Article

Data Mining Algorithms Based on an Optimization Model

Mirpouya Mirmozaffari [1]*, Noorbakhsh Amiri Golilarz [2] and Shahab S. Band [3,4]*, Amir Mosavi[5]*

[1] Department of Industrial Manufacturing and Systems Engineering, the University of Texas at Arlington, Arlington, TX, USA
[2] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; noorbakhsh.amiri@std.uestc.edu.
[3] Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
[4] Future Technology Research Center, College of Future, National Yunlin University of Science and Technology 123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan, R.O.C.
[5] Faculty of Civil Engineering, Technische Universität Dresden, 01069 Dresden, Germany

Abstract: The main purpose of this paper is to propose a novel optimization model with a new data mining approach in the first section to achieve the best results in financial institutions in the second section. Since the constancy of efficacy derived from parametric and non-parametric is not significant, this paper provides a scientific assessment at the optimization section and proposes a novel combined parametric and non-parametric model which will be a new experiment in literature perception. A scientific assessment of banks based on a combination of the efficiency measurement method of CCR (Charnes, Cooper and Rhodes model) or CRS (Constant Return to Scale) BCC (Banker, Charnes, and Cooper model) or VRS (Variable Return to Scale) in Data Envelopment Analysis (DEA), as well as Stochastic Frontier Approach (SFA) for 65 banks during Feb to July 2020, are introduced. For analyzing the performance of the parametric and non-parametric approaches we have considered the linear regression and Unreplicated Linear Functional Relationship (ULFR). At the data mining section, a novel four-layers data mining filtering pre-processes for selected supervised classification as well as unsupervised clustering algorithms to increase the accuracy and to remove unrelated attributes and data are applied. For the four kinds of preprocessing approaches of unsupervised attributes, supervised attributes, supervised instances, and unsupervised instances, we have chosen discretization, attribute selection, stratified remove folds, and resample filters respectively. Based on the nature of the suggested financial institution's dataset and attributes, the most appropriate preprocessing filter in each layer to achieve the highest performance is suggested. Finally, the superior bank, best performance model, and the most accurate algorithm are introduced. The results indicate that the bank number 56 is the superior bank. Among the proposed techniques, the novel recommended CVS compared with CCR-BCC and SFA models, has a more positive correlation with profit risk, and show a higher coefficient of determination values. Sequential Minimal Optimization (SMO) algorithm receives the highest accuracy in all four suggested filtering layers.

Keywords: Data Envelopment Analysis; Data mining; Optimization; Parametric and non-parametric methods; Supervised and unsupervised models; CVS model

1. Introduction

There are various applications for data mining which are classified in “supervised” and “unsupervised” learning tools. In the first one, it is possible to predict the output concerning one or more input using a statistical model. In the latter, there is no need for a dependent variable estimation for data analyzing. The data analysis will be done for defining the data set’s structure. Cluster analysis is categorized as unsupervised learning for recognizing the complex relationships among the variables. Admittedly, specifying the target variable is not required for the initial dataset. The importance of all the variables is the same since the certain value’s prediction is not the target of the analysis. Establishing the predestined set of classes or presenting a previous data collection-based training stage is not required here. The qualitative (discrete) dependent variable’s prediction is included in the classification problems.

 doi:10.20944/preprints202009.0729.v1

A unique idea is conducted in this paper to measure the efficiency of financial firms utilizing the DEA combined with the SFA model applied on a unique dataset concerning the stock market. Additionally, regression analysis and ULFR are used for identifying the profit risk's impact on efficiency. Admittedly, these achievements are very beneficial in the management of doing proper decision making. Based on the abovementioned statements, the objective of this research is as follows: In the first stage, for obtaining the technical efficiency we have used DEA and SFA. Next, the combination of DEA combined with SFA is evaluated using DEA and SFA average scores for acquiring technical efficiency. Then, we examined the efficiency impact on the profit risk for obtaining the most effective technique. At the end, forgetting the error-free efficiency, ULFR model is utilized.

2. Background

Making decisions is a very complex process and needs the consideration of many different factors [1-3]. Identifying the most productive company is a kind of critical decision in the stock market since the variables are correlated to the evaluation procedure of the performance of the companies. Investors' earning may rise dramatically sometimes because of taking advantage of each discernible trend in the series of stock price. Nowadays, considering only one stock price is not enough for recognizing the most productive company. Investors have tried various methods for optimizing their return and minimizing the investing risks [4]. Only a productive company or market can provide the investors with a precise signal from the stock price. Therefore, efficiency provides us with burgeoning economically, but inefficiency causes lots of ups and downs in the market [5]. Recently, analyzing frontier efficiency has attracted lots of attention. Admittedly, since corporate performance evaluation deals with a variety of inputs/outputs then this measurement becomes a kind of an important task [6]. This analysis provides investors with lots of advantages regarding nonfinancial performance as well [7].

A variety of techniques already existed for performance evaluation of parametric/ nonparametric approaches [8]. The SFA model turned into one of the most common techniques recently [9]. Once input/output selection is not right thereafter the score will not be valid anymore [10]. On the other hand, the DEA model does not have these kinds of drawbacks, and also there are no random errors in this model [11-15]. Thus, these two methods have both advantageous and disadvantageous [16]. Some researchers (e.g. [17-19]) are of the opinion that these models' efficiency persistence is not critical. A plus point is that Fernandes, *et al.* [20] and Altunbas, *et al.* [21]indicated that profit risk and efficiency are related to each other. Moreover, they have realized that efficiency is influenced by the profit risk positively. To make it clear, the ratio of net income on the whole asset is defined by profit risk. There may be some errors in the chosen data because of the manipulation of the datasheet or the availability of some missing values. Therefore, in this research, finding the error-free efficiency using an un-replicated linear functional relationship model (ULFR) firstly proposed by Adcock in 1877 [22] is the main contribution.

Nowadays, in the various industries, particularly in financial sector, optimization and machine learning widely are used [23-37]. Financial institutions are always trying to find the most valuable and promising techniques for managing and mining the data for having better risk management (machine learning employing as "RegTech") or for competing with other FIs and FinTechs. Machine learning and optimization are employed in this study and also, we will discuss some application cases in the FIs.

Lots of the researchers are of the opinion that choosing a proper data mining method is dependent to the analyst's experience [12,38]. Ozekes and Camurcu [39], proposed an application for classification and prediction in the data mining field in which the decision tree will be created by the credits which the bank gives to the customers. Thus, the repayment status of customers' credits will be predicted easily.

Albayrak and Yilmaz [40], has done a research about the data mining and the data of Istanbul Securities Exchange (İMKB) in which the financial indicators have been benefited in the period of 2004-2006 of the 100 businesses in industry and service sector operating in İMKB 100 indexes and the decision tree is utilized as a data mining method. The CHAID algorithm has been employed for securing the data with the companies' financial information. The decision tree helps to determine the enterprise's positions and the critical variable which influences the sector with respect to the outcomes of the research.

A research based on the clustering techniques about the commercial banks' financial portions available in Turkish Banking Sector during the years of (1998–2006) has been done by Doğan (2008). The analysis is done based on the suitability of the results based on the financial portions.

Aşan [41], collected the customers' socio-economic characteristics with credit cards. In this study, the clustering method is utilized, and credit cards and individual banking have been specified first, and then the customers utilized the credit cards have been put into set. These customers have been categorized in three main groups based on socio-economic characteristics. Finally, novel optimization and data mining approaches can be addressed through different recent studies in the literature such as [34-36,42-54]

However, far too little attention has been paid to developing an efficient solution method to cope with DEA and data mining associated with finding the superior model, algorithm, and DMU. In conclusion, the main contributions of this study can be summarized as follows:

1. This research aims to study a comprehensive comparison of several efficiencies deliver insight into the bank's efficiency based on data mining combined with innovative optimization models. This comparison is of considerable significance to banks' practitioners who desire to assess productivity and efficiency at a proper step of its progression.
2. An exclusive and a novel CVS optimization model is introduced which will be a new experiment in literature perception. The proposed approach has a more positive correlation with profit risk and shows a higher coefficient of determination values.
3. After applying the abovementioned optimization part, best data mining supervised and unsupervised algorithms are introduced. DEA inputs and outputs as potential attributes for data mining algorithms in WEKA is considered. Besides, data play the role of instances, and finally, efficient DMUs are applied for class yes and inefficient DMUs for class no.
4. Unique filtering in the preprocessing approach designed by experts based on the nature of data and algorithms is introduced to increase the accuracy of algorithms.
5. After using the above novel combined optimization and data mining approaches, the superior model, bank, and algorithm, are proposed. Thus, it can be beneficial for managers to remove unrelated data and conduct more effective processes.

