بسم الله الرحمن الرحيم-

Sudan University of Science and Technology

Collage of Graduate studies

A Model for Prediction of Financial Distress in Sudanese Banking System Using a newly built data set.

A thesis Submitted in partial fulfillment for the Requirements of the Degree of Doctor Philosophy in computer sciences

By

Mohammed Awad SirElkhatim Nasir

Supervised by

Prof. Dr. Naomie Salim

2017

Dedication

To my Lovely parents.

Acknowledgement

All praise to Allah (s.w.t) the most Gracious and most Merciful, by whose grace and blessing this work has been completed. I would like to take this opportunity while relying on the instruction of the Prophet to the effect that "whoever does not thank people does not thank Allah" to express my thanks and gratitude to those who have contributed in one way or another through their advice, criticism, and support in strengthening the quality of this work. I would like to express my gratitude to those who have helped me in my pursuit for knowledge.

I would especially like to express my deep and sincere gratitude to my supervisor **Prof. Dr. Naomie Bt Salim**, for her attention, continuous guidance, and support throughout the length of this study. She has greatly helped me in a lot of ways I needed to go through this study. I am grateful to her for giving her wide knowledge, time and guidance to help me overcome the challenges in my study. I am also immensely grateful to **Prof Elzzedin M Osman** for his kind cooperation, as well as to all staff of Sudan University who extended their best cooperation during my study and their professionalism of tackling my personal obstacles.

My deepest thanks go to my parents and my brothers and lovely sister. Their influence made me realize the importance of education from a very early age. I also offer the deepest gratitude to my sweet hearts – my Wife, my sons (Zyad & Kenan) – for bearing my ignorance towards them during the course of this study

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. This is originally a classification problem to categorize banks as healthy or non-healthy ones in order to design the required measures and policies to mitigate the risks for non-healthy banks. This study aims to apply Discriminant analysis and Support Vector Machines methods to the bank failure prediction problem in Sudan, 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 have 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. Feature 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.

Financial analysts are focusing on finance sectors in order to determine which sector is subject to special study. Transportation Sector and Short Term Local finance Sector is considered the most significant sector in bank default probability. Payment type was not found the best predictors for Islamic credit risk analysis. The research outputs can be utilized by monetary policy regulator to

monitor commercial banks by focusing on the discovered important predictors as well as review all polices with regard to deferred credit finance mode as well as transportation sector finance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CBOS	-	Central Bank of Sudan
ROE	-	Return on equity measured as a ratio of net IncometoCapitalequity.
T1C	-	Tier 1 capital Ratio measured as a ratio of Tier 1 capital to riskweighted assets.
TCR	-	Total Capital ratio measured as a ratio of (Tier 1 + Tier 2 capital) to risk weighted assets
ROA	-	Return on assets measured as a ratio of net Income to total assets.
EAS	-	Ratio of equity capital to total asset.
LAS	-	Ratio of net loans to total assets.
LLP	-	Ratio of loan loss provisions to total loans.
LADF	-	Ratio of liquid assets to deposits and short term funds.
IDIVER	-	Finance related income to total income
FA	-	Factor Analysis
DA	-	Discriminate Analysis
LOGIT	-	Logistic Regression
NPL	-	Non-Performing Loan
SVM	-	Support Vector machine
BP-NN	-	Back-Propagation Neural Network
RFR	-	Rain Fall Ratio
FDIC		Federal Deposit Insurance Corporation

LIST OF APPENDICES

Chapter One

1.1 Introduction

The banking sector of Sudan 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 intermediation and the sector's small size and relative isolation from global financial markets. (CBOS, 2016)

Bank failure occurs when a bank could not 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 be used to pay on-demand deposits for its 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. A bank may also be taken over by the regulating government agency through central or reserve banks if shareholders equity is less than regulatory minimum.

It is often feared that 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 at the same time panicked

customers try to withdraw their deposits. Thereby, the effect of bank fear has a multiplier effect on all other firms. 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 healthy and non-healthy banks which can help on design different policies for each group.

The acronym CAMELS stand for the following factors that examiners use to rate bank institutions: Capital Adequacy, Asset Quality, Management, Earnings, Liquidity and Sensitivity.

CAMELS are used to rate the financial institution according to six factors as the 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 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 gross domestic product (GDP)(CBOS,2016). 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 underutilization and misallocation 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.

The organization of this chapter is described as follow: section 1.2 reviews the background of the problem while section 1.3 presents the problem statement

Then, section 1.4 discusses the objectives of study. Section 1.5, 1.6 and 1.7 revolve on the scope of study, significance of study, and expected contribution, respectively. Finally, section 1.8 describes the organization of thesis.

