Predicting Bank Financial Failures Using Discriminant Analysis, And Support Vector Machines Methods: A Comparative Analysis In Commercial Banks In Sudan (2006-2014)

Mohammed A. SirElkhatim, Naomie Salim

Abstract: Bank failures threaten the economic system as a whole. Therefore, predicting bank financial failures is crucial to prevent and/or lessen its negative effects on the economic system. Financial crises, affecting both emerging markets and advanced countries over the centuries, have severe economic consequences, but they can be hard to prevent and predict, identifying financial crises causes remains both science and art, said Stijn Claessens, assistant director of the International Monetary Fund. While it would be better to mitigate risks, financial crises will recur, often in waves and better crisis management is therefore important. Analyses of recurrent causes suggest that to prevent crises, governments should consider reforms in many underlying areas. That includes developing prudent fiscal and monetary policies, better regulating the financial sector, including reducing the problem of too-big-to-fail banks, and developing effective macro-prudential policies. Despite new regulations and better supervision, crises are likely to recur, in part because they can reflect deeper problems related to income inequality, the political economy and common human behavior. As such, improvements in crisis management are also needed. This is originally a classification problem to categorize banks as healthy or non-healthy ones. This study aims to apply Discriminant analysis and Support Vector Machines methods to the bank failure prediction problem in a Sudanese case, and to present a comprehensive computational comparison of the classification performances of the techniques tested. Eleven financial and non-financial ratios with six feature groups including capital adequacy, asset quality, Earning, and liquidity (CAMELS) are selected as predictor variables in the study. Credit risk also been evaluated using logistic analysis to study the effect of Islamic finance modes, sectors and payment types used by Sudanese banks with regard to their possibilities of failure. Experimental results are evaluated using accuracy of prediction. Features selection has shown that new groups can be identified from CAMELS ratios and narrowing the data set space to 11 factors instead of eighteen. Discriminant analysis has identified 3 ratios with highest predictive power which are: EAS (Ratio of equity capital to total asset), LADF (Ratio of liquid assets to deposits and short term funds) and RFR (Rain Fall Ratio), the later ratio is a novel one used for the first time by this research.

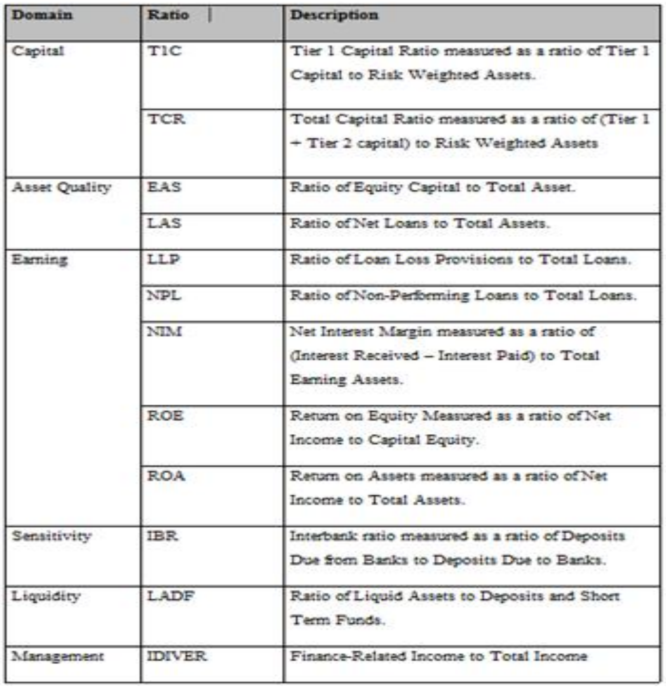
Index Terms: Bank Fail, Discriminant Analysis, Data Mining, Support Vector Machines, Sudan Bank System, Predicting Bank Distress