The rest sections of this paper are: Section 3 provide a flowchart or an overview and an assessment process of combining optimization models and data mining algorithms. Section 4 presents a clarification of the data set description and inputs and outputs evaluation. Section 5 reviews the five parts of the research methodology (part1: non-parametric model (consist of four subsections of: 1.CCR-BCC model 2. Parameters of non-parametric model 3. Primal proposed model 4. Dual proposed model), part2; parametric model (consist of five subsections of: 1. Stochastic Frontier Analysis (SFA) 2. Technical efficiency for SFA 3. Translog function 4. Parameters of parametric Model 5. Cost function translog form in the current study) part3: Combination of non-parametric and parametric models or CCR-BCC and SFA models or CVS proposed model, part4: The profit-risk evaluator or linear regression in this study, part5: ULFR model, Section 6 presents an evaluation with discussion and consist of five parts(part1: technical efficiency evaluation resulting from CCR-BCC , SFA and CVS proposed models, part2: a comparison of average technical efficiency of CCR-BCC , SFA and CVS models, part 3: evaluation of regression and ULFR, part 4: efficiency evaluation of CVS proposed model after applying ULFR error-free method, part 5: Evaluation of three suggested models after applying preprocessing approach for supervised and unsupervised algorithms). The conclusion and the future works of the experimental consequences are presented in Section 7.

3. Assessment process of combining optimization models and and data mining algorithms

Figure 1 shows the step by step of the flowchart for the proposed approaches in the current study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

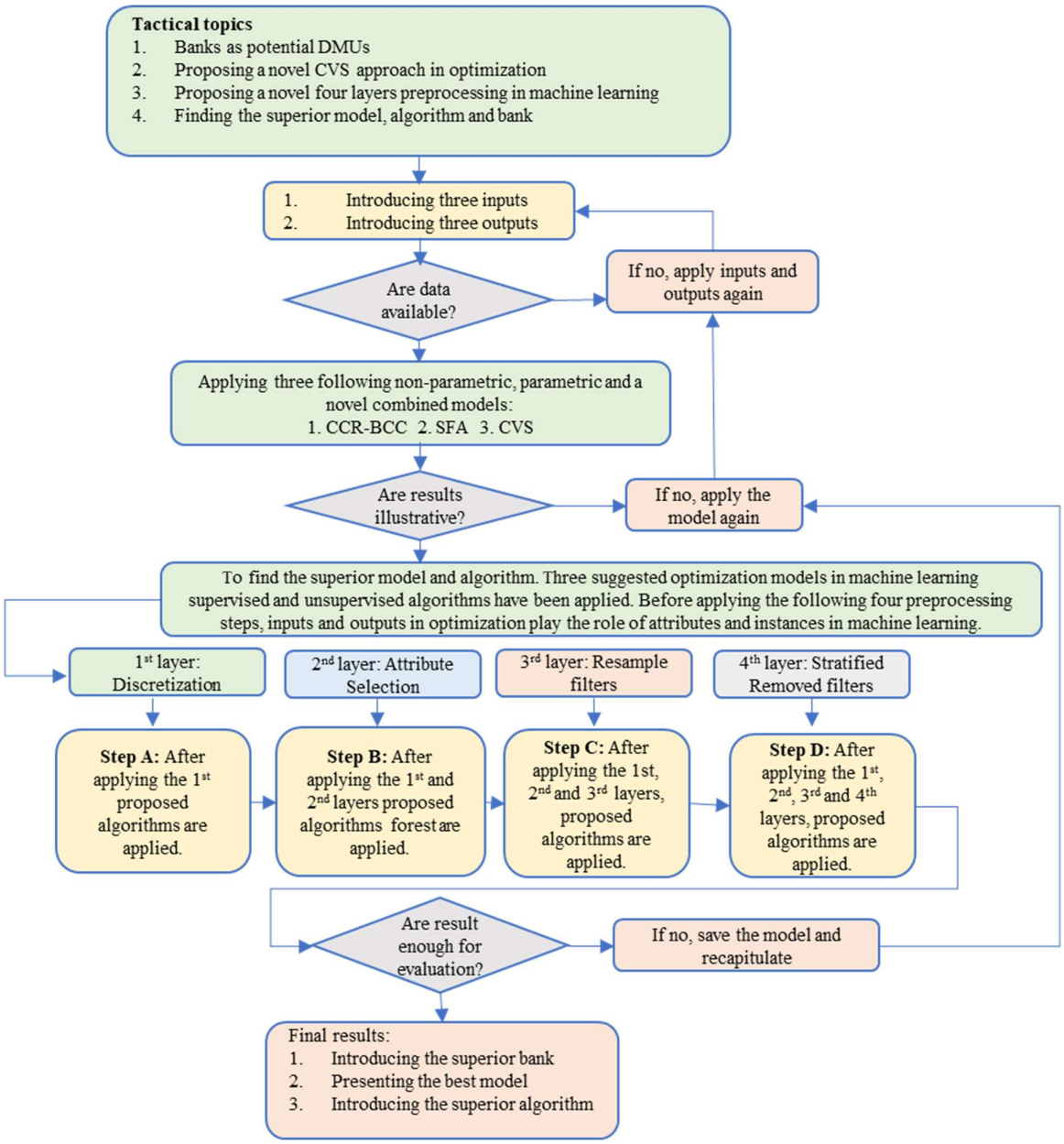

Figure 1. Flowchart of proposed combined optimization and data mining algorithms.

4. Dataset and inputs and outputs description

The standard data set, collected in this study covers six months from Feb to July 2020, which is collected from 65 banks. Three inputs and three outputs are presented in Table 1.

Table 1 Three inputs and three outputs.

Type	Variable (stat)	Unit
Input	Total deposits (X)	10000 USD
Input	Total operating expenses (N)	10000 USD
Input	Total provisions (M)	10000 USD

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

Output	Total loans (Y)	10000 USD
Output	Total other earning assets (E)	10000 USD
Output	Other operating income (F)	10000 USD

Table 2 shows the descriptive analysis of data from big data extraction during the six months.

Table 2 Expressive analysis of the data.

Stat	Max	Min	Mean	SD
X	880621.89	889.87	51389.89	115906.14
N	9719.44	8.91	611.64	1189.16
M	17589.38	0.78	605.19	1787.89
Y	688499.78	5.73	36017.13	80876.14
E	359019.34	14.96	17109.25	43017.28
F	5461.98	0.79	159.48	462.28

After applying the DMU j (j=1…n) which are 65 banks, the following three inputs and three outputs in our study which are demonstrated in figure 2 is introduced:

The three following inputs show a substitution of the investment by banks:

X_{ij} (i= 1, . . .,m): Total deposits (total deposits + total money market funding + total other funding)

N_{cj} (c= 1, . . ., k): Total operating expenses ((personnel expenses + other administrative expenses + other (i.e., those relating to the non-traditional activities noted above) operating expenses))

M_{hj} (h = 1, . . .,d): Total provisions (loan loss provisions + other provisions)

The three following outputs show the substitution of the main goal line by banks:

Y_{rj} (r = 1, . . .,s): Total loans (total customer loans + total other lending)

E_{tj} (t = 1, . . .,v): Total other earning assets

F_{zj} (z = 1, . . .,q): Other operating income

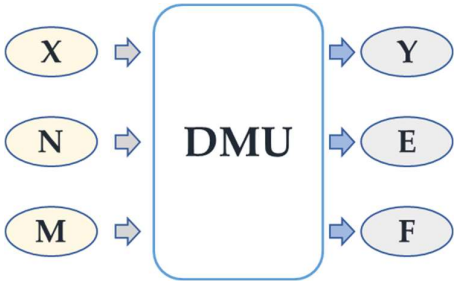

Figure 2. Three inputs and three outputs for the three suggested models.

5. Research methodology

This study aims to compare the efficiency of companies properly. To get the features of the bank, a comparative parametric combined with non-parametric methodology is settled accordingly in terms of some DMUs with three suggested models through Feb to July 2020. Then the whole progression is described below:

5.1. Non-parametric model

The non-parametric part is divided into the following four sub-sections:

5.1.1. CCR-BCC Model:

Note that in the manufacturing advances once X_0 and Y_0 are generated then λY_0 can be generated by λX_0 when $\lambda \leq 1$. A set of production possibilities consisting

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

of observations can be made for employing the convexities and feasibilities' fundamentals. This series is described below.

$$T_{CCR-BCC} = T_{ND} = \left\{(X, Y) \mid X \geq \sum_{j=1}^{n} \lambda_j X_j \,\&\, Y \leq \sum_{j=1}^{n} \lambda_j Y_j \,\&\, \sum_{j=1}^{n} \lambda_j \leq 1 \,\&\, \lambda \geq 0\right\} \quad \textit{Eq. 1}$$

Note that the DMU's measurement objective with input X and output Y with respect to the cited advances is fully described below:

T is the set of possible production

In the input-oriented approach, the main aim is obtaining a virtual unit in which the input θX_0 is not more than X_0 and at least produce Y_0. In fact:

$$Min\theta$$
$$St.$$
$$(\theta X_0, Y_0) \in T_{ND} \quad \text{Eq. 2}$$

Based on the T_{ND} structure for $CCR_{IO} - BCC_{IO}$:

$$Min\theta$$
$$St.$$
$$\sum_{j=1}^{n} \lambda_j x_{ij} \leq \theta_p \qquad ,i = 1, \dots\dots, m$$
$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{rp} \qquad ,r = 1, \dots\dots, s \quad \textit{Eq. 3}$$
$$\sum_{j=1}^{n} \lambda_j \leq 1$$
$$\lambda_j \geq 0 \qquad ,j = 1, \dots, n$$

5.1.2. Parameters of non-parametric Model

Description of dimensionless parameters in nomenclature for the primal and dual non-parametric models are provided in Table 3.