1.2 Background of Problem

Bank distress prediction is an inevitable study for monetary policy makers and regulators in the banking industry because the failure of one bank can lead to dramatic consequences. If bank failures are perfectly predicted, early warnings can be spread out to the people responsible of crisis remedy. Currently Central Bank of Sudan (CBOS) relies basically on On-Site inspection to control and monitor the underlying banking system. This type of protection requires a high volume of cost and effort as well time consuming. There is off-site supervision tasks performed as well but unfortunately cannot protect the institution before the hazard occurs. It depends totally on analyzing of the collected data and generate a report to reflect the current status. There is no overall banking failure prediction model in place.

(Detragiache, 1998) explores the reasons of the probability of a banking crisis around the world in 1980—1994 using a multivariate Logit model. They discover that bank crises are more undoubtedly in countries with low GDP growth, high real profit rates, high inflation rates, and explicit deposit insurance system. Those countries are more liable to balance of payments crises also has a higher probability of experiencing banking crises. Sudan is a typical example of the identified type of countries which indicates high possibility of banking crisis unless there are some actions in place.

(Wheelock, 2000) analyzes what factors can be used to predict bank failure in the United States, particularly. The authors use competing-risks hazard models with time-varying covariates. They find that banks with lower capitalization, higher ratios of loans to assets, poor quality loan portfolios and lower earnings have higher risk of failure. Banks located in states where branching is permitted are less likely to fail. This may indicate that an ability to create a branch network, and an associated ability to diversify, reduces banks' liability to failure. Further, the more efficiently a bank operates, the less likely the bank is to fail. This explains the power of CAMELS ratio system represents on assets and capital in predicting bank failure.

Our research utilizes the Features Selection (FS) techniques to obtain the most optimal set of factors that can predict banks failure. Four types of filter-based and wrapper based features selection will be studied in order to propose the best techniques to use. This research also tries to maximize the accuracy of factors quality by relying on subject matter expert's opinions or knowledge experts to determine which factors should be used to predict the likelihood of Sudan's banking system collapse. The selected factors have been used as an initial input vector.

(Canbas,2005) assembled four different statistical techniques (principal component analysis (PCA), discriminant analysis (DA), logit analysis (LA) and probit analysis (PA)) to develop the integrated early warning system (IEWS) that can be used in prediction of bank distress. At the beginning, principal component analysis (PCA) was used to discover the preliminary financial characteristics of the banks. Further on, discriminant analysis (DA), logit analysis (LA) and probit analysis (PA) models were estimated based on highlighted previous characteristics to construct the IEWS model.

Neural networks (NN) are probably the most widely used model among the intelligent techniques (Demyanyk, 2010). Its principle is to mimic the biological neural networks of the human nervous system through an algorithm. This latest technique offers two intrinsic advantages compared to classic statistical ones that we mentioned earlier. First of all, neural networks as non-parametrical models do not depend on specific assumptions like the distribution of predictors or properties of data. This makes it theoretically more reliable than models that would have their assumptions violated (as it is often the case and not the exception with financial data (Bardos, 2001). The other advantage is the dependence on nonlinear approaches, which offers extended possibilities for testing complex data patterns. A downside is that NN models may be more influenced by temporal or cyclical changes in the economy than classic statistical techniques (Bardos, 2001). Neural networks may also be difficult to interpret (Paliwal and Kumar, 2009).

This research tries to propose a highly efficient classification model by utilizing parameters optimization techniques and ensemble SVM classification model.

Although most researches recommended Neural Network as a better intelligent technique to be used at bankruptcy prediction tasks, this research tries to find out if Support vector machines (SVM) can show 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 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 BP neural network design that minimize over-fitting. Over-fitting is minimized because the neural network's training is halted when performance starts to decline.

(SALMAN, 2007) and (DIANA KIRIGO, 2014) studied the Islamic finance with regard to commercial banks failure , they discovered that Islamic banking system is relatively stable compared to conventional one, however, some Islamic banks[1] have shown signs of financial distress and few were forced to close their operations.

This research tries to study the relationship between non-healthy banks and Islamic finance modes, sectors and payment types to advice about which one of them require special policies and regulations.

The process of creating the financial ratio reports is cumbersome; manual spreadsheets are being edited by financial analyst after getting the data from data acquisition system (DAS). This operation 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.

[1] For example, Ihlas Finance House, an Islamic financial institution, in Turkey was closed in 2001 due to liquidity problems and financial distress. Bank Taqwa was closed in 2001. Faisal Islamic Bank closed its operations in the UK for regulatory reasons.

1.3 Problem Statement

The backbone question of this research is: how can we define a model to predict commercial banks failures in Sudan in order to make the nation economy well and stable?

This question is supported by many sub-questions:

1. What are the important factor(s) which represent the highest risks for Sudan's banking sector?
2. What are the risk rankings of the Islamic finance factors?
3. What methods can be designed to effectively predict banks distress?

1.4 Objectives of the study

The ultimate goal of this research is to develop a model to predict the distress of banks operating in Sudan in order to enable decision makers to take appropriate action to eventually help save the banking sector.

