Financial distress	*t*-3
Vochozka et al. (2020)	Albertina: 5500	2014-2018	FF	Bankruptcy	N/A
Shrivastav & Ramudu (2020)	N/A: 59	2000-2017	FF	Bankruptcy	*t*-(1 to 4)

1a: Only UCI: 43405

Appendix 2. Description of datasets used in the studies (continued)

Reference	Source: # samples	Period	Input	Output	Forecast horizon
Shen et al. (2020)	CSMAR: 1496	2007-2017	FFM, MF, OF	Financial distress	*t*-(2 to 6)
Lahmiri & Bekiros (2019)	UCI: 250	N/A	OF, MF	Bankruptcy	N/A
Son et al. (2019)	NICE: 977940	2011-2016	FF, OF	Bankruptcy	N/A
Valencia et al. (2019)	Superintendencia de Sociedades de Colombia: 2922	2012-2013	FF	Bankruptcy	*t*-1
Nyitrai & Virag (2019)	N/A: 2996 UCI: N/A	2001-2016	FF	Bankruptcy	*t*-2
Huang & Yen (2019)	TEJ: 64	2010-2016	FF	Financial distress	*t*-(1Q to 4Q)
Korol (2019)	N/A: 600	2004-2017	FFM	Bankruptcy	*t*-(1 to 10)
Sun et al. (2019)	N/A: 712	2000-2015	FFM	Financial distress	*t*-1
Wang et al. (2018)	CSMAR: 1726	2011-2015	FFM, TF	Financial distress	*t*-(3 to 5)
Liang et al. (2018)	UCI: 690, 1000 TEJ: 440, 688	N/A	N/A	Bankruptcy, Credit risk	N/A
Lin et al. (2018)	UCI: 689, 1000 TEJ: 440	N/A	N/A	Bankruptcy, Credit risk	N/A
Le et al. (2018)	N/A: 120355	2016-2017	FF	Bankruptcy	*t*-1

Appendix 2. Description of datasets used in the studies (continued)

Reference	Source: # samples	Period	Input	Output	Forecast horizon
Zhu et al. (2017)	CSMAR: 57	2012-2013	FF, OF	Credit risk	*t*-1
Zhao et al. (2017)	Pietruszkiewicz: 240	1997-2001	FF	Bankruptcy	*t*-(2 to 5)
Mselmi et al. (2017)	DIANE: 212	2010-2013	FF	Financial distress	*t*-(1 to 2)
Barboza et al. (2017)	NYU: 14331	1985-2013	FFM	Bankruptcy	*t*-1
Antunes et al. (2017)	DIANE: 1334, 2000, 2000 UCI: 690, 1000, 653	2002-2006[2a]	FF, OF[2a]	Bankruptcy	*t*-1[2a]
Figini et al. (2017)	UniCredit: 38036	2014-2015	FF, OF	Bankruptcy	*t*-1
Sun et al. (2017)	CSMAR: 932	2000-2012	FF	Financial distress	*t*-2
Wang et al. (2017)	Pietruszkiewicz: 240 JPNBDS: 152	1995-2009[3a]	FF	Bankruptcy	*t*-3[3b]
Jones et al. (2017)	Standard and Poor's Capital IQ: 30129	N/A	FF, OF	Bankruptcy	*t*-(1 and 3)
Liu & Wu (2017)	CCER: 326	2005-2014	FFM	Financial distress	*t*-(1 to 3)
Du Jardin (2017)	DIANE: 190700, 111350, 194360	1996-2010	FFM	Bankruptcy	*t*-(1 to 6)
Sun et al. (2016)	N/A: 29, 29	1996-2010	FFM	Financial distress	*t*-(1 to 10)