Table 3. Explanation of dimensionless parameters in nomenclature for the primal and dual proposed model.

Dimensionless parameters	Units
DMU_j	Decision making units
λ_j	Non-negative scalar (dual variables that categorize the benchmarks for inefficient parts)
X_{ij}	m^{th}input (Total deposits) for n^{th}DMU
N_{cj}	k^{th}input (Total operating expenses) for n^{th}DMU
M_{hj}	d^{th}input (Total provisions) for n^{th}DMU
Y_{rj}	s^{th}output (Total loans) for n^{th}DMU
E_{tj}	v^{th}output (Total other earning assets) for n^{th}DMU
F_{zj}	q^{th}output (Other operating income) for n^{th}DMU
w	Free of sign for variable return to scale
j	n^{th}DMU
n	DMU observation
m	Input (Total deposits) observation
k	Input (Total operating expenses) observation
d	Input (Total provisions) observation
s	Output (Total loans) observation
v	Output (Total other earning assets) observation
q	Output (Other operating income) observation
i	m^{th}input (Total deposits)
c	k^{th}input (Total operating expenses)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

h	d^{th} input (Total provisions)
r	s^{th}output (Total loans)
t	v^{th}output (Total other earning assets)
z	q^{th}output (Other operating income)
v_i	Weight assigned to input i (Total deposits)
g_c	Weight assigned to input c (Total operating expenses)
l_h	Weight assigned to input h (Total provisions)
u_r	Weight assigned to output r (Total loans)
b_t	Weight assigned to output t (Total other earning assets)
a_z	Weight assigned to output z (Other operating income)
ϕ	Scalar and real primal-variable representing the value of efficiency score
θ	Scalar and real dual-variable representing the value of efficiency score
θ_p	Free of sign dual scalar for p^{th}DMU
X_{ip}	m^{th}dual input (Total deposits) for p^{th}DMU
N_{cp}	k^{th} dual input (Total operating expenses) for p^{th}DMU
M_{hp}	d^{th} dual input (Total provisions) for p^{th}DMU
Y_{rp}	s^{th} dual output (Total loans) for p^{th}DMU
E_{tp}	v^{th} dual output (Total other earning assets) for p^{th}DMU
F_{zp}	q^{th} dual output (Other operating income) for p^{th}DMU

5.1.3. Primal proposed model in $\mathrm{CCR_{IO}} - \mathrm{BCC_{IO}}$:

$$Max\ \phi\ \sum_{r=1}^{s} Y_r u_r + \sum_{t=1}^{v} E_t b_t + \sum_{z=1}^{q} F_z a_z$$

$$\sum_{i=1}^{m} X_i v_i + \sum_{c=1}^{K} N_c g_c + \sum_{c=1}^{d} M_h l_h \ \leq 1$$

$$\sum_{r=1}^{s} Y_r u_r + \sum_{t=1}^{v} E_t b_t + \sum_{z=1}^{q} F_z a_z - \sum_{i=1}^{m} X_i v_i \qquad \text{Eq. 4}$$

$$- \sum_{c=1}^{K} N_c g_c - \sum_{c=1}^{d} M_h l_h \ \leq 0 \quad j = 1, \dots, n$$

$$v_r, g_c, l_h, b_t, u_r, a_z \geq 0 \ , j = 1, \dots, n$$

5.1.4. Dual proposed model in $\mathrm{CCR_{IO}} - \mathrm{BCC_{IO}}$:

$$Min\ \theta$$

$$St.$$

$$\sum_{j=1}^{n} \lambda_j X_{ij} \leq \theta_p X_{ip}$$

$$\sum_{j=1}^{n} \lambda_j N_{cj} \leq \theta_p N_{cp} \qquad \text{Eq. 5}$$

$$\sum_{j=1}^{n} \lambda_j M_{hj} \leq \theta_p M_{hp}$$

$$\sum_{j=1}^{n} \lambda_j E_{tj} \geq E_{tp}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

$$\sum_{j=1}^{n} \lambda_j Y_{rj} \geq Y_{rp}$$

$$\sum_{j=1}^{n} \lambda_j F_{zj} \geq F_{zp}$$

$$\sum_{j=1}^{n} \lambda_j \leq 1$$

$$\lambda_j \geq 0 \quad \theta_p \; free$$

Finally, to measure the technical efficiency of CCR-BCC model, DEA-SOLVER is applied in this study.

5.2. Parametric model

The parametric parts are divided to the following five sub-sections:

5.2.1. Stochastic Frontier Analysis (SFA)

The SFA model is specified as:

$$Y_{it} = f(X_{it}; \beta) + \varepsilon_{it} \qquad \textit{Eq. 6}$$

Where, Y_{it} is the output of bank i (i = 1, 2, . . , N) at time t (t = 2, . . . , T); $f(\beta)$ is the production technology; X_{it} is a vector of n inputs; and β is the vector of unknown parameters to be assessed. Based on the unbalance factors during the various period. The error term is specified as:

$$\varepsilon_{it} = V_{it} - U_{it} \qquad \textit{Eq. 7}$$

Where V_{it} is statistical noise (expected to be individualistically and identically disseminated); characterizes those effects which cannot be measured by the banks, such as quality, access to raw material, labor market conflicts, trade problems, measurement errors in the dependent variable, and left-out illustrative variables. $U_{it} \geq 0$, demonstrates technical inefficiency. On the other hand, V_{it} shows those factors which can be controlled by banks.

5.2.2. Technical efficiency for SFA

Based on the above equation 8, U_{it} demonstrates technical inefficiency and according to the following equation 9, U_i is the inefficiency level of the i^{th} bank at time T and ρ is an unknown parameter:

$$U_{it} = U_i \, e^{-\rho(t-T)} \qquad \textit{Eq. 8}$$

Equation 10 shows TE_{it} which is technical efficiency for the i^{th} bank at t^{th}time period based on the suggested SFA model:

$$TE_{it} = e^{-U_i \, e\hat{}(-\rho(t-T))} = e^{-U_{it}} \qquad \textit{Eq. 9}$$

5.2.3. Translog function

Production function assessment generally provides an assessment of Cobb-Doubles (CD) or Constant Elasticity of Substitution (CES) functions. Both CD and CES functions fit in the class of production functions which satisfy the condition of quasi-concavity and positive monotonicity. However, both aforementioned functional forms have some restrictions on the parameters such as the elasticity of exchange. Recent studies have demonstrated that the translog function to be more suitable in the assessment of production associations since it has more flexibility. This function has both linear and quadratic terms with the advantages of applying multiple inputs and outputs. The four-inputs and two outputs of the translog production function can be applied in terms of logarithms as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

$$\ln(Y_{it}) = \beta_0 + \beta_1 \ln(X_{it}) + \beta_2 \ln(N_{it}) + \beta_3 \ln(M_{it}) + \frac{1}{2}\beta_4 \ln(X_{it}^2) + \frac{1}{2}\beta_5 \ln(N_{it}^2) + \frac{1}{2}\beta_6 \ln(M_{it}^2) + \beta_7 (\ln(X_{it}) \times \ln(N_{it})) + \beta_8 (\ln(X_{it}) \times \ln(M_{it})) + \beta_9 (\ln(N_{it}) \times \ln(M_{it})) \quad Eq. 10$$

Y_{it} are the variables of the outputs for the i^{th} unit at time t. X_{it} is the first input for the i^{th} unit at time t. N_{it} is the second input for the i^{th} unit at time t. M_{it} is the third input for the i^{th} unit at time t. β_0 is the intercept or the constant term. β_1, β_2, *and* β_3 are first derivatives. β_4, β_5 *and* β_6 are own second derivatives. β_7, β_8, *and* β_9 are cross second derivatives.

5.2.4. Parameters of parametric Model

Description of dimensionless parameters in nomenclature for the parametric model are provided in Table 4:

Table 4. Explanation of dimensionless parameters in nomenclature for the proposed parametric model.