The aim of the research is supported by the following objectives:

1. To study the factors that represents the high potential risk areas for banks.
2. To propose a model that can be used to predict the bank financial distress depending on the factors on the first objective.
3. To identify the most trusted Islamic finance methods that have the low risk impacts on bank failure.

1.5 Scope of Study

The former section has clarified the aims of this study which concentrate on how to propose a distinctive model that can be used to identify troubled banks, the following aspects are the scope of research for those objectives:

1. The study focus only on factors produced by international bank-rating system called CAMELS as well as NEW novel factor (Rainfall Rate).

Sudan is an agricultural country, most of banking finance operations are directed to agriculture sector. Farmers are highly depend on rain in their farming , it has been known that many agricultural projects have failed due to few rate of rainfall. This create the likelihood of risk on banks which are financing such kind of projects.

2. The study proposes a hybrid features selection technique called DAGA-FS that combine the benefits of discriminant analysis and genetic algorithms. Those techniques have been selected after thoroughly reviewing the literature and compare its performance with other methods.
3. Evaluation of results has been done with comparing the model results with the original reports produced by regulatory authority (Central Bank of Sudan), as well as using Accuracy, Specificity and Sensitivity metrics.
4. The Study uses 16 Islamic finance modes which are commonly implemented in any Islamic banking system.

1.6 Significance of the study

The study investigates bankruptcy prediction task using the capabilities of statistic and machine learning methods, which can help to predict banking sector collapse and assist in off-site supervision by regulatory authorities. The significance of this research is to identify the highest risk CAMELS factors which can lead banks to bankruptcy by using principal component analysis as a dimensionality reduction method besides conduct supervised learning through collecting those significant factors from subject matter experts. This research tries to propose new factor which is (Rainfall ratio) and discover its relationship with specialized agricultural banks due to special financing operations practiced by Sudan banking sector which heavily directed on agricultural finance. The research also contribute on evaluating the Islamic banking finance modes by

identifying which of them are the main causes of bank's operation deterioration. The performance of the proposed model is evaluated based on the pre-existing reports prepared by regulatory authority and compared as well compare the model results with other previous researches in same area, Ultimately the research try to give recommendations which can Reduce the Cost that 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 bearing in mind that 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.

1.7 Expected Contributions

The expected contributions can be described as follow:

1. Design the first prediction solution for the Sudanese banking sector.
2. Identify the factors which assist on predicting bank's financial distress.
3. Help in stabilization of Sudan economy by maintain healthy banking system.
4. Enhances CBOS' ability to predict Potential crisis, and it enables them to manage, coordinate and supervise more efficiently.
5. Early distinction between troubled and sound banks allows for appropriate actions to prevent failure and to protect healthy institutions.
6. Identify the Islamic finance features which have a connection with bankruptcy problem.

1.8 Thesis Organization

This thesis is organized into seven chapters that show as follow:

- Chapter 2, Literature Review: this chapter presents the state-of-the-art approaches in Bank's distress problem. An overview of the survey in the research areas is covered by this chapter. Some information and issues that

related to statistical modeling, and Intelligent modeling , this chapter also discusses data sets and evaluation measures of bank distress.

- Chapter 3, Methodology: this chapter describes the methodology used to achieve the objectives of this research. Also, and specifying the operational framework and discussing the experiments and models.
- Chapter 4, Features Selection Techniques: this chapter discusses possibilities of proposing new features selection techniques (DAGA-FS) that can be used to identify the most significant factors that affect the prediction ability of classification models.
- Chapter 5, Hybrid Classification of Evolutionary Algorithms and Bootstrap Aggregation Support Vector Machine: this chapter covers the prediction models that have been design to fulfill the research objective which are: NN, optimized single SVM, as well as the ensemble system SVM.
- Chapter 6, Islamic Credit Analysis: this chapter covers the part of identification the risky finance factors related to commercial banks performance and weather they might make the Islamic credit process risky.
- Chapter 7, Conclusion and future work: this chapter discusses and highlights the contributions and findings of the research work and presents suggestions and recommendations for future study.

Chapter Two

Literature Reviews

2.1 Introduction

This chapter presents some of the existing approaches in banks distress prediction, and the literature reviews that are provided here established the background for this research. An overview of the survey in the research areas is covered by this chapter. Furthermore, this chapter reviews a theoretical explanation on the fundamental methods on which the current study is expected to rely. The most important evaluation measures of bankruptcy prediction are also presented. This chapter is organized into nine sections. Section 2.2 is the basic introduction of banks distress prediction, while Section 2.3 explains banks distress prediction methodologies. Section 2.4 gives the related work regarding the prediction of bank failure and methods that will be used in this study. Section 2.5 presents an evaluation measure for models proposed. Finally, in Section 2.8 and 2.9, a discussion and the chapter summary are given respectively.