2a: Only DIANE
3a: Only JPNBDS
3b: Only Pietruszkiewicz

Appendix 3. Methods applied in the studies

Reference	Feature selection	Dataset type	Type of model(s)	Model(s)	Benchmark(s)
Bussmann & Giudici (2021)	N/A	Imbalanced (N/A)	Ensemble	XGBoost	LR
Perboli & Arabnezhad (2021)	RFE (wrapper)	Balanced (N/A)	Ensemble & single/hybrid	XGBoost, RF, LR, NN	N/A
Jan (2021)	CHAID (wrapper)	Imbalanced (N/A)	Single/hybrid	CNN, DNN	N/A
Sreedharan et al. (2020)	N/A	Imbalanced (N/A)	Single/hybrid	NN (MLP)	SVM, DT
Zoricak et al. (2020)	Multiple (wrapper & filter)	Imbalanced (N/A)	Single/hybrid	OCSVM, IF, LSAD	SVM
Smiti & Soui (2020)	SAE (wrapper)	Imbalanced (Borderline-SMOTE)	Single/hybrid	BSM-SAE	KNN, DT, SVM, ANN...
Tang et al. (2020)	RFE (wrapper)	Balanced (N/A)	Ensemble & single/hybrid	SVM, XGBoost, GBDT, DT…	N/A
Uthayakumar et al. (2020)	N/A	Imbalanced & balanced (N/A)	Single/hybrid	k-means-FSCGACA	Olex-GA, ACA, GACA…
Gregova et al. (2020)	N/A	Imbalanced (N/A)	Ensemble & single/hybrid	RF, NN, LR	N/A
Zeng et al. (2020)	Sparse PCA (filter)	Balanced (N/A)	Single/hybrid	GSPCA-SVM	OF-SVM, PCA-SVM…
Vochozka et al. (2020)	N/A	N/A	Single/hybrid	NN (LSTM layer)	N/A
Shrivastav & Ramudu (2020)	Relief (filter)	Imbalanced (N/A)	Single/hybrid	SVM	N/A

Appendix 3. Methods applied in the studies (continued)

Reference	Feature selection	Dataset type	Type of model(s)	Model(s)	Benchmark(s)
Shen et al. (2020)	Multiple (filter)	Imbalanced (Adaptive neighbor-SMOTE)	Ensemble	ANS-REA (multiple base)	SMOTE, ANS, RWO…
Lahmiri & Bekiros (2019)	N/A	Balanced (N/A)	Single/hybrid	BPNN, PNN, RBFNN, GRNN	Regression trees
Son et al. (2019)	N/A	Imbalanced (N/A)	Ensemble & single/hybrid	XGBoost, LightGBM, NN, RF, LR…	N/A
Valencia et al. (2019)	GAMSEL (embedded)	Imbalanced (SMOTE)	Single/hybrid	GAMSEL-GAM	RF, SVM, LR, LDA, GAM
Nyitrai & Virag (2019)	Multiple (wrapper)	Balanced (N/A)	Single/hybrid	CHAID, CART, NN (MLP)…	N/A
Huang & Yen (2019)	Genetic algorithm (wrapper)	Balanced (N/A)	Ensemble & single/hybrid	XGBoost, hybrid GA-fuzzy clustering…	N/A
Korol (2019)	Correlation test (filter)	Balanced (N/A)	Single/hybrid	NN (MLP), NN (recurrent), fuzzy sets…	N/A
Sun et al. (2019)	Plus-L-minus-R (wrapper)	Imbalanced (SMOTE)	Ensemble	SMOTE-Adaboost	N/A
Wang et al. (2018)	Lasso (Embedded)	Imbalanced (N/A)	Ensemble	IST-RS (SVM base)	Ensemble SVMs
Liang et al. (2018)	N/A	Imbalanced & balanced (N/A)	Ensemble	Unanimous voting	Ensemble & single models
Lin et al. (2018)	Multiple (wrapper & filter)	Imbalanced & balanced (N/A)	Ensemble & single/hybrid	Bagging, Boosting, NN, SVM…	N/A
Le et al. (2018)	N/A	Imbalanced (multiple)	Ensemble & single/hybrid	RF, DT, NN, SVM	N/A

Appendix 3. Methods applied in the studies (continued)