1 INTRODUCTION

Ddevelop of Sudan's gross domestic product (GDP) has been expected at 3.4 percent in 2014 and is predicted at 3.1 percent and 3.7 percent in 2015 and 2016, respectively. GDP should be driven by rain-fed agriculture, minerals and oil-transit fees with its neighbor country South of Sudan. Local services grew by 3.1 percent and accounted for 45.6 percent of GDP in 2014. Services, however, typically provide low-productivity jobs, so in general Sudanese economy has shown sharp decline in exports as a consequence to less production. Inflation in Sudan, the highest in Africa and averaging 36.9 percent in 2013-14, is due to exchange-rate devaluations, unsterilized gold purchases and supply disruptions owing to civil conflicts. Although it is projected to drop to 21.8 percent in 2015 on the back of a tight policy stance, build-ups of inflationary pressures will increase the already high rates of poverty and unemployment. In the short and medium term, growth will be sustained by the revitalization of agriculture and increased production of mineral and non-mineral exports, in addition to restraining inflation [9] A new IMF Staff-Monitored Program (SMP) and a five-year program of economic reform (FYPER, 2015-19) were adopted in 2014, aimed at enhancing macroeconomic stability and sustaining inclusive growth. Policy makers must nonetheless face challenges stemming from the structural weaknesses of the economy and limited market penetration. The slow growth of credit to the private sector due to low financial inter-mediation and the crowding-out effects of fiscal operations have further restrained the formalization of business and job creation. Refusal, since 2014, of foreign correspondent banks to process transfers to and from Sudan in order to avoid violating US sanctions has tightened the foreign-exchange market and raised the costs of imported inputs. In this respect, effective outreach is needed to remove the US sanctions. Additionally, Sudan's heavy external debt and volatile internal and external political environments could effectively weaken progress towards meeting the Millennium Development Goals (MDGs) According to African Economic Outlook (AEO 2016). The banking sector forms the backbone of Sudan's financial system and is the primary source of financing for the domestic economy. As of 2016, 37 commercial banks were active in the country, including 8 foreign banks and 22 others partially owned by foreign shareholders. However, public banks dominate the sector and account for around 50 percent of total banking sector assets. Bank deposits and credit to the private sector nearly doubled between 2005 and 2009, but still represented only 16 percent and 12 percent of GDP respectively by the end of the period. Systemic risk is estimated to be low, largely due to low levels of inter-mediation and the sector's small size and relative isolation from global financial markets [9] Bank failure occurs when a Bank is couldn't commit to perform its duties toward its customers specially depositors because it has become insolvent or too illiquid to meet its liabilities. To be more precise, a bank often fails economically when the market value of its liabilities becomes higher than its assets. The failed bank either ask other solvent bank to grant a loan or sells its assets at a lower price than its market value to gain a sufficient liquidity which can use it to pay on-demand deposits for it clients. If solvent bank for any reason declare its inability to grant the loan this creates a bank panic among the depositors as more of them try to take out cash deposits from the bank. As such, the bank is unable to meet the requests of all of its depositors on time. Also, a bank may be taken over by the regulating government agency which is typically central or reserve banks if Shareholders Equity is less than regulatory minimum. The failure of a bank is generally considered to be of more crucial than the collapse of other types of business firms because of the dependency and fragility of banking institutions. It is often feared that the bubble over effects of a failure of one bank can quickly spread throughout the economy and possibly lead to the failure of other firms there is no difference if those firms are solvent or not as the same time of those few numbers of panicked customers try to withdraw their deposits. Thereby, the bubble over effect of bank fear has a multiplier effect on all other firms that why reserve and central banks should have an effective policy in place to prevent and minimize those effects. Bank supervisory authorities have to maintain a reliable rating system for those financial institutions which can be used to classify them into different zones resulted in different policies targeted different zones. CAMELS used to rate the financial institution according to sex factors as name letters represents, each bank will have score on a scale, and a rating of one is considered the best and the rating of five is considered the worst for each factor. The acronym CAMELS stand for the following factors that examiners use to rate bank institutions: Capital Adequacy, Asset Quality, Management, Earnings, Liquidity and Sensitivity. The study of bank distress is an important issue. First, it enhances regulators' ability to predict potential crisis, and enables them to manage, coordinate and supervise more efficiently. Second, the early distinction between troubled and sound banks allows for appropriate actions to prevent failure and to protect healthy institutions. Third, the direct fiscal cost of recapitalizing and restructuring a troubled sector is high, and may amount to as large as half of the country's GDP. Fourth, the crisis in the financial sector may create other crisis, such as currency crisis, which may further weaken the economy, and aggravates the cost of distress. Finally, bank distress is accompanied with a credit crunch that leads to under-utilization and miss-allocation of funds, which may further hamper growth in the economy. For all five reasons, we are going to study how we could predict the commercial banks distress that occurs in the republic of Sudan. This Research tries to combine the usage of statistical and intelligent technique and to be the first research study the effect of using Discriminant Analysis (DA) and Support Vector Machines (SVM). DA is one of the most popular techniques used for analyzing financial distress [11]. This method assesses the predictive ability of several financial ratios. (Wilson and Jones, 1987) described this method as a technique which assigns a Z score to each company in a sample by using a combination of independent variables. The main advantage of this approach is its ability to reduce a multidimensional problem to a single score and provide a high level of accuracy. The DA approach has been used to develop a number of prediction models, including (Altman, 1968), (Altman, Haldeman, and Narayanan, 1977) [4]. Memic [10] has pointed out that DA outperformed its peers. Although most of researches are recommending Neural Network as sole intelligent techniques to be used at bankruptcy prediction tasks but this research tries to prove that Support vector machines (SVM) is shows very good learning and prediction capabilities, which makes it an efficient tool to deal with uncertainties encountered in this venture. The significance of the proposed model is the ability to predict the financial strength of banks at any future time. The implementation of SVM model is less complicated than that of sophisticated identification and optimization procedures. SVM has automated identification algorithm and easier design and compared to neural networks it has less number of parameters and faster adaptation. Neural networks take time to learn but once trained, they can offer an equitable performance because they are model independent and flexible enough to adapt any functional forms. Moreover, this model adopts a neural network design that minimize over-fitting. Over-fitting is minimized because the neural network's training is halted when performance starts to decline. There is no researches found about logical connection between banking system collapse and Islamic finance, this research tries to study such relationship and recommend advices that can mitigate such risks. The process of creating the financial ratio reports is cumbersome; it takes long time and great effort, as well as it exposes many weaknesses in terms of the accuracy of the manual operations. Designing a more sophisticated and professional method to generate those ratios is highly required such as a data warehouse. The structure of this paper is as follows: the second part describe the process of bank ratios selection which will be used in this research. The third part covers factor analysis which is required as dimensional reduction techniques; The fourth and fifth parts describe the selected models DA and SVM respectively and finally discussion and conclusion will be presented.2 Procedure for Paper Submission

2 Selected Variables

We have conducted interviews with the departments which are concerned by the results of this research as well as possess the required knowledge to act as subject matter experts (SMES) represented in Prudential Supervision Department (PSD) at the Central Bank of Sudan. 27 employees were interviewed using online questionnaire distributed through local intranet. We came to the conclusion that their goal mainly: is to predict the status of banks, to see whether they are solvent or not. The most used method to determine this status is the CAMELS rating system adopted by the BASEL Committee. The BASEL Committee is the primary global entity that provides supervisory and prudential standards for banks; as it also provides a forum for cooperation on banking supervisory matters. Its mandate is to strengthen the regulation, supervision and practices of banks worldwide with the purpose of enhancing financial stability. We have analyzed the Responses and came up with the following results, based on the highest frequencies for each factor under each domain which identical to CAMELS ratios recommended by International monetary fund to supervision tasks.