2.2. Prediction of Banks Distress

Recently, a mass amount of studies has been dedicated to explaining financial distress and failure of financial institutions, more specifically banks, because of their nature, high volatility and fragility with extreme ability to fail. Determining which bank is going bankrupt is a very complex task but remain an important concern for the all associated parties. The importance of this task can be exemplified in following dimensions First, the findings help regulatory

authorities (e.g. Central Bank of Sudan) in their duty to maintain a guarded and stable system. Second, the early detection of potential problems is likely to reduce the expected cost of a bank failure and to decrease the possibility of the problem pervasion more widely through the banking services. If the features of potential failure can be specified, this aids the regulators to concentrate on their limited resources. Instead of wasting much time in supervising sound banks, prudential supervision teams can devote the time for the rehabilitation processes and adopt close monitoring programs to the distressed and fragile banks. Since most banking regulatory and supervisory authorities employ such early warning systems (EWS), apparently they find them sufficiently useful. Nevertheless, there is a considerable scope for improvement as the number of surprise failures in the global financial crisis suggested.

In the late 1970s, US developed and implemented an Early Warning System (EWS) to assess the financial, managerial, and operational strength and weaknesses of financial institutions. Its quality and success is explained by the fact that this model is valid for nearly all theoretical researches as well as industrial experiments in all countries. The EWS is featured by a set of factors or ratios obtained from financial statements.

These ratios are classified into six categories: capital adequacy (C), asset quality (A), management competence and expertise (M), earning ability and strength (E), liquidity (L) and sensitivity to market risk (S), jointly referred as CAMEL(S).For each component the regulatory authority assigned a rating using a scale from 1 (good) to 5 (bad), where ratings of 1 or 2 are considered to present no or little supervisory concern which put the institution under the green zone with low to no assistance and ratings of 3-5 are the subject of moderate to

major concern. These individual scores for each individual ratio are combined to provide an overall rating also with a scale from 1 to 5. An overall rating of 3 to 5 sends a signal to regulatory authority that immediate intervention is highly required. The development and the implementation of similar Early Warning Systems in Europe took until the early 1991 and a substantial number of EWS have now been developed round the world (Yap, 1998).

Recently, a massive variety of early warning systems have been designed. There are two main Broad types to these EWS, as follows

1. The type of organization or institution they cover.
2. The type of factors or ratios that they adopted.

The old EWS basically concentrated on single banks level and were concerned with individual banking failures, but there are other EWS that consider the strength of the whole banking or financial sector and try to predict systemic crises. This research follows the modern type that deal with soundness of the whole banking sector of Sudan.

The original CAMELS model focuses on accounting and financial data for individual banks. Some researchers' also added macroeconomic data to process economic effect that could trigger a banking failure , Variables such as GDP growth, inflation, inter-bank profit rates or exchange rates are included to capture those effects.

(Flannery, 1998) was probably the first to add market information, driven by market forces and disciplines, he includes price-based indicators such as market expectations through stock prices, volatility of returns and bond spreads.

Market information-based approaches are only applicable for publicly listed and traded banks. However, in most countries, the majority of financial institutions are not publicly traded like the majority of deposits. There are other non-accounting or market information indicators and information that can be taken into account instead, such as rating agency assessments as recently applied in Sudan through the credit and rating agency as a partnership between a central bank and commercial banks, soon later was turned to an independent body. The idea is to capture the effects of risk and financial strength that are reflected in other indicators. There are also indicators of depositor behavior and bank credit ratings of ratings agencies considered.

2.2.1. CAMELS Factors

The Uniform Financial Institution Rating system, commonly referred to the acronym CAMEL rating, was officially sponsored by the Federal Financial Institution Examination Council on 1979. After many experiments it was known as a superior method to evaluate any financial institution, with assumption those institutions required special monitoring and supervision (Uniform Financial Institutions Rating System)

(Barr, 2002) states that "CAMEL rating has become a concise and indispensable tool for examiners and regulators", this rating ensures a bank's soundness conditions by revising various aspects of a bank based on diversity of information sources such as financial statement, funding sources, macroeconomic data, budget and cash flow.

According to (Mazzillo, 1993), CAMELS ratings are highly useful in identifying banks which need a large amount of supervisory attention.

CAMELS adopt the notion of composite rating that designed to take into account and reflect all major financial factors required by inspectors in their assessment of an institutions performance. Institutions are classified using a combination of specific financial ratios and inspector qualitative judgments.

The five ratings that constitute this composite rating can be summarized as follows:

Rating	**Description**
Strong Performance	Management clearly identifies all risks and employs compensating factors mitigating concerns
Satisfactory Performance	Management identifies most risks and compensates accordingly
Moderate Performance	Risk management practices may be less than satisfactory relative to the bank's or credit union's size, complexity, and risk profile
Poor Performance	Risk management practices are generally unacceptable relative to the bank or credit union's size, complexity and risk profile. Key performance measures are likely to be negative.
Unsatisfactory Performance	Critically deficient and in need of immediate remedial

	attention. Such performance, by itself or in combination with other weaknesses, directly threatens the viability of the bank or credit union.