Reference	Feature selection	Dataset type	Type of model(s)	Model(s)	Benchmark(s)
Zhu et al. (2017)	N/A	Imbalanced (N/A)	Ensemble & single/hybrid	RS-boosting, bagging…	N/A
Zhao et al. (2017)	N/A	Balanced (N/A)	Single/hybrid	KELM	SVM, ELM, RF...
Mselmi et al. (2017)	Stepwise regression (wrapper)	Balanced (N/A)	Single/hybrid	ANN, SVM, PLS, PLS-SVM, Logit	N/A
Barboza et al. (2017)	N/A	Balanced (N/A)	Ensemble & single/hybrid	RF, SVM, ANN, LR, MDA, Boosting...	N/A
Antunes et al. (2017)	N/A	Imbalanced & balanced (N/A)	Single/hybrid	Gaussian Process	SVM, LR
Figini et al. (2017)	N/A	Imbalanced (N/A)	Ensemble & single/hybrid	GLM, LDA, BGEV, KNN, Bagging...	N/A
Sun et al. (2017)	Stepwise MDA (wrapper)	Balanced (N/A)	Ensemble	DEVE-AT, ADASVM-TW	SVM, BE-LWS…
Wang et al. (2017)	N/A	Balanced (N/A)	Single/hybrid	GWO-KELM	GSKELM, GA-KELM...
Jones et al. (2017)	Multiple (wrapper & embedded)	Imbalanced (N/A)	Ensemble & single/hybrid	Adaboost, Generalized boosting…	N/A
Liu & Wu (2017)	N/A	Balanced (N/A)	Ensemble	GA-based SBE	IB, EW, ES
Du Jardin (2017)	Multiple (filter)	Imbalanced (undersampling)	Ensemble & single/hybrid	Bagging, Boosting…	N/A
Sun et al. (2016)	EBW (filter)	Balanced (N/A)	Single/hybrid	EBW-VSTW-SVM	SFDP

Appendix 4. Results of the studies

Reference	PM*	TM*	PMO*	ST*	Robust
Bussmann & Giudici (2021)	AUC	XGBoost	Yes	N/A	Yes
Perboli & Arabnezhad (2021)	AUC, F-score, Log-Loss…	XGBoost	N/A	N/A	Yes
Jan (2021)	Accuracy, F-score, Precision…	CHAID-CNN	N/A	N/A	Yes
Sreedharan et al. (2020)	Accuracy, F-score	MLP-GA	Yes	N/A	Yes
Zoricak et al. (2020)	AUC, GM	LSAD	Yes	N/A	No: LSAD was not the best model in all datasets.
Smiti & Soui (2020)	AUC	BSM-SAE	Yes	Paired t-test	Yes
Tang et al. (2020)	Accuracy, AUC, F-score…	Ensemble & deep learning models	N/A	N/A	No: Other models yielded comparable results. No clear winners.
Uthayakumar et al. (2020)	Accuracy, Error rate…	FSCGACA (with k-means clustering)	Yes	N/A	Yes
Gregova et al. (2020)	AUC, Error rate, Gini, Log-Loss…	NN	N/A	N/A	No: NN model did not outperform in all metrics.
Zeng et al. (2020)	Accuracy, F-score, Precision…	GSPCA-SVM	Yes	Variance, Friedman	No: GSPCA-SVM was not the best model in all datasets.
Vochozka et al. (2020)	Accuracy	NN (14-940-Tanh-Tanh-2-1)	N/A	N/A	No: Other models yielded comparable results. No clear winners. Overfitting occurred in the test set.
Shrivastav & Ramudu (2020)	Accuracy	SVM (Linear kernel)	N/A	N/A	Yes

**PM=Performance metric; TM=Top model; PMO=Proposed model outperformed; ST=Statistical tests*

Appendix 4. Results of the studies (continued)

Reference	PM*	TM*	PMO*	ST*	Robust
Shen et al. (2020)	AUC, F-score, G-score, Kappa	ANS-REA (RF)	Yes	N/A	Yes
Lahmiri & Bekiros (2019)	Accuracy, Recall, Specificity	GRNN	N/A	N/A	No: GRNN did not outperform in all metrics.
Son et al. (2019)	AUC, False positive rate, Recall	XGBoost	N/A	Friedman, Nemenyi	No: Other models yielded comparable results. No clear winners.
Valencia et al. (2019)	AUC	RF	No	N/A	No: Proposed model yielded comparable results and improved interpretation.
Nyitrai & Virag (2019)	AUC	NN (MLP)	N/A	Wilcoxon test	No: Other models yielded comparable results. No clear winners.
Huang & Yen (2019)	Accuracy	XGBoost	N/A	N/A	Yes
Korol (2019)	Overall effectiveness …	Fuzzy sets	N/A	N/A	Yes
Sun et al. (2019)	Accuracy	Dynamic hybrid models	Yes	N/A	Yes
Wang et al. (2018)	Accuracy, AUC, Type I-II error	IST-RS	Yes	Paired t-test	No: Other models yielded comparable results. No clear winners.
Liang et al. (2018)	Accuracy, Misclassification cost…	UV ensemble	Yes	Wilcoxon	Yes
Lin et al. (2018)	Type I error	GA-SVM	N/A	Wilcoxon	No: GA-SVM was not the best model in all datasets.
Le et al. (2018)	AUC	SMOTE +ENN-RF	N/A	N/A	Yes