Domain	Ratio	Description
Capital	T1C	Tier 1 Capital Ratio measured as a ratio of Tier 1 Capital to Risk Weighted Assets.
	TCR	Total Capital Ratio measured as a ratio of (Tier 1 + Tier 2 capital) to Risk Weighted Assets
Asset Quality	EAS	Ratio of Equity Capital to Total Asset.
	LAS	Ratio of Net Loans to Total Assets.
Earning	LLP	Ratio of Loan Loss Provisions to Total Loans.
	NPL	Ratio of Non-Performing Loans to Total Loans.
	NIM	Net Interest Margin measured as a ratio of (Interest Received – Interest Paid) to Total Earning Assets.
	ROE	Return on Equity Measured as a ratio of Net Income to Capital Equity.
	ROA	Return on Assets measured as a ratio of Net Income to Total Assets.
Sensitivity	IBR	Interbank ratio measured as a ratio of Deposits Due from Banks to Deposits Due to Banks.
Liquidity	LADF	Ratio of Liquid Assets to Deposits and Short Term Funds.
Management	IDIVER	Finance-Related Income to Total Income

Figure 1: *CAMELS Ratios*

2.1 Data Warehouse

We faced a challenge that our data is distributed in many location as well as should be gathered in very lengthy way, the decision we have made to design a data ware house (DW) that can serve even the enterprise (central bank) in its future, its main responsibility to extract those information, transform it (calculate the equations) load and cleanse it We started the DW design by identifying the source of information involved in calculating the selected financial ratios. We elicit the information required to calculate the ratios from the SMEs (Subject Matter Experts) in the targeted department (PSD) that they use to design their reports and get an idea about the mapping between banks columns and the required source of calculation. Then we designed our source database under the name (LZ)or the landing zone where we do all the data loading (weather new or modified). It also serves in committing with time allowed to execute the data loading packages from the database administrators. We designed a second layer (staging area) which performs all the data quality tasks as well as writing all the calculation expressions of the financial ratios. From the staging database, we cloned our DW to be the last destination of our loading process shaping the star schema

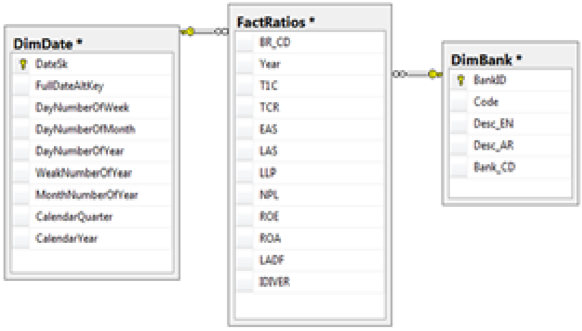
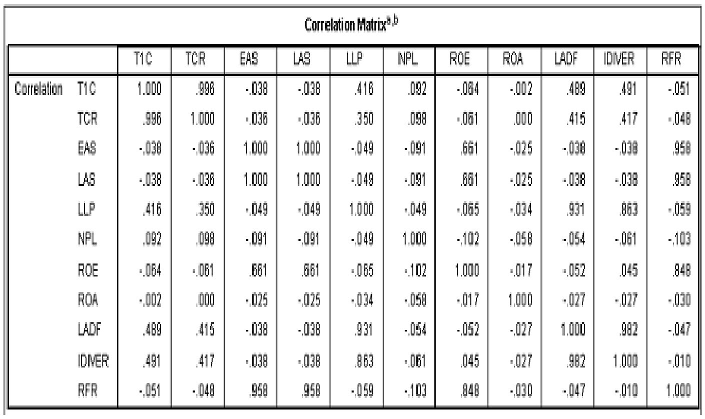


Figure 2: *Ratios Fact Table*

Two data sets have been configured using holdout method to train and test our designed models respectively, total of 559 healthy banks distributed to 493 as training data set and 66 as validation set, total of 67 non-healthy banks distributed to 33 training and 34 validation data sets.

3 Factor Analysis

The broad usage of factor analysis is to condense data so that relationships and patterns can be easily interpreted and understood. It is normally used to regroup variables into a limited set of other groups based on shared variance. Hence, it helps to isolate constructs and concepts. Factor analysis tries to bring inter-correlated variables together under more general, underlying variables. More specifically, the goal of factor analysis is to reduce "the dimensionality of the original space and to give an interpretation to the new space, spanned by a reduced number of new dimensions which are supposed to underlie the old ones" [1] When the data are appropriate, it is possible to create a correlation matrix by calculating the correlations between each pair of variables



Correlation Matrix[a,b]

		T1C	TCR	EAS	LAS	LLP	NPL	ROE	ROA	LADF	IDIVER	RFR
Correlation	T1C	1.000	.996	-.038	-.038	.416	.092	-.064	-.002	.489	.491	-.051
	TCR	.996	1.000	-.036	-.036	.350	.098	-.061	.000	.415	.417	-.048
	EAS	-.038	-.036	1.000	1.000	-.049	-.091	.661	-.025	-.038	-.038	.958
	LAS	-.038	-.036	1.000	1.000	-.049	-.091	.661	-.025	-.038	-.038	.958
	LLP	.416	.350	-.049	-.049	1.000	-.049	-.065	-.034	.931	.863	-.059
	NPL	.092	.098	-.091	-.091	-.049	1.000	-.102	-.058	-.054	-.061	-.103
	ROE	-.064	-.061	.661	.661	-.065	-.102	1.000	-.017	-.052	.045	.848
	ROA	-.002	.000	-.025	-.025	-.034	-.058	-.017	1.000	-.027	-.027	-.030
	LADF	.489	.415	-.038	-.038	.931	-.054	-.052	-.027	1.000	.982	-.047
	IDIVER	.491	.417	-.038	-.038	.863	-.061	.045	-.027	.982	1.000	-.010
	RFR	-.051	-.048	.958	.958	-.059	-.103	.848	-.030	-.047	-.010	1.000