Table 2.1: CAMELS Rating. (UNIFORM FINANCIAL INSTITUTIONS RATING SYSTEM ,1997)

In the next section we will introduce each component and its important ratios.

Capital Adequacy

Capital adequacy is the capital expected to maintain balance with the risks exposure of the financial institution such as credit risk, market risk and operational risk, in order to mitigate the potential losses and risks and prevent the financial institution‘s debt holder. “Meeting statutory minimum capital requirement is the key factor in deciding the capital adequacy, and maintaining an adequate level of capital is a critical element” (The United States, Uniform Financial Institutions Rating System 1997, p. 4).

Credit unions that are less than "adequately capitalized" must operate under an approved net worth restoration plan prepared by regulatory authorities. Inspectors examine capital adequacy by evaluating progress toward objectives set forth in the plan.

(Karlyn and Mitchell, 1984) define the capital adequacy in term of capital-deposit ratio because the primary risk is depository risk derived from the sudden and considerably large scale of deposit withdrawals. In 1930, Federal Deposit Insurance Corporation *(*FDIC) created a new capital model as capital-asset ratios since the default on loans came to expose the greatest risk instead of

deposit withdrawals. To measure the capital adequacy, bank supervisors currently use the capital-risk asset ratio. The adequacy of capital is assessed based upon the two most important measures such as Capital Adequacy Ratio (CAR) or Capital to Risk-weighted Assets.

Capital Adequacy Ratios

The capital adequacy ratio (CAR) is a measure of a bank's capital. It is expressed as a percentage of a bank's risk weighted credit exposures.

Also known as capital-to-risk weighted assets ratio (CRAR), it is used to protect depositors and promote the stability and efficiency of financial systems around the world. Two types of capital are measured: tier one capital, which can absorb losses without a bank being required to cease trading, A good example of a bank's tier one capital is its ordinary share capital. And tier two capital, which can absorb losses in the event of a winding-up and so provides a lesser degree of protection to depositors.

$$\text{Capital Adequacy Ratio (CAR)} = \frac{Tier\ One\ Capital + Tier\ Two\ Capital}{Risk\ Weghted\ Assets} \quad (2.1)$$

Risk Weighting Assets is fund based assets such as cash, loans, investments and other assets. Degrees of credit risk expressed as percentage weights have been assigned by the national regulator to each such assets , sometimes other regulatory authorities used Non-fund based which directly attached to the Off-Balance sheet items.

Asset Quality

An asset quality rating is an evaluation assessing the credit risk associated with a particular asset. These assets usually require profit margin payments - such as a loans and investment portfolios

According to (Grier, 2007), “poor asset quality is the major cause of most bank failures”. A most crucial asset category is the loan portfolio; the highest risk facing the bank is the risk of loan losses derived from the bad loans. The credit analyst should perform the asset quality assessment by carrying out the credit risk management and assessing the quality of loan portfolio using trend analysis and peer comparison. Gauging the asset quality is difficult because it is mostly derived from the analyst’s subjectivity.

Non-Performing Loans (NPL) ratios considered as distinctive predictor in assessing asset quality according to (Frost, 2004) as defined in usual classification system, loans include five categories: standard, special mention, substandard, doubtful and loss. NPLs are regarded as the three lowest categories which are past due or for which profit margin has not been paid for international norm of 90 days. In some countries regulators allow a slack time period, typically five to six months. The bank is regulated to back up the bad debts by providing adequate provisions to the loan loss reserve account. The allowance for loan loss to total loans and the provision for loan loss to total loans should also be taken into account to estimate thoroughly the quality of loan portfolio.

Asset Quality Ratios

A nonperforming loan is a loan that is either in default or close to default. A loan goes into default when a borrower fails to repay the loan according to the terms set forth in the loan contract. Ideally, loans become nonperforming when the debtor fails to make payments for (90 -180) days. An allowance for loan and lease losses, commonly called ALLL, is a bank's estimated credit losses, which is the amount of the loans the bank will unlikely be able to collect from the debtor.

$$\text{NPLs for total loans} = \frac{\text{NPLs}}{\text{Total Loans}} \qquad (2.2)$$

$$\text{Provision for loan loss ratio} = \frac{\text{Provison for loan loss}}{\text{Total Loans}} \qquad (2.3)$$

Management Quality

Management quality is all about the capability of the board of directors and top management, to identify, measure, and control the risks of an institution‘s activities and to ensure the safe, sound, and efficient operation in compliance with applicable laws and regulations (Uniform Financial Institutions Rating System 1997, p. 6).

(Grier, 2007) suggests that management is considered to be the single most important element in the CAMEL rating system because it plays a substantial role in a bank’s success; however, it is subject to measure as the asset quality examination.

It is difficult to measure the management quality component generally, and especially on the developing country, so this research is going to omit this

component due to the lack of raw data that can help to calculate such ratios, hopefully we can find a way in future to better and reliable evaluation.