**PM=Performance metric; TM=Top model; PMO=Proposed model outperformed; ST=Statistical tests*

Appendix 4. Results of the studies (continued)

Reference	PM*	TM*	PMO*	ST*	Robust
Zhu et al. (2017)	Accuracy, F-score, Precision…	RS-boosting	N/A	N/A	Yes
Zhao et al. (2017)	Accuracy, AUC, Type I-II error	Two-step grid search KELM	Yes	N/A	No: Other models yielded comparable results. No clear winners.
Mselmi et al. (2017)	Accuracy, AUC, Kappa…	PLS-SVM	N/A	N/A	No: Other models yielded comparable results. No clear winnesr.
Barboza et al. (2017)	Accuracy, AUC, Recall…	Machine learning models	N/A	N/A	Yes
Antunes et al. (2017)	Accuracy, F-score, Precision…	Gaussian process	Yes	Friedman	No: Other models yielded comparable results. No clear winners.
Figini et al. (2017)	AUC, Gini, H-measure, KS…	GBM & RF	N/A	DeLong	No: Other models yielded comparable results. No clear winners.
Sun et al. (2017)	Accuracy	ADASVM-TW	Yes	Paired t-test	Yes
Wang et al. (2017)	Accuracy, AUC, Type I-II error	GWO-KELM	Yes	Paired t-test	Yes
Jones et al. (2017)	AUC	Generalised boosting (CT)	N/A	N/A	Yes
Liu & Wu (2017)	AUC, Type I-II error	GA-SBE (NN)	Yes	Paired t-test	No: GA-SB (NN) was not the best model in all datasets.
Du Jardin (2017)	AUC, H-measure…	"New framework" with ensemble	Yes	N/A	Yes
Sun et al. (2016)	Error rate	EBW-VSTW-SVM-DFDP	Yes	N/A	Yes

**PM=Performance metric; TM=Top model; PMO=Proposed model outperformed; ST=Statistical tests*

Appendix 5. Conclusions and future work of the studies

Reference	Conclusions	Future work
Bussmann & Giudici (2021)	Network based explainable AI models can produce more understandable results in credit risk predictions.	Apply to imbalanced data.
Perboli & Arabnezhad (2021)	Two-phase training procedure improved the performance of the considered ML methods. Accurate forecasts were made up to 60 months ahead.	Apply to extra datasets; Dynamic evolution
Jan (2021)	All four FDP models achieved a high accuracy rate. CNN had higher prediction accuracy than DNN. CHAID improved accuracy of models.	N/A
Sreedharan et al. (2020)	Neural network (MLP) model with genetic algorithm (hyperparameter tuning) had higher FDP performance compared to classic machine learning models.	N/A
Zoricak et al. (2020)	The proposed bankruptcy prediction model based on one-class LSAD achieved prediction scores from 76% to 91%.	FS & other approaches with imbalanced data
Smiti & Soui (2020)	BSM-SAE had the best performance compared to other machine learning techniques. BSM technique improved all classification models. The proposed model had inferior training time performance.	N/A
Tang et al. (2020)	Employing wrapper-based FS method with ensemble or deep learning models improved prediction performance compared to single ML models in the FD context.	Apply to imbalanced data + longer time horizon; To study "early-warning" signs.
Uthayakumar et al. (2020)	The integration of improved K-means clustering with FCSGACA increased the classification accuracy. The proposed FDP model showed the best results in all performance metrics.	FS methods.
Gregova et al. (2020)	The machine learning models showed higher performance compared to traditional LR model. Particularly, NN model yielded better results measured by almost all performance metrics (2%-22% improvement).	Other approaches; Apply to longer time horizon.
Zeng et al. (2020)	The proposed method (GSPCA-SVM) had better forecast effect. By selecting fewer, relatively more important variables through sparse PCA, a FDP model's performance can be improved.	Apply to imbalanced datasets + fuzzy theory methods.
Vochozka et al. (2020)	The proposed NN model can predict financial distress of a company. The model is flexible and can be trained on different datasets for different environments.	Model simplification and applicability enhancement.
Shrivastav & Ramudu (2020)	SVM with linear kernel had better accuracy in bankruptcy prediction compared to radial basis kernel.	N/A