Dimensionless parameters	Units
$\ln(X_{it})$	Natural logarithm of Total deposits for the i^{th}bank at time t
$\ln(N_{it})$	Natural logarithm of Total operating expenses for the i^{th} bank at time t
$\ln(M_{it})$	Natural logarithm of Total provisions for the i^{th} bank at time t
$\ln(Y_{it})$	Natural logarithm of the variable of outputs (Total loans for the first output, Total other earning assets for the second output and other operating income for the third output) for the i^{th} bank at time t
β_0	Intercept or the constant term
β_1	First derivatives of the first input or natural logarithm of Total deposits for the i^{th} bank at time t
β_2	First derivatives of the second input or natural logarithm of Total operating expenses for the i^{th} bank at time t
β_3	First derivatives of the third input or natural logarithm of Total provisions for the i^{th} bank at time t
β_4	Own second derivatives of the first input or natural logarithm of Total deposits for the i^{th} bank at time
β_5	Own second derivatives of the second input or natural logarithm of Total operating expenses for the i^{th} bank at time t
β_6	Own second derivatives of the third input or natural logarithm of Total provisions for the i^{th} bank at time t
β_9	Cross second derivatives of the first and the second inputs or combination of natural logarithm of Total deposits for the i^{th}bank at time t with natural logarithm of Total operating expenses for the i^{th} bank at time t
β_{10}	Cross second derivatives of the first and the third inputs or combination of natural logarithm of Total deposits for the i^{th}bank at time t with natural logarithm of Total provisions for the i^{th} bank at time t
β_{11}	Cross second derivatives of the second and the third inputs or combination of natural logarithm of Total operating expenses for the i^{th}bank at time t with natural logarithm of Total provisions for the i^{th} bank at time t
X_{it}	Total deposits for the i^{th}bank at time t
N_{it}	Total operating expenses for the i^{th} bank at time t
M_{it}	Total provisions for the i^{th} bank at time t
Y_{it}	The variable of outputs (Total loans for the first output, Total other earning assets for the second output and other operating income for the third output) for the i^{th} bank at time t
V_{it}	Random error for the i^{th} bank at time t
U_{it}	Non-negative random variable (or technical inefficiency) for the i^{th} bank at time t

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

5.2.5. Cost function translog form in the current study

Implementing a proper functional form is kind of a critical task for completing the model assessment. Translog and Cobb-Douglas cost functions are the most recognized methods for research, particularly in estimating the units' efficiency. The Translog function is utilized frequently. It is a simplification of the Cobb-Douglas function and it's a flexible functional form based on the second-order approximation. Cobb-Douglas and translog functions are linear in terms of parameters and are measured via least-squares techniques. Based on our study the translog function is employed due to the availability of multi-inputs and outputs. Here the translog form is utilized with four inputs and two outputs provided in the following equation:

$$\ln(Y_{it}) = \ln(Y_{it}) = \beta_0 + \beta_1 \ln(X_{it}) + \beta_2 \ln(N_{it}) + \beta_3 \ln(M_{it}) + \frac{1}{2}\beta_4 \ln(X_{it}^2) + \frac{1}{2}\beta_5 \ln(N_{it}^2) + \frac{1}{2}\beta_6 \ln(M_{it}^2) + \beta_7 (\ln(X_{it}) \times \ln(N_{it})) + \beta_8 (\ln(X_{it}) \times \ln(M_{it})) + \beta_9 (\ln(N_{it}) \times \ln(M_{it})) + (V_{it} - U_{it}) \quad \text{Eq. 11}$$

5.3. Combination of non-parametric and parametric models or CCR-BCC and SFA models or CVS proposed model

Based on the equation 11, for measuring the efficiency score of the novel proposed CVS model, the averge efficiency score of CCR-BCC and SFA models should be considered:

$$CVS = \frac{Efficiency\ score\ amount\ of\ CCR - BCC\ model + Efficiency\ score\ amount\ of\ SFA\ model}{2} \quad \text{Eq. 12}$$

5.4. The profit risk evaluator or linear regression in this study

To evaluate the positive or negative effects of profit risk on the efficiency score of CCR-BCC ,SFA and CVS models, three following linear regression forms have been applied:

$$EF(CCR - BCC)_i = \beta_{CCR-BCC(0)} + \beta_{(CCR-BCC)_i} Apr_{(CCR-BCC)_i} + \varepsilon_{(CCR-BCC)_i} \quad \text{Eq. 13}$$

$$EF(SFA)_i = \beta_{SFA(0)} + \beta_{(SFA)_i} Apr_{(SFA)_i} + \varepsilon_{(SFA)_i} \quad \text{Eq. 14}$$

$$EF(CVS)_i = \beta_{(CVS)0} + \beta_{(CVS)_i} Apr_{(CVS)_i} + \varepsilon_{(CVS)_i} \quad \text{Eq. 15}$$

Equations 13,14 and 15 represent the average technical efficiency scores of the bank i based on CCR-BCC, SFA and the proposed CVS models correspondingly. $Apr_{(BCC-CCR)_i}, Apr_{(SFA)_i}, and\ Apr_{(CVS)_i}$ shows the average profit risk of i^{th}bank for CCR-BCC , SFA and CVS models correspondingly. $\beta_{BCC-CCR(0)}, \beta_{SFA(0)}, and\ \beta_{(CVS)0}$ are intercept or the constant term or the slope parameter of CCR-BCC , SFA and CVS models respectively.

$\beta_{(CCR-BCC)_i}$, $\beta_{(SFA)_i}$, and $\beta_{(CVS)_i}$ are orderly derivatives of CCR-BCC, SFA and CVS models correspondingly. Finally $\varepsilon_{(BCC-CCR)_i}, \varepsilon_{(SFA)_i}, and\ \varepsilon_{(CVS)_i}$ signifies the error terms of CCR-BCC , SFA and CVS models respectively.

5.5. ULFR model

Assume that $APR_{(CCR-BCC)_i}$, $APR_{(SFA)_i}, and\ APR_{(CVS)_i}$ are linearly related unobservable variables associated with $EF(CCR - BCC)_i$, $EF(SFA)_i, and\ EF(CVS)_i$ respectively. So, the three functional forms of CCR-BCC , SFA and CVS are:

$$\varphi_{(BCC-CCR)_i} = EF(CCR - BCC)_i = \beta_{(CCR-BCC)_\alpha} + \beta_{(CCR-BCC)_f} APR_{(CCR-BCC)_i} \quad \text{Eq. 16}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

$$\varphi_{(SFA)_i} = EF(SFA)_i = \beta_{(SFA)_\alpha} + \beta_{(SFA)_f} APR_{(SFA)_i} \quad \textit{Eq. 17}$$

$$\varphi_{(CVS)_i} = EF(CVS)_i = \beta_{(CVS)_\alpha} + \beta_{(CVS)_f} APR_{(CVS)_i} \quad \textit{Eq. 18}$$

and the two equivalent random variables for CCR-BCC , SFA and CVS are detected with errors $d_{(BCC-CCR)_i}, d_{(SFA)_i}, d_{(CVS)_i}$ and $e_{CCR-BCC_i}, e_{(SFA)_i}, e_{(CVS)_i}$, (i = 1,2, … . n) correspondingly as:

$$Ef(CCR-BCC)_i = EF_{(CCR-BCC)_i} + d_{(CCR-BCC)_i}$$
$$Apr_{(CCR-BCC)_i} = APR_{(CCR-BCC)_i} + e_{(CCR-BCC)_i} \quad \textit{Eq. 19}$$

$$Ef(SFA)_i = EF_{(SFA)_i} + d_{(SFA)_i}$$
$$Apr_{(SFA)_i} = APR_{(SFA)_i} + e_{(SFA)_i} \quad \textit{Eq. 20}$$

$$Ef(CVS)_i = EF_{(CVS)_i} + d_{(CVS)_i}$$
$$Apr_{(CVS)_i} = APR_{(CVS)_i} + e_{(CVS)_i} \quad \textit{Eq. 21}$$

Equations number 16, 17, 18 and 19, 20, 21 represent ULFR model in which $EF(CCR-BCC)_i, EF(SFA)_i, EF(CVS)_i$ and $APR_{(CCR-BCC)_i}$, $APR_{(SFA)_i}, APR_{(CVS)_i}$ are only the two variables of CCR-BCC , SFA, CVS models respectively and there are only one relation between $EF(CCR-BCC)_i$ with $APR_{(CCR-BCC)_i}$, $EF(SFA)_i$ with $APR_{(SFA)_i}$ and , $EF(CVS)_i$ with $APR_{(CVS)_i}$. Finally, $d_{(CCR-BCC)_i}, d_{(SFA)_i}, d_{(CVS)_i}$ and $e_{(CCR-BCC)_i}, e_{(SF\ )_i}, e_{(CVS)_i}$ are random variables of CCR-BCC , SFA, CVS correspondingly, which are mutually independent and normally distributed.