Earning Ability

(Eccles , 2014) define Earnings are the net benefits of a corporation's operation. In accordance with Grier's opinion, a consistent profit not only builds the public confidence in the bank but absorbs loan losses and provides sufficient provisions. It is also important for a balanced financial structure and helps provide shareholder reward. So, consistently and progressively earnings are crucial to the stability of banking institutions. Profitability ratios measure the ability of a bank to generate profits from revenue and assets.

Earnings are also the amount on which corporate tax is due. For an analysis of specific aspects of corporate operations several more specific terms are used as EBIT -- earnings before interest and taxes, EBITDA - earnings before interest, taxes, depreciation, and amortization

This rating reflects not only the quantity and trend in earning, but also the criteria that may affect the stability of earnings. Incompetent management may result in loan losses and in return require higher loan allowance or pose high level of market risks. The future performance in earning should be given equal or greater value than past and present performance. (Uniform Financial Institutions Rating System 1997, p.7).

Earning Ability Ratios

The return on assets (ROA) ratio, or return on total assets ratio, relates a Bank's after tax net income during a specific year, to the Bank's average total assets during the same year and it can be defined also as an indicator of how profitable a Bank is relative to its total assets. ROA gives an idea as to how efficient management is at using its assets to generate earnings. Calculated by dividing a Bank's annual earnings by its total assets, ROA is displayed as a percentage. Sometimes this is referred to as "return on investment".

Since Bank assets' sole purpose is to generate revenues and produce profits, this ratio helps both management and investors see how well the bank can convert its investments in assets into profits. We can look at ROA as a return on investment for the bank since capital assets are often the biggest investment for most banks. In this case, the bank invests money into capital assets and the return is measured in profits.

The return on equity ratio (ROE) is a profitability ratio that measures the ability of a firm to generate profits from its shareholders investments in the bank, Return on equity measures a corporation's profitability by revealing how much profit an institution generates with the money shareholders have invested

$$\text{Returns on Assets (ROA)} = \frac{Net\ Income}{Total\ Assets} \qquad (2.4)$$

$$\text{Returns on Equity (ROE)} = \frac{Net\ Income}{Shareholder's\ Equity} \qquad (2.5)$$

Liquidity

The bank is classified in a strong liquidity position when well-developed funds-management practices are present, the bank has reliable access to sufficient sources of funds on favorable terms to meet present and anticipated liquidity needs

(Duttweiler, 2009) points out that "the liquidity expresses the degree to which a bank is capable of fulfilling its respective obligations". Banks makes money by mobilizing short-term deposits at lower profit margin rate, and lending or investing these funds in long-term at higher rates, so it is un-safe for banks mismatching their lending profit margin rate.

Liquidity Ratios

The liquidity coverage ratio (LCR) refers to highly liquid assets held by financial institutions to meet short-term obligations. The ratio is a generic stress test that aims to anticipate market-wide shocks. The liquidity coverage ratio is designed to ensure financial institutions have the necessary assets on hand to ride out short-term liquidity disruptions

$$\text{Liquidity Coverage Ratio (LCR)} = \frac{\text{Liquid Assets}}{\text{Deposits and Short term funds}} \qquad (2.6)$$

Sensitivity to Market Risk

Low risk is exposed to bank when Market-risk sensitivity is well controlled and there is minimal potential that the earnings performance or capital position will be adversely affected. Risk-management practices are strong for the size, sophistication, and market risk accepted by the institution. The level of earnings

and capital provide substantial support for the degree of market risk taken by the institution.

It is generally described as the degree to which changes in interest rates, foreign exchange rates, commodity prices, or equity prices can adversely affect earnings and/or capital. Market risk for a bank involved in credit card lending frequently reflects capital and earnings exposures that stem from changes in interest rates. These lenders sometimes exhibit rapid loan growth, a lessening or low reliance on core deposits

Sensitivity to Market Risk Ratio

One way to calculate the risk pertaining to market is use income related to finance to overall finance for the bank

$$\text{Finance Related Income Risk (FRI)} = \frac{\text{FInance Related Income}}{\text{Total Income}} \qquad (2.7)$$

2.2.2 Non- CAMELS factors:

There are other factors which can assist on predicting the bank's system distress, those factors are not included on financial and accounting statements of banks, therefor are used with specific orientation to predict the future status of banks in term of non-financial information, this research will try to propose a novel ratio other than CAMELS one to study its effect on bank's performance.

2.2.2.1. Macroeconomic Factors

Early studies on bank performance were conducted by (Short, 1979) and (Bourke, 1989) Then, in order to specify the causes of bank's distress, many

studies were held. In recent literature, the determinant of bank profitability is defined as a function of internal and external determinants. Internal determinants are connected to bank management and termed micro or bank specific determinants of profitability (GÜNGÖR, 2007). The external determinants (Non-Bank Specific) are reflecting economic and legal environment that affects the operation and performance of banks. There are different variables could be used. Among the internal determinants, there are bank specific financial ratios representing capital adequacy, cost efficiency, liquidity, asset quality which we have discussed on the former sections, furthermore there are some external factors such as size, Economic growth, inflation, and real interest rates are external determinants that affect bank performance, following we exhibit the threats and prediction power of each external factors.