Table 1: *Factors Correlation Matrix.*

In this matrix five clusters of variables with high intercorrelations are represented. Which can be grouped as following (T1C, TCR), (EAS, LAS, RFR), (LLP, LADF, IDIVER), (ROE), (ROA). Although the correlation coefficient between ROA, ROE and other some factors is high but cannot be grouped into one cluster as their negative correlation with some other group members. As has already been said before, these clusters of variables could well be “manifestations of the same underlying variable” [1]. The data of this matrix could then be reduced down into these five underlying variables or factors. With respect to the correlation matrix, two things are important: the variables have to be intercorrelated, but they should not correlate too highly (extreme multicollinearity and singularity) as this would cause difficulties in determining the unique contribution of the variables to a factor. Inter-correlations can be checked by using Bartlett’s test of spherity, which tests the null hypothesis that the original correlation matrix is an identity matrix. This test has to be significant: when the correlation matrix is an identity matrix, there would be no correlations between the variables.

4 Discriminant Analysis

Discriminant analysis is a classification problem, where two or more groups or clusters or populations are known a priori and one or more new observations are classified into one of the known populations based on the measured characteristics. Once data are loaded in the data warehouse, the next important part is the development of the final model and the selection of discriminatory variables that go in it. Then, from several combinations of the 11 discriminant ratios, few final models are proposed for validation. Discriminant analysis (DA) tries, as the name implies, to discriminate between different groups. In this case it discriminates between the group of healthy banks and the group of non-healthy banks In order to do this DA will take into account various samples from both groups. DA tries to separate these two groups based on the financial ratios of each sample. Each ratio becomes a variable X and gets its own coefficient V.

4.1 Discriminant Variable Selection

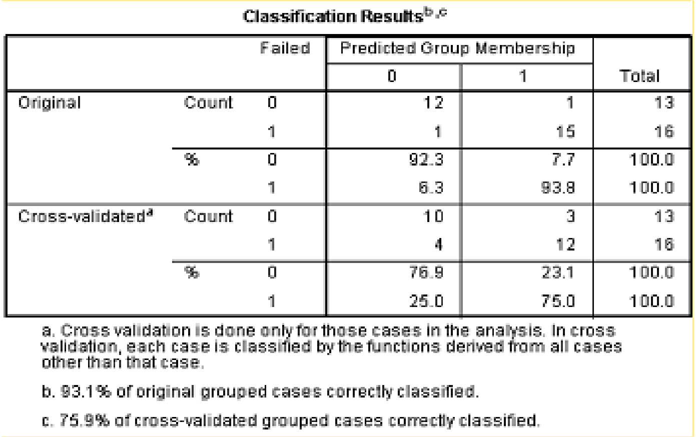
Bankruptcy prediction is a challenging exercise. Due to the multitude of factors that influence the bankruptcy's process, discriminations or differences between groups are difficult to determine. In particular, because of the nature of the data (financial statement data), discrimination will not be clear in the most of the cases15. Therefore, small oppositions and differences between groups are satisfying. More importantly, because of the assumption of multinormality, distributions that follow a Gaussian distribution are preferred, or at least approached due to the frequent violation of this latest assumption in LDA. However, the model is very robust to the loss of the assumption of multinormality as long as distributions between opposing groups approach a Gaussian distribution [3]

4.2 Variables Combination

In the previous section, discriminant variables were selected and conducted to 11 final selected ratios. The next task is to associate these 11 ratios in a multivariate model in order to obtain the ultimate discriminatory model. Since this process is mainly iterative, there is no claim regarding the optimality of the determined discriminant model [4]. However, the function developed aims to be the best among the ones tested. In order to develop different sets of discriminant models, different approaches are utilized. All models are developed on the training data. A combination of different approaches is a valuable help in selecting variables for the discriminant model. Thus, decision trees are performed in order to distinguish strong discriminant variables and to detect certain behaviors. This process offers an interesting exploratory analysis of variables. Then, correlation of variables, CAMELS sense and knowledge of the data, subjective judgment and a combination of all approaches are used to help find the best combination of discriminant variables.

4.3 Validation

From the previously mentioned work, different models of interests are developed through the classification analysis solution in SPSS. After testing how the models performed on the training data, few last models are kept for final consideration on the testing data. Thus, only one last model is kept as the discriminant model because of its overall better performances and characteristics. The performances achieved by the final model during each steps of the selection process are presented in the following parts: (i) to begin, the results from the validation of a function evaluation are presented. The idea in this latest part is to select within all models the best evaluation based on the training sample. (ii) Then, the results from the validation of the final function are presented. At this stage, the best model is selected. The validation of a function estimate is the first step performed in the validation and confirmation of the final model. At this stage, different year evaluations of models developed are compared for each model to determine what year appraisal is the best as well as to check for the overall healthiness to model fragility as general. Accurately, a good selected function should have the two potentials: (i) a good rate of good classification and (ii) the steadiness of its performance over time. Therefore, the selected function should license discrimination for several years after usage. Different approximations are built and tested on the training data. In particular, for each model eight assessments are obtained and tested, almost one for each year sample of the training data (2006-2014). In the case of the final model, the "2014 results" is selected for further consideration in the validation of the estimated function's part. Its results and comments are presented below. It can be concluded that the final model and its different estimates are performing correctly within the period of study. However, two estimates (2007, 2009 and 2013) do not demonstrate results as good as the others. Regarding the (2008, 2010 and 2012) estimates, the results are not high, especially for the rate of good classification of non-healthy banks; this result is explained by the possibility of misleading data provided by some banks in their financial information.