To summarize our calculation, CVS proposed model is considered. CCR-BCC and SFA have the same evaluation and they have not considered in the following equations. The following conditions are considered for CVS proposed model:

$$E(d_{(CVS)_i}) = E(e_{(CVS)_i}, \mathrm{Var}(d_{(CVS)_i}) = \sigma^2_{d_{(CVS)}}, \mathrm{Var}(e_{(CVS)_i}) = \sigma^2_{e_{(CVS)}} \quad \forall i$$

$$\mathrm{Cov}(d_{(CVS)_i}, d_{(CVS)_j}) = \mathrm{Cov}(e_{(CVS)_i}, e_{(CVS)_j}) = 0, \quad i \neq j \quad \textit{Eq. 22}$$

$$\mathrm{Cov}(d_{(CVS)_i}, e_{(CVS)_j}) = 0 \qquad \forall i, j$$

The ratio of error variance in CVS proposed model is recognized as:

$$\frac{\sigma^2_{e_{(CVS)}}}{\sigma^2_{d_{(CVS)}}} = \lambda \quad \textit{Eq. 23}$$

Consider the following 24, 25, 26, 27 and 28 assumptions in CVS proposed model before introducing the equations 29, 30, 31 and 32:

$$\overline{Ef(CVS)} = \frac{\sum Ef(CVS)_i}{n} \quad \textit{Eq. 24}$$

$$\overline{Pr(CVS)} = \frac{\sum Pr(CVS)_i}{n} \quad \textit{Eq. 25}$$

$$S_{yy} = \sum (Ef(CVS)_i - \overline{Ef(CVS)})^2 \quad \textit{Eq. 26}$$

$$S_{xx} = \sum (Pr(CVS)_i - \overline{Pr(CVS)})^2 \quad \textit{Eq. 27}$$

$$S_{xy} = \sum (Ef(CVS)_i - \overline{Ef(CVS)})(Pr(CVS)_i - \overline{Pr(CVS)}) \quad \textit{Eq. 28}$$

Based on the 24, 25, 26, 27 and 28 assumptions in CVS proposed model, maximum possibility evaluator of parameters in CVS proposed model are introduced in equations 29, 30, 31 and 32:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

$$\hat{\beta}_{(CVS)_\alpha} = \overline{Ef(CVS)} - \beta_{(CVS)_f}\overline{Apr_{(CVS)}} \quad \text{Eq. 29}$$

$$\hat{\beta}_{(CVS)_f} = \frac{(S_{yy}+\lambda S_{xx})+\left\{(S_{yy}+\lambda S_{xx})^2+4\lambda S_{xy}^2\right\}^{\frac{1}{2}}}{2S_{xy}} \quad \text{Eq. 30}$$

Eq. 31

$$APR_{(CVS)_i} = \frac{\lambda Apr_{(CVS)_i} + \hat{\beta}_{(CVS)_f}(Ef(CVS)_i - \hat{\beta}_{(CVS)_\alpha})}{\lambda + \hat{\beta}_{(CVS)_f}} \quad \text{Eq. 32}$$

The sum of squared distances of the detected parts from the close-fitting line or the residual sum of squares (S_E) in CVS proposed model is specified as:

$$SS_E = \frac{\sum\left\{Ef(CVS)_i - (\hat{\beta}_{(CVS)_\alpha} + \hat{\beta}_{(CVS)_f}Pr(CVS)_i)\right\}^2}{(\lambda + \hat{\beta}_{(CVS)_f}{}^2)} = \frac{S_{yy} - 2\hat{\beta}_{(CVS)_f}S_{xy} + \hat{\beta}_{(CVS)_f}{}^2 S_{xx}}{(\lambda + \hat{\beta}_{(CVS)_f}{}^2)} \quad \text{Eq. 33}$$

Consider the ratio of error variance in CVS proposed model is equal to one ($\lambda = 1$). For special cases in which $\lambda \neq 1$, it should be reduced with the case of $\lambda = 1$ by dividing the detected amounts of $Ef(\text{CAS})$ by $\lambda^{\frac{1}{2}}$. So, the following model is specified:

$$SS_E = \frac{S_{yy} - 2\hat{\beta}_{(CVS)_f}S_{xy} + \hat{\beta}_{(CVS)_f}{}^2 S_{xx}}{(1 + \hat{\beta}_{(CVS)_f}{}^2)} \quad \text{Eq. 34}$$

Finally, the coefficient in determination of ULFR, for free value of λ in CVS proposed model is defined as:

$$R^2_{(CVS)_f} = \frac{SS_R}{S_{yy}} \quad \text{Eq. 35}$$

SS_R is the regression sum of squares in CVS proposed model and can be represented as the following relation:

$$SS_R = S_{yy} - SS_E = S_{yy} - \frac{S_{yy}-2\hat{\beta}_{(CVS)_f}S_{xy}+\hat{\beta}_{(CVS)_f}{}^2 S_{xx}}{(1+\hat{\beta}_{(CVS)_f}{}^2)} = \frac{\hat{\beta}_{(CVS)_f}{}^2 S_{yy}+2\hat{\beta}_{(CVS)_f}S_{xy}-\hat{\beta}_{(CVS)_f}{}^2 S_{xx}}{(1+\hat{\beta}_{(CVS)_f}{}^2)} = \frac{\hat{\beta}_{(CVS)_f}{}^2(S_{yy}-S_{xx})+2\hat{\beta}_{(CVS)_f}S_{xy}}{(1+\hat{\beta}_{(CVS)_f}{}^2)} \quad \text{Eq. 36}$$

6. Discussion and evaluation

Technical efficiency evaluation for the three suggested models are presented at the first step.

6.1. Technical efficiency evaluation resulting from CCR-BCC, SFA and CVS proposed models

Based on the nature of input-oriented model, a bank is technical efficient if it can be reduced inputs in providing the specified outputs. Staying at the best frontier line shows the efficiency score of one. Technical efficiency evaluation of CCR-BCC, SFA and a novel CVS models through are presented in the following three sub-sections.

6.1.1. Technical efficiency evaluation of CCR-BCC model

Figure 3 shows the technical efficiency scores for 65 banks in CCR-BCC model. Based on the given information, the average technical efficiency scores of banks is 0.8599. This shows that the banks have an inefficiency score of 14.01% in using the existing resources. Banks numbers 5, 11, 16, 41, 56 and 64 with the efficiency scores of 1 are efficient. Banks number 51, 52 and 57 are the least efficient banks with the efficiency scores of 0.511, 0.593 and 0.582, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

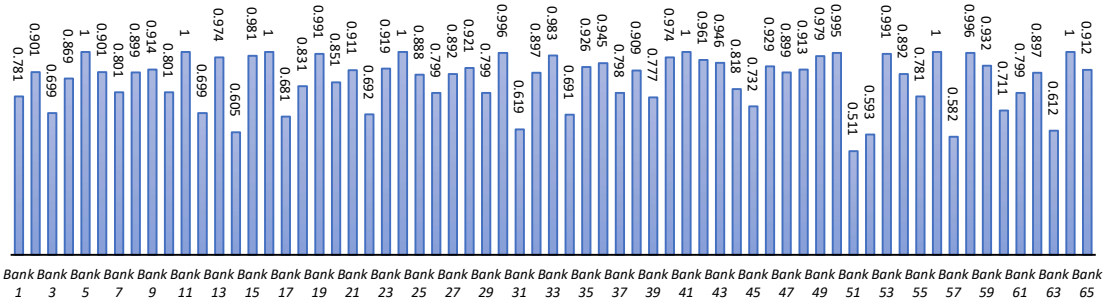

Figure 3. Technical efficiency scores for 65 banks in CCR-BCC model.

6.1.2. Technical efficiency evaluation of SFA model

Figure 4 shows the technical efficiency scores for 65 banks in SFA model. Based on the given information, the average technical efficiency scores of banks is 0.8462. This shows that the banks have an inefficiency score of 15.38% in using the existing resources. Bank number 50 with the efficiency scores of 0.996 is the most efficient bank. Banks number 51, 52 and 57 are the least efficient banks with the efficiency scores of 0.525, 0.581 and 0.588 respectively.

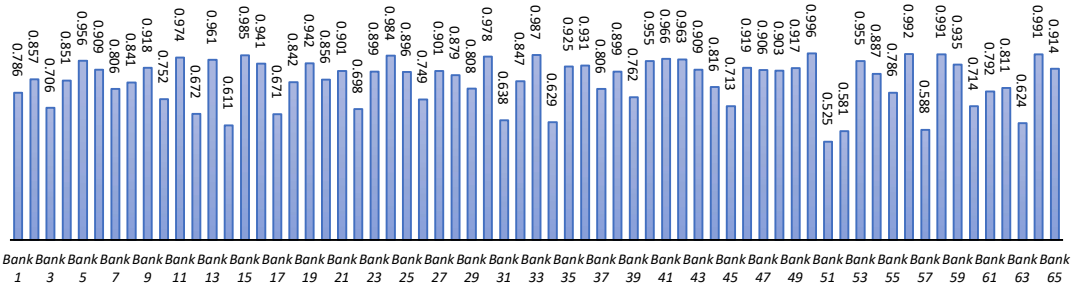

Figure 4. Technical efficiency scores for 65 banks in SFA model.

6.1.3. Technical efficiency evaluation of CVS model

Figure 5 shows the technical efficiency scores for 65 banks in CVS model. Based on the given information, the average technical efficiency scores of banks is 0.85305. This shows that the banks have inefficiency score of 14.69% in using the existing resources. Banks number 50 and 64 with the efficiency scores of 0.995 is the most efficient banks. Banks number 51, 52 and 57 are the least efficient banks with the efficiency scores of 0.518, 0.587 and 0.585, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

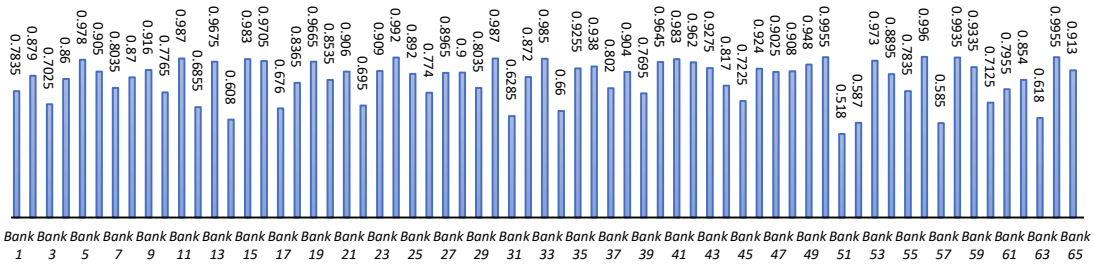

Figure 5. Technical efficiency scores for 65 banks in CVS proposed model.