Appendix 5. Conclusions and future work of the studies (continued)

Reference	Conclusions	Future work
Shen et al. (2020)	RF outperformed other classifiers. The ANS-REA algorithm was superior to the benchmark methods.	Model optimization; Apply extra features + longer time horizon.
Lahmiri & Bekiros (2019)	A model based on GRNN architecture can yield accurate bankrupt predictions.	N/A
Son et al. (2019)	By solving the skewness problem of financial data, AUC was improved 17% on average. The model produced high accuracy results, hence it is applicable to the industry.	N/A
Valencia et al. (2019)	Results showed that the GAM model with embedded variable selection has a good prediction performance. Classic ML methods outperformed the proposed model by a small margin.	Apply extra features and data; Highly correlated variables.
Nyitrai & Virag (2019)	In the presence of outliers, DTs are robust, NN and linear models are not. Outlier handling improves the prediction performance. CHAID-based categorization was the best method for outlier handling.	Apply to public companies; Other data binning methodologies & classifiers.
Huang & Yen (2019)	The XGBoost model produced the most accurate prediction performance in the FDP context. The hybrid model, DBN-SVM, yielded reasonable financial distress predictions.	Apply other ML approaches & datasets.
Korol (2019)	Dynamic elements positively affect the effectiveness and the stability of the forecast of bankruptcy. The fuzzy sets model was superior to others.	Apply macroeconomic variables.
Sun et al. (2019)	The model was effective for corporate financial risk management. An industry's relative FD concept drift was discovered via PCA. Operating capacity and solvency features had the greatest impact on the predictions.	N/A
Wang et al. (2018)	Feature integration may increase the prediction performance. The introduced model is capable of managing high-dimensional features and the class imbalance problem.	Apply different data; In-depth analysis; Parallel computing techniques.
Liang et al. (2018)	the UV ensemble approach outperformed other methods in all four datasets.	N/A
Lin et al. (2018)	A feature selection step leads to better performance in most classifiers. GA outperformed the information gain method. GA-Naïve Bayes and GA-SVM yielded the best performance.	N/A
Le et al. (2018)	Oversampling techniques may enhance the prediction performance. SMOTE + ENN with RF classifier achieved the greatest AUC values.	Apply to imbalanced data.

Appendix 5. Conclusions and future work of the studies (continued)

Reference	Conclusions	Future work
Zhu et al. (2017)	Two IEML (integrated and embedded ML) methods obtained better performance results than the others. RS–boosting showed the best results.	N/A
Zhao et al. (2017)	The proposed model performed better than other five advanced models. It can be used as an early warning system in financial decision-making.	Improve the model; Explore financial ratios.
Mselmi et al. (2017)	None of the models outperformed. Two years prior to financial distress produced the best prediction results.	Apply macroeconomic and corporate governance features.
Barboza et al. (2017)	The three machine learning techniques (boosting, bagging and RF) yielded the best results. In most cases, RF outperformed other models.	Apply growth rates and time effects to the models.
Antunes et al. (2017)	The probabilistic GP classifier outperformed both SVM and LR methods in many datasets.	Apply to different datasets with different balance ratios; Test other kernel functions.
Figini et al. (2017)	Multivariate outlier detection techniques enhance predictive power in bankruptcy prediction of SMEs. BGEV obtained similar performance results than ensemble methods.	Apply ensemble methods; Apply qualitative and textual features.
Sun et al. (2017)	ADASVM-TW and DEVE-AT yielded the highest average testing accuracy. Their performance was significantly better than the benchmark models.	Apply dynamic SVM.
Wang et al. (2017)	the proposed GWO-KELM significantly outperformed other advanced models.	Apply different datasets & other meta-heuristics.
Jones et al. (2017)	“New age” classifiers, such as AdaBoost, outperformed all other classifiers on test data. Prediction performance can be improved by Box-Cox transformation.	Apply different classifiers.
Liu & Wu (2017)	The proposed model enhanced dynamic prediction performance.	N/A
Du Jardin (2017)	The model with integrated financial evolution yielded the best forecasts. The proposed segmentation method with ensemble models improved accuracy.	Improve the model. Apply different segmentation criteria.
Sun et al. (2016)	EBW-VSTW-SVM-DFDP is effective in the VRFD prediction. The model requires relatively long historical data to perform optimally.	N/A