Classification Results[b,c]

		Failed	Predicted Group Membership		Total
			0	1	
Original	Count	0	12	1	13
		1	1	15	16
	%	0	92.3	7.7	100.0
		1	6.3	93.8	100.0
Cross-validated[a]	Count	0	10	3	13
		1	4	12	16
	%	0	76.9	23.1	100.0
		1	25.0	75.0	100.0

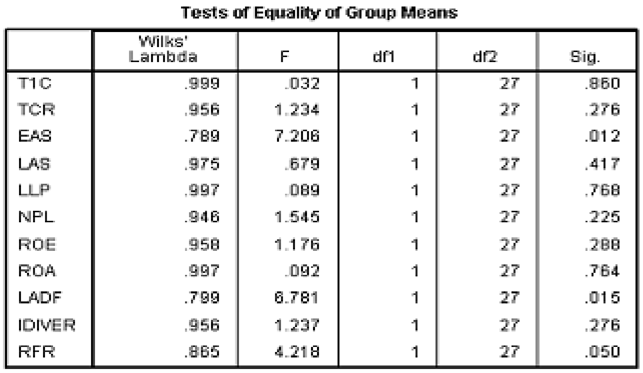
a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 93.1% of original grouped cases correctly classified.

c. 75.9% of cross-validated grouped cases correctly classified.

Table 2: *Classification Results for 2014 estimates.*

Still, in the case of the final model, three last estimates remain at this stage of the process: 2006, 2011, and 2014 estimates. The results of these estimates show that they are for some extent close to each other as shown in Appendix E but 2014 estimate give higher results with same count of data samples (29) contrast to 2006 performance is lowest comparing to others although with fewest number of available data samples. 2011 estimate is promising but overall performance is lower than 2014 one and the wide spread between opposing group is noticed in its results however it is considered the best candidate to replace the selected model in case of bad validation performance. Between other models with their best estimates, this 2014 estimate is considered for further testing as illustrated in table 2. The results of the selected function on the training data are promising with an overall percentage of correct prediction equal to 93.1 percent. However, two clarifications should be taken into consideration: first, even if results are promising they should be confirmed on a testing data (performed later in this thesis), the second observation concerns the particular rates of good classification of healthy and non-healthy banks. Certainly, when attention is paid to these rates, it can be concluded that they are balanced and that there is a non-influencing spread of 1.4 percent between their performances. Seen from another angle, type I (6.3 percent) and type II (7.7 percent) errors are quite identical for each other. Type I errors represent the misclassification of non-healthy banks as healthy ones. Type II errors are the misclassification of healthy banks as non-healthy banks. Therefore, the model appears to be not biased in favor of healthy prediction and to the detriment of non-healthy banks, which are the ones of particular interest. However, this rate is still largely acceptable and is highly performing on the training data. The model gives the highly discriminant variables represented in (EAS, LADF, RFR), as their p-value is less than or equal to .05, then among 11 ratios only three of them are used by DA


Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
T1C	.999	.032	1	27	.860
TCR	.956	1.234	1	27	.276
EAS	.789	7.206	1	27	.012
LAS	.975	.679	1	27	.417
LLP	.997	.089	1	27	.768
NPL	.946	1.545	1	27	.225
ROE	.958	1.176	1	27	.288
ROA	.997	.092	1	27	.764
LADF	.799	6.781	1	27	.015
IDIVER	.956	1.237	1	27	.276
RFR	.865	4.218	1	27	.050

Table 3: *Test of Equality for Group Means.*

When using equal prior probabilities (as was done in this experiment) the cut-off score for Z is calculated by taking the average of the group averages Z-score (refer to appendix E10). Here that is ((1.464-1.190)/2) = 0. 137. The positive number is the Z-score of the non-healthy banks average; the negative one of the healthy banks. In order to qualify a sample, its Z-score is calculated using the aforementioned formula. If the score is below 0.137, it is classified as healthy. If the score is above 0.137, the sample is classified as non-healthy. Discriminant analysis uses two financial ratios (LADF, EAS) and one non-financial ratio (RFR). LADF is ratio of liquid assets to deposits and short term funds (liquid assets / deposits and short term funds) that represent the importance of liquidity management for commercial banks and their relevant measures to on-demand types of deposits, EAS is ratio of equity capital to total asset (Equity Capital / Total Asset) which holds implications of the necessity of raising the equity capital to a convenient portion from the whole value of assets for the banks.

5 Support Vector Machines

Support vector machines (SVMs) are a set of new supervised learning methods used for binary classification. Among all classification algorithms SVM is strong because of its simple structure and it requires less number of features. SVM is a structural risk minimization classifier algorithm derived from statistical learning theory by Vladimir Vapnik and his colleagues in 1992. Support Vector Machines were first introduced to solve the pattern classification and regression problems.