Based on the information provided in figure 2, 3 and 4, bank number 51, 52 and 57 have the lowest efficiency scores in all suggested models. In fact, for all CCR-BCC, SFA, and CVS models these three banks have more than 40.0 % inefficiency score using the existing resources.

6.2. A comparison of average technical efficiency of CCR-BCC , SFA and CVS models

Figure 5 shows the average technical efficiency scores for 59 banks in CCR-BCC , SFA and CVS model. Based on the information provided in previous subsections, the following relation is presented:

$$ATE_{CCR-BCC}(0.8599) \geq ATE_{CVS}(0.85305) \geq ATE_{SFA}(0.8462) \qquad Eq.\ 37$$

It is obvious that $ATE_{BCC-CCR}$ (Average Technical Efficiency of CCR-BCC model) is more than ATE_{CVS}(Average Technical Efficiency of CVS model) and ATE_{CVS} is more than ATE_{SFA}(Average Technical Efficiency of SFA model). Unlike SFA model, CCR-BCC or DEA models cannot measure statistical noise. Moreover, SFA lets DMUs to depart from efficiency frontier line because of statistical noise and inefficiency scores. Finally, among the three suggested models, CCR-BCC presents the similar efficiency scores compare with the two other models.

6.3. Evaluation of regression and ULFR

Table 5 shows the statistical evaluation for the three suggested models in SPSS and SAS sources.

Table 5. Statistical evaluation of CCR-BCC, SFA and CVS models.

Model	Coefficients (β_1)	Coefficient Determination of simple linear regression (R^2)	Coefficient Determination of ULFR (R_f^2)	P-value
CCR-BCC	0.4799	0.2287	0.9971	0.0141
SFA	0.2996	0.0825	0.9956	0.1914
CVS	0.5649	0.2999	0.9999	0.0018

Based on the information provided in table 7, after applying the statistical evaluation under 5% significant for CCR-BCC and CVS proposed models, the p-value is less than 5% (0.0141 and 0.0018 for CCR-BCC and CVS models respectively) and this shows that the relation among efficiency and profit risk is statistically noteworthy. Moreover, based on the aforementioned low amount of p-value, profit risk has a positive effect on financial procedure of the banks and provide more appropriate conditions for banks. In addition, the managers of banks can get more profits and they face lower challenge of wasting a large amount of profits. Alternatively, the p-value for the SFA model which is 0.1914 (is more than 0.05) and this shows the weak relationship

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

among SFA and profit risk in the current study. In addition, with low coefficient (β_1) amount of 0.2996 compare with two other suggested models which are 0.4799 and 0.5649 for CCR-BCC and CVS respectively, it has the lowest efficiency growth. In fact, the 1% growth in profit risk of SFA model provide the growing in the efficiency by only 0.29%. On the other hand, the 1% growth in profit risk of CVS model provide the growing in the efficiency by 0.56%. Thus, this shows that the efficiency growth of the suggested CVS model is approximately twice the efficiency growth of SFA model. Finally, the coefficient of determination of simple linear regression (R^2) and the coefficient of determination for ULFR (R_f^2) for CVS suggested, is more and better than the two other models. This shows the strong connection between profit risk and coefficient in the suggested CVS model compare with the two other models. As a final point, the coefficient of determination for ULFR (R_f^2) in the three suggested models (0.9971, 0.9956, 0.9999 for CCR-BCC , SFA and CVS respectively) are significantly more than (0.2287, 0.0825 and 0.2999 for CCR-BCC, SFA and CVS correspondingly) and this demonstrates the fact that ULFR plays a more important role compare with linear regression.

6.4. Efficiency evaluation of CVS proposed model after applying ULFR error-free method

After combining the CCR-BCC and SFA models, the efficiency amount of the proposed CVS model may have some errors as a result of some unrelated data and missing values. In order to remove some errors and missing values, ULFR has been applied. Table 6 shows error-free efficiency amount of the CVS proposed model before and after applying ULFR model. Moreover, the final ranking of the banks after applying error-free ULFR method are presented in table 6. After considering the error free ULFR method, some banks receive the higher efficiency scores and some banks obtain the lower efficiency scores. Finally, bank number 56 by obtaining the extra efficiency amount of 0.03% and increasing from 0.996 to 0.999 efficiency score, is introduced as the highest efficient bank. On the other hand, bank number 51 by obtaining the subtractive efficiency amount of 0.15% and decreasing from 0.518 to 0.503 efficiency score, is introduced as the lowest efficient bank.

Table 6. Efficiency evaluation of CVS proposed model before and after applying ULFR error-free method.

Banks	Technical efficiency score before applying ULFR error free method	Minimal error free efficiency after applying ULFR	Ranking of banks after applying ULFR	Banks	Technical efficiency score before applying ULFR error free method	Minimal error free efficiency after applying ULFR	Ranking of banks after applying ULFR
1	0.783	0.789	48	34	0.660	0.681	57
2	0.879	0.878	36	35	0.925	0.921	22
3	0.702	0.741	53	36	0.938	0.946	19
4	0.860	0.868	39	37	0.802	0.798	46
5	0.978	0.979	11	38	0.904	0.895	29
6	0.905	0.892	31	39	0.769	0.763	52
7	0.803	0.814	43	40	0.964	0.959	15
8	0.870	0.875	37	41	0.983	0.982	9
9	0.916	0.915	24	42	0.962	0.948	18
10	0.776	0.781	49	43	0.927	0.929	21

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

11	0.987	0.990	5	44	0.817	0.831	42
12	0.685	0.699	56	45	0.722	0.701	55
13	0.967	0.968	13	46	0.924	0.918	23
14	0.608	0.632	61	47	0.902	0.898	28
15	0.983	0.980	10	48	0.908	0.904	27
16	0.970	0.975	12	49	0.948	0.950	17
17	0.676	0.661	58	50	0.995	0.989	6
18	0.836	0.851	40	51	0.518	0.503	65
19	0.966	0.964	14	52	0.587	0.555	64
20	0.853	0.848	41	53	0.973	0.952	16
21	0.906	0.911	25	54	0.889	0.880	35
22	0.695	0.652	59	55	0.783	0.771	51
23	0.909	0.906	26	56	0.996	0.999	1
24	0.992	0.991	4	57	0.585	0.596	63
25	0.892	0.885	33	58	0.993	0.995	3
26	0.774	0.779	50	59	0.933	0.936	20
27	0.896	0.882	34	60	0.712	0.722	54
28	0.900	0.889	32	61	0.795	0.796	47
29	0.803	0.809	44	62	0.854	0.801	45
30	0.987	0.985	7	63	0.618	0.606	62
31	0.628	0.641	60	64	0.995	0.998	2
32	0.872	0.874	38	65	0.913	0.894	30
33	0.985	0.984	8				

6.5. Efficiency evaluation of CVS proposed model after applying ULFR error-free method

Filtering in WEKA environment is applied. In filter selection, there are two classifications of filters: Supervised and unsupervised. In both groups, filters are for attributes and instances distinctly. Utilizing attribute and instance filters, all attributes and instances can be increased, excluded, and improved. In this study, a special pre-processing is applied by experts. In fact, the best filtering approach for each category has been selected. Four layers filtering Pre-process is applied to the dataset to make imbalanced data balanced. This process is applied in four steps as follows:

Step A: Discretization (**unsupervised attribute filter):** Consistent with the points, for a systematic plan, this step should be done. Discretization adapts one form of data to another form. There are many techniques used to explain these two data types, such as 'quantitative' vs. 'qualitative', 'continuous' vs. 'discrete', 'ordinal' vs. 'nominal', and 'numeric' vs. 'categorical'. It is so valuable to choose the most fitting method for discretization. In this paper, data is classified in to quantitative or qualitative.

Step B: Attribute Selection (supervised attribute filter): With the aim of choosing the best attributes for defining the best scenario, this step can be applied.

Step C: Resample filters (Unsupervised instances): Creates a random subsample of a dataset applying sampling with a substitute. The original dataset must fit totally in memory. The number of instances in the created dataset may be stipulated.

Step D: Stratified Remove Folds (supervised instance filter): In our specific data set, this filter plays a vital role in improving the accuracy of all algorithms. This filter takes a data set and outputs a recognized fold for cross-validation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 September 2020 doi:10.20944/preprints202009.0729.v1

The classes to clusters evaluation in WEKA using the primary evaluator to each output of steps defined above. It is applied because it is the individual clustering and classification surveyor which generates numeric accuracy as a principle of evaluation within numerous algorithms. The following formula shows the final numerical accuracy:

$$Accuracy = \frac{Truepositive + Truenegative}{Truepositive + Truenegative + Falsepositive + Falsenegative} \quad Eq.38$$

K-means and Make Density Based Cluster in clustering unsupervised algorithms as well as Sequential Minimal Optimization (SMO) and Naïve Bayes in classification supervised algorithms have been applied to find the superior algorithm

K-means is a method of cluster analysis which drives to split N descriptions into K clusters where each reflection fits into the cluster with the nearest mean. Originally, the k centroid prerequisite to be designated at the beginning. These centers should be designated productively, employing them as much as possible secluded from each other since an assorted place causes the various outcome.