Bank Size

Large banks have grown significantly in size and become more involved in market-based activities (those outside traditional bank lending) since the late 1990s. The advance of information technology and deregulation, which has led to a proliferation of financial markets, may have been the key driver of this process.

Large banks tend to have lower capital, less-stable funding, more market-based activities, and be more organizationally complex than small banks. This suggests that large banks may have a distinct, possibly more fragile, business model. Large banks are riskier, and create more systemic risk, when they have lower capital and less-stable funding. Large banks create more systemic risk

(but are not individually riskier) when they engage more in market-based activities or are more organizationally complex.

This has triggered a heated debate on the optimal size, organizational complexity, and range of activities of banks (Vinãls , 2013). This debate takes place against the backdrop of a financial landscape that has evolved markedly over the past two decades, spurred by financial innovation and deregulation. Large banks have increased in size, complexity, and involvement in market-based activities, as banking systems have grown in size and become increasingly global and interconnected.

Failures of large banks tend to be more disruptive to the financial system than failures of small banks. The failures of large banks generate liquidity stress in the banking system, their activities that rely on economies of scale and scope cannot easily be replaced by small banks, and the marginal cost of taxpayer support may increase in the volume required.

Economic Growth (GDP)

It is a measure of the total economic activity and it is adjusted for inflation. It is expected to have an impact on numerous factors related to the demand and supply for banks deposits and loans. According to the literature on the association between economic growth and financial sector profitability, GDP growth is expected to have a positive relation on bank profitability (Demirguc-Kunt, 1999); (Bikker, 2002) So generally a positive relationship is exist between bank profitability and GDP development as the demand for lending is increasing (decreasing).

(Calomiris, 1995) argue that firms may not properly anticipate how aggregate economic circumstances may affect the value and liquidity of their assets. As a result, firms may have a tendency to be excessively optimistic regarding their ability to avoid financial distress and therefore, take on excessive leverage during periods of economic expansion.

(Kent, 1997) provided that , In Australia, there is a long history of slumps in building activity leading to banking problems As the share of activity taken up by construction grows, therefore, the economy's overall credit quality is likely to decline. Particularly during the early 1990s a large proportion of banks problem loans were associated with the financing of commercial property. The sharp peak in the share of construction in GDP in early 1990 led the peak in impaired assets. The spike in construction activity in 1998 was not, however, reflected in an increase in impaired assets. This reflected, in part, the fall in the share of commercial property finance provided by banks as listed property trusts took on a greater role.

Inflation

This measures the overall percentage increase in Consumer Price Index (CPI) for all commodities and services. Inflation affects the real value of costs and incomes. The relationship between the inflation and profitability may have a positive or negative effect on profitability depending on whether it is anticipated or unexpected (Perry, 1992). If an inflation rate is anticipated, banks can adjust profit rate in order to increase incomes than expenses. At the contrast, if inflation rate is not expected, banks cannot make proper adjustments of profit rate that costs may increase faster than revenues. But most studies

observe a positive impact between inflation and profitability (Bourke, 1989); (Molyneux, 1992); (Hassan, 2003); (Kosmidou, 2006).

Real Interest Rate

Referring to previous studies, there is a positive relationship between interest rates and banks performance, bank profits increase with rising interest rates (Samuelson,1945).

Surprisingly, the link between monetary policy and bank profitability is an under-researched area. Many researchers analyze the link between bank profitability and business conditions, producing results on the link between the interest rate structure and bank profitability only as a by-product. In particular, (Demirguc-Kunt, 1999) were among the first to relate bank profits to macroeconomic indicators, such as real interest rates. They find that high real interest rates are associated with higher interest margins and profitability, especially in developing countries where demand deposits frequently pay below-market interest rates.

Recent examples from this strand of literature include (Albertazzi , 2009), who use aggregate data for the banking sector in 10 OECD (Organization for Economic Co-operation and Development) countries and find a significant relationship between net interest rate income and the yield curve slope. They also find a positive relationship between bank loss provisions and the short-term interest rate. (Bolt, 2012) obtain similar results using bank-level data and allowing for asymmetrical effects over the business cycle.

2.2.2.2. Other sectors

There could be factors in other sectors which can affect the banks performance such like factors on the agricultural sector for instance annual rain rate, indirectly affects the banks because they reduce farmer's productivity.

Other sectors or domain that might have the similar affects are political issues, social issues we can summarized as follow.