5.1 Model development

Compared with the limitations of other intelligent models, the major advantages of the SVM are as follows: first, SVM has only two experimental parameters, namely the upper bound and the kernel parameter. Obtaining an optimal combination of parameters that produce the best prediction performance is an easier task [5] Second, the SVM guarantees the existence of a unique, optimal, and global solution because SVM training is equivalent to solving a linearly constrained QP [5]. Third, the SVM implements the SRM principle that is known to have good generalization performance, Finally, the SVM can be constructed with small training data sets to obtain prediction performance [7]. These four advantages support our proposed hybrid model that adopts the SVM technique for financial distress prediction. The approach in developing prediction models rests on determining whether information from the outcomes, as reflected in the data prior to bank wen non-healthy, can provide signals of that impending event. In a dichotomous classification setting, that is, to predict one or the other class from a combined set of two classes (e.g., healthy and non-healthy), the development of a support vector machines model, as with other models of prediction, begins with the design of a training sample is the input information for the training object i on a set of m independent variables and corresponding outcome (dependent variable). Formally, the aim of the analysis is the development of a function that distinguishes between the two classes of healthy and non-healthy banks In the simplest case, f(x) is defined by the hyperplane XW = y as follows:

f(x)=sgn(xw-y)

Where w is a normal vector to the hyperplane and y is a constant. Since f is invariant to any positive rescaling of the argument inside the sign function, the canonical hyperplane is defined by separating the classes by a “distance” of at least 1. The analysis of the generalization performance of the decision function f(x) has shown that the optimal decision function f is the one that maximizes the margin induced in the separation of the classes, which is 2/||W|| Hence, given a training sample of n observations, the maximization of the margin can be achieved through the solution of the following quadratic programming problem:

$$\min \frac{1}{2W^T} W + Ce^T y \text{ ,s: t: } \mathrm{D(Xw - \ ey) + \ ye} \text{ ,} \mathrm{y} \geq 0 \qquad (1).$$

W , y belongs to R , where D is n x n matrix such that $D_i j = d_i$

And $D_i j = 0 \text{ , } i \neq j \text{ , } X \text{ is an } n \text{ x } m \text{ matrix}$ with the training data, e is a vector of ones, y is an n x 1 vector of positive slack variables associated with the possible misclassification of the training objects when the classes are not linearly separable, and C > 0 is a parameter used to penalize the classification errors. From the computational point of view But it is more convenient to consider its dual Lagrangian formulation

$$\max e^T u - \frac{1}{2} DXX^T \mathrm{Du; (S.t)} \ e^T D u = 0 \text{ ,} 0 <= u <= Ce. \qquad (2)$$

The decision function is then expressed in terms of the dual variables u as follows:

$$f(x) = sgn(xX^T Du - y).$$

Various kernel functions exist, such as the polynomial kernel, the radial basis function (RBF) kernel, the sigmoid kernel, etc. [8]. The representation of the data using the kernel function enables the development of a linear model in the feature space H. Since H is a nonlinear mapping of the original data, the developed model is nonlinear in the original input space. The model is developed by applying the above linear analysis to the feature space H. We will explore the development of both linear and nonlinear SVMs models with a polynomial and an RBF kernel. The width of the RBF kernel was selected through a cross-validation analysis to ensure the proper specification of this parameter. A similar analysis was also used to specify the tradeoff constant C. All the data used during model development was normalized to zero mean and unit variance. As with any supervised learning model, we first train a support vector machine, and then cross validate the classifier. Use the trained machine to classify (predict) new data.

5.2 Results and Validation

The process will be decomposed in two parts. Firstly, the SVM machines will be built with the full set of variables (11 ratios) and study the best combinations of machines configuration that lead to best prediction model. Secondly, the DA-SVM model will be examined, which focused on the variables that proofed high discrimination power in the discriminant analysis phase which are (EAS, LADF, RFR) from the 2014 data set, those variables will be studied through the best SVM model reached in the first part. Finally, comparison between two models will be presented and discussed.

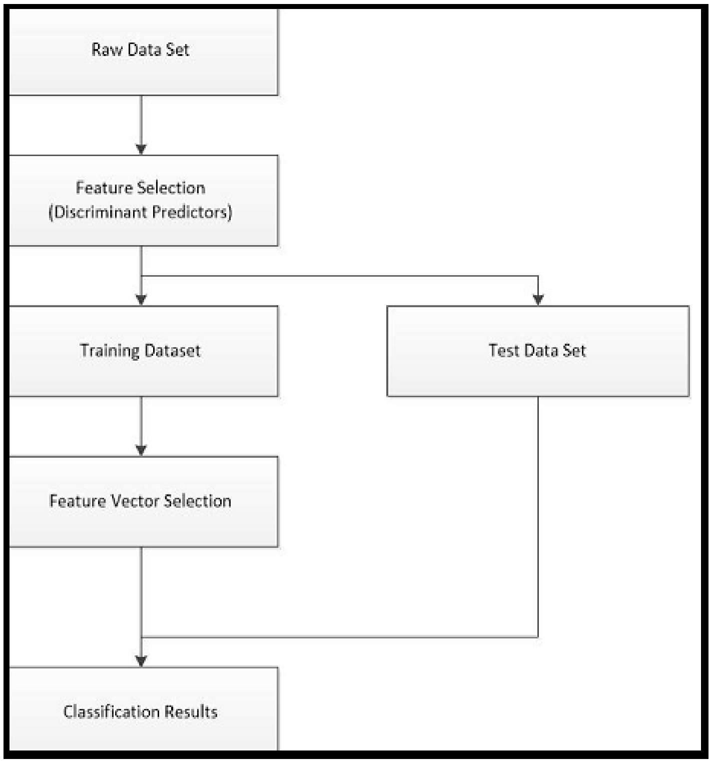


Figure 4: *Flowchart of the proposed DA-SVM framework for bank distress prediction.*