Make Density Based Cluster finds numerous clusters beginning from the expectable density scattering of corresponding nodes. It is one of the most communal clustering algorithms and stated in the technical literature. Supposed a set of themes in some space, it clusters together items that are methodically packed together (items with various adjacent neighbors), pattern as outliers' themes that lie singlehanded in low-thickness areas.

SMO is an upgraded method for training Support vector machines (SVM) which accepted good performance in many problems. Nonetheless, the usage of SVM was partial because of application difficulties and training intricacy. Thus, SMO is silently upgraded by being theoretically simple, easy to implement, and in general cases faster than SVM.

Naive Bayes is a well-known classifier that contains conditional probabilities. So, it applies Bayes' theorem and it assumes that structures have a solid individuality among each other. Additionally, it has many advantages such as straightforwardness of use, quick union, and high scalability. In conclusion, Naïve Bayes needs less training data for building a model.

After applying steps (Step A, Step B, Step C and Step D), the accuracy and average accuracy in each stage are presented in Tables 7-10.

Table 7. Accuracy comparison contained by K-means algorithm (All Numbers Are in Percent).

Model	Step A	Step B	Step C	Step D
CCR-BCC	86.69	88.71	90.65	92.73
SFA	82.67	83.18	84.25	86.16
CVS	90.34	92.16	93.27	97.93

Table 8. Accuracy comparison contained by Make Density based Cluster algorithm (All Numbers Are in Percent).

Model	Step A	Step B	Step C	Step D
CCR-BCC	73.78	75.29	77.18	79.78
SFA	68.29	70.08	72.99	74.19
CVS	80.56	81.29	82.64	84.11

Table 9. Accuracy comparison contained by SMO algorithm (All Numbers Are in Percent).

Model	Step A	Step B	Step C	Step D
CCR-BCC	88.96	90.67	93.95	95.35
SFA	85.58	87.36	89.45	90.59

CVS	92.58	95.39	97.59	98.99

Table 10. Accuracy comparison contained by Naive Bayes algorithm (All Numbers Are in Percent).

Model	Step A	Step B	Step C	Step D
CCR-BCC	81.47	83.16	85.99	88.93
SFA	76.25	78.27	80.39	83.78
CVS	84.39	86.11	88.98	90.03

It can be concluded from Tables 7–10 as the layers of filtering rises:

- The maximum of accuracy within four valuation approaches are increased.
- The average accuracy within four models, links to each filtering step is increased.
- The accuracy of all algorithms is augmented as well.

Finally, the two following relations for four suggested algorithms and for three suggested models are applicable:

$$\text{SMO} \geq \text{K} - \text{means} \geq \text{Naive Bayes} \geq \text{Make Density based Cluster algorithm} \qquad Eq.\,39$$

$$\text{CVS} \geq \text{CCR} - \text{BCC} \geq \text{SFA} \qquad Eq.\,40$$

CVS at Steps A–D has the highest accuracy. In fact, according to our unique data, attributed, and instances using the K-means based on CVS model in proposed combining optimization and data mining methodology has the best performance.

7. Conclusion

To introduce the most efficient model, algorithm and bank, this paper provides a scientific assessment of three CCR-BCC, SFA and a novel combination of the CCR-BCC and SFA models or CVS at the first step. At data mining or the second step, a novel four-layers data mining filtering pre-processes for selected supervised classification as well as unsupervised clustering algorithms to increase the accuracy and to remove unrelated attributes and data are applied. In the optimization part, the average technical efficiency score of SFA was lower than CVS and the average technical efficiency score of CVSs was lower than CCR-BCC model. The average technical efficiency score of CCR-BCC model indeed gets the highest score, but DEA non-parametric models unlike SFA parametric model have not measured statistical noise. So, to introduce the most efficient model and bank the positive or negative correlation between profit risk and efficiency score based on statistical analysis has been proposed and it would be the most reliable approach for evaluating and ranking models and banks. The coefficient determination of ULFR and simple linear regression show that CVS model has the most positive correlation between the profit risk and efficiency score compare with other suggested models. Finally, to remove unrelated, noisy, missing data, and introduce the most and the least efficient bank ULFR error-free method has been applied. After applying the ULFR error-free method some banks received higher efficiency scores and some hospitals obtained lower efficiency scores. Bank number 58 get the highest and the best efficiency score and hospital number 51 obtained the lowest efficiency score. Besides, the highest coefficients, coefficient determination of simple linear regression, and coefficient determination of ULFR and the lowest p-value among the other suggested models show that CVS proposed model should be the most appropriate approach. In the data mining part, SMO algorithm receives the highest accuracy in all four suggested filtering layers. In the future, a comparison of combining other well-known nonparametric operation research approaches in DEA as well as parametric operation research models such as the deterministic frontier approach (DFA) and thick frontier approach (TFA) will be helpful. In addition, based on the nature of the data, the best preprocessing approach should apply by experts. So, other preprocessing approaches may have a more positive effect on accuracy.

References

1. Aghapour, A.H.; Yazdani, M.; Jolai, F.; Mojtahedi, M. Capacity planning and reconfiguration for disaster-resilient health infrastructure. *Journal of Building Engineering* **2019**, *26*, 100853.
2. Chen, Y.; Yazdani, M.; Mojtahedi, M.; Newton, S. The impact on neighbourhood residential property valuations of a newly proposed public transport project: The Sydney Northwest Metro case study. *Transportation Research Interdisciplinary Perspectives* **2019**, *3*, 100070.
3. Khalili, S.M.; Babagolzadeh, M.; Yazdani, M.; Saberi, M.; Chang, E. A Bi-objective Model for Relief Supply Location in Post-Disaster Management. In Proceedings of 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS); pp. 428-434.
4. Saad, N.M.; Idris, N.H.; Edzalina, N. Efficiency of life insurance companies in Malaysia and Brunei: a comparative analysis. *International Journal of Humanities and Social Science* **2011**, *1*, 111-122.
5. Hubbard, R.G.; Hubbard, R.G. *Money, the financial system, and the economy*; Addison-Wesley Reading, MA: 1994.
6. Adrienn, H. FINANCING ASPECTS OF THE HUNGARIAN GENERAL MANUFACTURERS IN 2010-2012. *Annals of the University of Oradea, Economic Science Series* **2014**, *23*.
7. Patrícia, B.-N.; Balázs, F. Returns of Private Equity–Comparative analyses of the Returns of Venture capital and Buyout funds in Europe and in the US. *THE ANNALS OF THE UNIVERSITY OF ORADEA* **2014**, 818.
8. Fenyves, V.; Tarnóczi, T.; Zsidó, K. Financial Performance Evaluation of agricultural enterprises with DEA Method. *Procedia Economics and Finance* **2015**, *32*, 423-431.
9. Kumbhakar, S.C.; Lovell, C.K. *Stochastic frontier analysis*; Cambridge university press: 2003.
10. Coelli, T.J.; Rao, D.S.P.; O'Donnell, C.J.; Battese, G.E. *An introduction to efficiency and productivity analysis*; Springer Science & Business Media: 2005.
11. Dong, Y.; Hamilton, R.; Tippett, M. Cost efficiency of the Chinese banking sector: A comparison of stochastic frontier analysis and data envelopment analysis. *Economic Modelling* **2014**, *36*, 298-308.
12. Mirmozaffari, M.; Yazdani, M.; Boskabadi, A.; Dolatsara, H.A.; Kabirifar, K.; Golilarz, N.A. A Novel Machine Learning Approach Combined with Optimization Models for Eco-efficiency Evaluation. *Applied Sciences* **2020**, *10*, 5210.
13. Mirmozaffari, M.; Alinezhad, A.; Gilanpour, A. Data Mining Apriori Algorithm for Heart Disease Prediction. *Int'l Journal of Computing, Communications & Instrumentation Engg* **2017**, *4*, 20-23.
14. Mirmozaffari, M. Presenting a Medical Expert System for Diagnosis and Treatment of Nephrolithiasis. *EJMED. May* **2019**, *1*, 1.
15. Mirmozaffari, M. Eco-Efficiency Evaluation in Two-Stage Network Structure: Case Study: Cement Companies. *Iranian Journal of Optimization* **2019**, *11*, 125-135.
16. Huang, T.H.; Wang, M.H. Comparison of economic efficiency estimation methods: Parametric and non–parametric techniques. *The Manchester School* **2002**, *70*, 682-709.
17. Casu, B.; Girardone, C.; Molyneux, P. Productivity change in European banking: A comparison of parametric and non-parametric approaches. *Journal of Banking & Finance* **2004**, *28*, 2521-2540.
18. Pasiouras, F.; Delis, M.D.; Papanikolaou, N.I. Determinants of bank efficiency: evidence from a semi - parametric methodology. *Managerial finance* **2009**.