Following table present the result of SVM classification for the full factor analyzed data set, the best result achieved with Fine Gaussian kernel function with 65.2 percent accuracy and lowest type I and II errors respectively compared to liner, quadratic, cubic functions with the capacity fixed at C = 1 and KernelScale = .83. Capacity C in the non-separable case (often called Soft-Margin SVM), one allows miss classifications, at the cost of a penalty factor C, Making C large increases the weight of miss classifications, which leads to a stricter separation. This factor C is called box constraint. The reason for this name is that in the formulation of the dual optimization problem, the Langrange multipliers are bounded to be within the range [0, C]. C thus poses a box constraint on the Lagrange multipliers. The kernel scale parameter is also called "gamma". If gamma is large, then the kernel will fall off rapidly as the point y moves away from x for the following kernel:

$$K(x,y) = \exp(-gamma * (x - y)^2) \qquad (3)$$

As gamma decreases, the kernel will fall off less and less rapidly. When gamma is 0, the kernel will be the same (=1) for all points y irrespective of where y is in the feature space. In this interpretation, gamma is related to how spread out our data points is. If they are very far from each other (which would happen in a very high dimensional space for example), then we don't want the kernel to drop off quickly, so we will use a small gamma as we suggest in this configuration (.83)

Phase	Kernel function	Accuracy %	Error Type I %	Error Type II %
Training	Linear	48.3	77	31
	Quadratic	62.1	38	38
	cubic	51.7	46	50
	Fine Gaussian	**65.2**	**59**	**25**
Validation	Linear	50	62	28
	Quadratic	60	38	43
	cubic	46.7	56	50
	Fine Gaussian	**66.5**	**54**	**24**

Table 4: *SVM Classification Result (All ratios).*

In order to make comparative study Logistic regression and BP-NNs were also carried out in the experiment. 5-fold method was used to test the validity of different models, because k-fold accuracy can objectively reflect the models' ability to predict financial distress for banks outside the training samples. . SPSS 18 was utilized for Logit analysis. MATLAB 9 was used for BP-NNs, and its structure was 11-10-1, and the learning rate was set as 0.1, LIBSVM software developed by Prof. Lin Chih-Jen in Taiwan University was used for SVM modeling and testing. The experiment results are list in following table

Model	Logit	SVM	BP-NN
Training	54.1	65.2	61.4
5-fold	60.9	66.5	63.1

Table 5: *SVM Comparative Study.*

SVM has the highest accuracy. Whether from the perspective of training accuracy or 5-fold accuracy, SVM performs better than BP-NN and logit models. The training accuracy of SVM is a slightly better than BP-NNs which show that SVM has better generalization ability than BP-NNs and can better avoid over-fitting phenomenon. So, SVM has a better balance among fitting ability, generalization ability and model stability. By finding out support vectors for financial distress prediction from training samples, SVM is suitable to predict financial distress for banks outside training sample, and it can keep the predictive accuracy relatively stable when the training samples change within a certain range. The figure below present the scatter plot of miss predicted values (x) compared with truly predicted which accounts more and shapes hyperplane clearly separates the two classes of healthy and non-healthy banks in the intersection of EAS and RFR ratios.

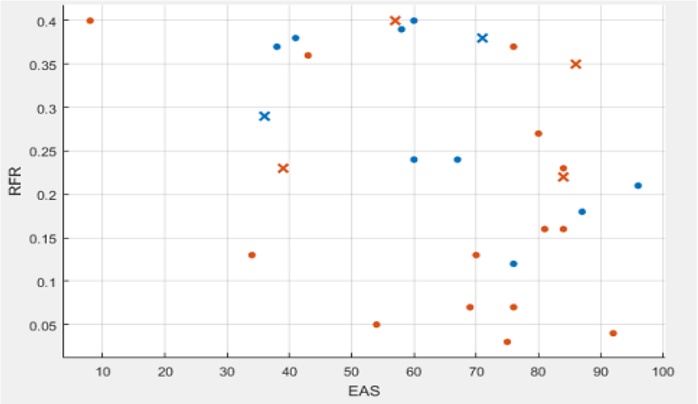


Figure 5: *DA-SVM Plot for EAS, RFR classification.*

Following Table presents the classification result of DA-SVM model which is outperform the full data set SVM model (83.7 percent and 65.5 percent respectively) and slightly better than DA classification , The best result of a specific performance measure is highlighted in boldface in the Table, all the selected kernels are also outperform the full data set SVM model with at least 15 percent accuracy increasing , that indicates the proposed DA-SVM model gives satisfactory results and the two financial ratios (LADF, EAS) and one non-financial ratio (RFR) are representing good predictors for Sudanese banking system distress. LADF is ratio of liquid assets to deposits and short term funds. EAS is ratio of equity capital to total asset. (RFR) is a new proposed ratio which genuinely introduced by this research, DA-SVM again considers it one of the good predictors.

Phase	Kernel function	Accuracy %	Type I Error %	Type II Error %
Training	Linear	75.9	41.2	28.1
	Quadratic	65.5	31.5	29.7
	cubic	72.4	38	29
	Fine Gaussian	**79.3**	**31**	**13**
Validation	Linear	76.8	39.8	26.4
	Quadratic	78.4	22.6	23.7
	cubic	73.8	27.4	25.7
	Fine Gaussian	**83.7**	**17.9**	**10.4**