19. Weill, L. Measuring cost efficiency in European banking: A comparison of frontier techniques. *Journal of Productivity Analysis* **2004**, *21*, 133-152.
20. Fernandes, F.D.S.; Stasinakis, C.; Bardarova, V. Two-stage DEA-Truncated Regression: Application in banking efficiency and financial development. *Expert Systems with Applications* **2018**, *96*, 284-301.
21. Altunbas, Y.; Carbo, S.; Gardener, E.P.; Molyneux, P. Examining the relationships between capital, risk and efficiency in European banking. *European financial management* **2007**, *13*, 49-70.
22. Sprent, P. Some history of functional and structural relationships. *Contemporary Mathematics* **1990**, *112*, 3-15.
23. Zhao, X.; Zhang, X.; Cai, Z.; Tian, X.; Wang, X.; Huang, Y.; Chen, H.; Hu, L. Chaos enhanced grey wolf optimization wrapped ELM for diagnosis of paraquat-poisoned patients. *Computational biology and chemistry* **2019**, *78*, 481-490.
24. Zhao, X.; Li, D.; Yang, B.; Ma, C.; Zhu, Y.; Chen, H. Feature selection based on improved ant colony optimization for online detection of foreign fiber in cotton. *Applied Soft Computing* **2014**, *24*, 585-596.
25. Yazdani, M.; Khalili, S.M.; Jolai, F. A parallel machine scheduling problem with two-agent and tool change activities: an efficient hybrid metaheuristic algorithm. *International Journal of Computer Integrated Manufacturing* **2016**, *29*, 1075-1088.
26. Yazdani, M.; Jolai, F.; Taleghani, M.; Yazdani, R. A modified imperialist competitive algorithm for a two-agent single-machine scheduling under periodic maintenance consideration. *International Journal of Operational Research* **2018**, *32*, 127-155.
27. Yazdani, M.; Aleti, A.; Khalili, S.M.; Jolai, F. Optimizing the sum of maximum earliness and tardiness of the job shop scheduling problem. *Computers & Industrial Engineering* **2017**, *107*, 12-24.
28. Xu, Y.; Chen, H.; Luo, J.; Zhang, Q.; Jiao, S.; Zhang, X. Enhanced Moth-flame optimizer with mutation strategy for global optimization. *Information Sciences* **2019**, *492*, 181-203.
29. Shahmansouri, A.A.; Yazdani, M.; Ghanbari, S.; Bengar, H.A.; Jafari, A.; Ghatte, H.F. Artificial neural network model to predict the compressive strength of eco-friendly geopolymer concrete incorporating silica fume and natural zeolite. *Journal of Cleaner Production* **2020**, 123697.
30. Xu, X.; Chen, H.-l. Adaptive computational chemotaxis based on field in bacterial foraging optimization. *Soft Computing* **2014**, *18*, 797-807.
31. Yazdani, M.; Jolai, F. Lion optimization algorithm (LOA): a nature-inspired metaheuristic algorithm. *Journal of computational design and engineering* **2016**, *3*, 24-36.
32. Azadeh, A.; Seif, J.; Sheikhalishahi, M.; Yazdani, M. An integrated support vector regression–imperialist competitive algorithm for reliability estimation of a shearing machine. *International Journal of Computer Integrated Manufacturing* **2016**, *29*, 16-24.
33. Yazdani, M.; Kabirifar, K.; Frimpong, B.E.; Shariati, M.; Mirmozaffari, M.; Boskabadi, A. Improving Construction and Demolition Waste Collection Service in an Urban Area Using a Simheuristic Approach: A Case Study in Sydney, Australia. *Journal of Cleaner Production* **2020**, 124138.
34. Yazdani, M.; Ghodsi, R. Invasive weed optimization algorithm for minimizing total weighted earliness and tardiness penalties on a single machine under aging effect. *International Robotics and Automation Journal* **2017**, *2*, 1-5.
35. Yazdani, M.; Jolai, F. A Genetic Algorithm with Modified Crossover Operator for a Two-Agent Scheduling Problem. **2015**.

36. Yazdani, M.; Mojtahedi, M.; Loosemore, M. Enhancing evacuation response to extreme weather disasters using public transportation systems: a novel simheuristic approach. *Journal of Computational Design and Engineering* **2020**, *7*, 195-210.
37. Shahmansouri, A.A.; Akbarzadeh Bengar, H.; Jahani, E. Predicting compressive strength and electrical resistivity of eco-friendly concrete containing natural zeolite via GEP algorithm. *Construction and Building Materials* **2019**, *229*, 116883, doi:https://doi.org/10.1016/j.conbuildmat.2019.116883.
38. Mirmozaffari, M.; Alinezhad, A.; Gilanpour, A. Data Mining Classification Algorithms for Heart Disease Prediction. *Int'l Journal of Computing, Communications & Instrumentation Engg* **2017**, *4*, 11-15.
39. Ozekes, A.; Camurcu, Y. Classification and prediction in a data mining application. *Journal of Marmara for pure and applied sciences* **2002**, *18*, 157-172.
40. Albayrak, A.S.; Yilmaz, S.K. Data mining: decision tree algorithms and an application on ISE data. *Suleyman Demirel University the Journal of Faculty of Economics and Administrative Sciences* **2009**, *14*, 31-52.
41. Aşan, Z. Examining the socioeconomic characteristics of customers using credit cards, with clustering analysis. *Dumlupınar University The Journal of Social Sciences* **2007**, *17*, 256-267.
42. Golilarz, N.A.; Demirel, H.; Gao, H. Adaptive generalized Gaussian distribution oriented thresholding function for image de-noising. *International Journal of Advanced Computer Science and Applications* **2019**, *10*, 10-15.
43. Golilarz, N.A.; Gao, H.; Kumar, R.; Ali, L.; Fu, Y.; Li, C. Adaptive wavelet based MRI brain image de-noising. *Frontiers in Neuroscience* **2020**, *14*.
44. Golilarz, N.A.; Mirmozaffari, M.; Gashteroodkhani, T.A.; Ali, L.; Dolatsara, H.A.; Boskabadi, A.; Yazdi, M. Optimized wavelet-based satellite image de-noising with multi-population differential evolution-assisted harris hawks optimization algorithm. *IEEE Access* **2020**, *8*, 133076-133085.
45. Yazdani, M.; Babagolzadeh, M.; Kazemitash, N.; Saberi, M. Reliability estimation using an integrated support vector regression–variable neighborhood search model. *Journal of Industrial Information Integration* **2019**, *15*, 103-110.
46. Yazdani, M.; Khalili, S.M.; Babagolzadeh, M.; Jolai, F. A single-machine scheduling problem with multiple unavailability constraints: A mathematical model and an enhanced variable neighborhood search approach. *Journal of Computational Design and Engineering* **2017**, *4*, 46-59.
47. Shahmansouri, A.A.; Akbarzadeh Bengar, H.; Ghanbari, S. Compressive strength prediction of eco-efficient GGBS-based geopolymer concrete using GEP method. *Journal of Building Engineering* **2020**, *31*, 101326, doi:https://doi.org/10.1016/j.jobe.2020.101326.
48. Akbarzadeh Bengar, H.; Shahmansouri, A.A.; Akkas Zangebari Sabet, N.; Kabirifar, K.; W.Y. Tam, V. Impact of elevated temperatures on the structural performance of recycled rubber concrete: Experimental and mathematical modeling. *Construction and Building Materials* **2020**, *255*, 119374, doi:https://doi.org/10.1016/j.conbuildmat.2020.119374.
49. Nematzadeh, M.; Shahmansouri, A.A.; Fakoor, M. Post-fire compressive strength of recycled PET aggregate concrete reinforced with steel fibers: Optimization and prediction via RSM and GEP. *Construction and Building Materials* **2020**, *252*, 119057, doi:https://doi.org/10.1016/j.conbuildmat.2020.119057.

50. Aranizadeh, A.; Niazazari, I.; Mirmozaffari, M. A novel optimal distributed generation planning in distribution network using cuckoo optimization algorithm. *European Journal of Electrical Engineering and Computer Science* **2019**, *3*.
51. Shen, L.; Chen, H.; Yu, Z.; Kang, W.; Zhang, B.; Li, H.; Yang, B.; Liu, D. Evolving support vector machines using fruit fly optimization for medical data classification. *Knowledge-Based Systems* **2016**, *96*, 61-75.
52. Wang, M.; Chen, H. Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. *Applied Soft Computing* **2020**, *88*, 105946.
53. Wang, M.; Chen, H.; Yang, B.; Zhao, X.; Hu, L.; Cai, Z.; Huang, H.; Tong, C. Toward an optimal kernel extreme learning machine using a chaotic moth-flame optimization strategy with applications in medical diagnoses. *Neurocomputing* **2017**, *267*, 69-84.
54. Chen, H.; Zhang, Q.; Luo, J.; Xu, Y.; Zhang, X. (2020). An enhanced Bacterial Foraging Optimization and its application for training kernel extreme learning machine. *Applied Soft Computing* **2020**, *86*, 105884.