Table 6: *DA-SVM Classification Result*

6 Discussion

This research has presented a process of design prediction model of Sudan's banking sector distress first of all factor analysis as feature selection methodology has applied, in the rotation task After applying varimax rotation method we can observe on the four components extracted (EAS, LAS, ROE, RFR) constitutes the first component with higher values, (LLP, LADF, IDIVER) the second component, (T1C, TCR) the third component and ROA the fourth component. If this results compared with CAMELS original distribution of factors there are some mismatches regarding the ROE factor which classified with Asset quality group while it's a profitability ratio likewise RFR (Rain Fall Ratio) which is basically not considered a CAMELS ratio yet and its one of this research novel factors, the rotated component matrix has classified it as one of asset quality ratios , this can lead to new factor groups rather than already identified by CAMELS ratio system. Instead of 18 initially selected financial ratios as a result of questionnaire answered by SMEs the analysis conclude to 11 ratios own the same predictive power which used throughout this research. Second, discriminant analysis has been performed,the model give the highly discriminant variables represented in (EAS, LADF, RFR), as their p-value is less than or equal to .05, then among 11 ratios only three of them are used by DA, Discriminant analysis uses two financial ratios (LADF, EAS) and one non-financial ratio (RFR). LADF is ratio of liquid assets to deposits and short term funds (liquid assets / deposits and short term funds) that represent the importance of liquidity management for commercial banks and their relevant measures to on-demand types of deposits, EAS is ratio of equity capital to total asset (Equity Capital / Total Asset) which holds implications of the necessity of raising the equity capital to a convenient portion from the whole value of assets for the banks.The third ratio used by discriminant analysis (RFR) is a novel ratio has never been used before in any published study, the researcher rely on the high dependency of Sudanese banks on agricultural and animal resources trade operations to maximize their profitability which have direct effect with the RFR (Rain fall ratio), DA proofs this expectations and consider it one of the discriminant variables. Third, SVM model has been developed with the full data set (11 ratios). Followed by a second model which built using the discriminant factors identified by DA process, the new model called DA-SVM model which outperformed SVM model and comparing to similar studies stated in literature review , the model performance is satisfactory with overall accuracy 83.7 where the highest related published study identified is 81.82 percent (Fu Shuen , 2012) , we meant by related study those exhibited in literature review and with main focus on macroeconomic perspective because the nature of data sets used throughout this research constitutes a comprehensive view for whole banking sector to truly predict its distress possibility factors. To fulfill this goal, the comparative study has been performed with other classification models namely (logistic regression and back propagation neural network), because they recorded satisfactory performance measures in previous studies, results confirmed the distinctive power of SVM model with slight better performance than BP-NN.

7 Conclusion and Future Work

This research aimed to apply and evaluate different statistical and intelligent models to predict commercial banks failures in Sudan. Based on the experimental results, we concluded the following: First, the research data applied to the techniques is very important for effective predictions. The performances of various techniques differ with respect to the form of data set applied. On the other hand, since different prediction performances are obtained in training and validation data sets, it is difficult to adopt a unique technique for this issue. The research starts with factor analysis task as feature selection methodology, followed by discriminant analysis to identify the ratios with superior predictive power then design the hybrid model of DA-SVM to study the possibility of having better performance. As many studies in a number of fields suggested, the superiority of SVM in prediction problems is proven once again here. As a newly developed learning algorithm, DA-SVM model gives promising results. Features selection has shown that new groups can be identified from CAMELS ratios and narrowing the data set space to 11 factors instead of 18. Discriminant analysis has identified 3 ratios with highest predictive power which are: EAS, LADF and RFR, the later ratio is a novel one used for the first time by this research. The quality of bank distress prediction can be improved. The current work has focused only on data mining techniques. Thus, in future work, we are planning to extend the proposed methods for applying text mining techniques as well by examining commercial banks web sites and web-published reports as well as client's complaints which received through central bank web site. Second, we are also planning to apply the current work on other countries has similar economic conditions to turn the designed model as regional oriented solution. The goals of future work will be achieved as the following objective:

I. To mine the texts pertaining to Sudan's banking sector specially the customer complaints weather in social media or direct posting on central bank web site.
II. To study applying the new text mining model into regional data set.
III. To identify other bank factors that can also affect and contribute on Sudan's banking failure.

REFERENCES

[1] Rietveld, T., van Hout, R., Rietveld, A.C. and Van Hout, R. (1993) Statistical techniques for the study of language and language behaviour. Germany: Mouton de Gruyter. (references)

[2] Erdogan, A. (2016) 'Applying factor analysis on the financial ratios of turkey's top 500 industrial enterprises', International Journal of Business and Management, 8(9). doi: 10.5539/ijbm.v8n9p134.

[3] Bardos M. (2001): Analyse discriminante: application au ri sque et scoring financier, Dunod.

[4] Altman, E.I., Marco, G. and Varetto, F. (1994) 'Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)', Journal of Banking and Finance, 18(3), pp. 505–529. doi: 10.1016/0378-4266(94)90007-8.

[5] Chang, Y.-W. and Hsieh, C.-J. (2010) 'Training and Testing Low-degree Polynomial Data Mappings via Linear SVM', Journal of Machine Learning Research, 11, pp. 1471–1490.

[6] Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', Machine Learning, 20(3), pp. 273–297. doi: 10.1007/bf00994018..

[7] Ribeiro, B., Silva, C., Chen, N., Vieira, A. and Carvalho das Neves, J. (2012) 'Enhanced default risk models with SVM+', Expert Systems with Applications, 39(11), pp. 10140–10152. doi: 10.1016/j.eswa.2012.02.142.

[8] Chang, Y.-W. and Hsieh, C.-J. (2010) 'Training and Testing Low-degree Polynomial Data Mappings via Linear SVM', Journal of Machine Learning Research, 11, pp. 1471–1490.

[9] http://www.CBOS.gov.sd.

[10] Memic, D. (2015) 'Assessing credit default using logistic regression and multiple Discriminant analysis: Empirical evidence from Bosnia and Herzegovina', Interdisciplinary Description of Complex Systems, 13(1), pp. 128–153. doi: 10.7906/indecs.13.1.13.

[11] Zavgren, C.V., Dugan, M.T. and Reeve, J.M. (1988) 'The association between probabilities of bankruptcy and market responses?A test of market anticipation', Journal of Business Finance and Accounting, 15(1), pp. 27–45. doi: 10.1111/j.1468-5957.1988.tb00118.x.