Sector	Factors	Affect
Agricultural sector	- Annual rain rate - Crops Diseases.	If the rain rate is small, banks might face troubles with farmer's obligations satisfaction, and ruined crops.
Political sector	- Wrong decisions. - Instability.	One wrong decision like changing the exchange rate policies in same area, resulted in public strikes might affect the banks situation.
Social Sector	- Consumer behavior - Customers awareness	Some customers have a common thought that finance shouldn't pay back specially in the micro finance operation; Customer's behavior might also affect the project's success.
Money Market	- Financial instruments	If for instance stock prices declined suddenly, it will affect banks

Table 2.2: Other sectors which may affect the bank's performance.

2.2.2. Discussion

(Duncan and Elliott, 2004) indicated that all financial performance measures such as interest margin return on assets, and capital adequacy as positively correlated with customer service quality. To assess the performance of the bank, it is necessary to report the financial reports which usually consist of a balance sheet, income statement, cash flow statement, and statement of changes in equity and notes to the financial statement (Salhuteru, 2015).

Providing a general framework in evaluating overall performance of banks is of great importance due to the increasing integration of global financial markets. CAMEL model reflects excellently the conditions and performances of banks over years as well as enriches the on-site and off-site examination to bring better assessments towards banks' conditions. Its purpose is to provide an accurate and consistent evaluation of a bank's financial condition and operations in the areas such as capital, asset quality, management, earning ability and liquidity. Some data from financial ratios are often compared to CAMELS in correctly identifying or predicting crisis events but sometimes, the relevant factors behind future failures or rating downgrades are satisfactorily captured by judiciously constructed risk sensitive summary statistics of conventional balance sheet data (Derviz, 2008)

By concentrating on the top line and bottom line, banks across the board have improved their profit while reducing their operational costs and more number of banks has improved their financial performance by using the concept of mergers and acquisitions. CAMEL rating is used by most banks across the world as a performance evaluation technique (Raiyani, 2010)

(Dzeawuni, 2009) claims that the strength of these factors would determine the overall strength of the bank. The quality of each component further underlines the inner strength and how far it can take care of itself against the market risks.

(Barker, 1993) find that the CAMEL system is useful, even after controlling for a wide range of publicly available information about the condition and performance of banks. This composite index further acts as a bank's failure predicting model. The rating is assigned based on both quantitative and qualitative information of the bank. If a bank's index is less than two, it is regarded as a high-quality bank, whereas institutions with grade four or five are rated to be insolvent (Curry, 2009) the up-to-date examination ratings help identify if the banks require increased supervisory attention well before they actually fail. Although (Gaytán, 2002) argues that the model is only parallel with the performance of the bank at the time of the examination, while variables in banks are highly volatile to market forces; the CAMEL model is still very much popular among regulators due to its effectiveness.

Capital is divided to two types, Tier 1 capital consists of shareholders' equity and retained earnings. Tier 1 capital is intended to measure a bank's financial health and is used when a bank must absorb losses without ceasing business operations. Under international accords such like BASEL III[2], the minimum tier 1 capital ratio is 10.5%, which is calculated by dividing the bank's tier 1 capital by its total risk-based assets, Tier 2 capital includes revaluation reserves, hybrid capital instruments and subordinated term debt, general loan-loss reserves, and undisclosed reserves. Tier 2 capital is supplementary capital because it is less reliable than tier 1 capital. In 2017, under Basel III, the minimum total capital

[2] "Basel III" is a comprehensive set of reform measures, developed by the Basel Committee on Banking Supervision, to strengthen the regulation, supervision and risk management of the banking sector.

ratio is 12.5%, which indicates the minimum tier 2 capital ratio is 2%, as opposed to 10.5% for the tier 1 capital ratio.(Bank for International Settlements , 2010).

IDC, which is an international bank rating agency, has been rating the safety and soundness of banks, savings institutions, and credit unions since 1985 the most important factors which are affect bank's performance listed as following order:

1. Capital risk is determined by Tier I capital as a percent of assets and as a percent of risk-based assets. Tier I & II capital as a percent of risk-based assets (risk-based capital ratios) measure credit and interest rate risk as well as estimate risk in the asset base. Government standards should be used, as well as, enforcement actions to evaluate well, or less than well, capitalized institutions.
2. Adequacy of capital and reserves measures asset quality as the levels of delinquent loans, nonaccrual loans, restructured and foreclosed assets relative to loan loss reserves and Tier I capital.
3. Margins are the best measurement of management's financial controls.
4. Earning returns measure the success of the bank's operating strategy. Ratios of revenue yields from investments, loans, and noninterest income compared to operating costs before interest expense are the major components of the after-tax net operating return on earning assets (ROEA). ROEA is a measure of operating strategy as if the institution was wholly funded by equity capital. Earnings from financial leverage (ROFL) measures the level of leverage and after-tax cost of funding compared to the after-tax return on

earning assets (ROEA). Leverage returns measure the efficiency of the bank's financial strategy.

5. Liquidity measures (i) balance sheet cash flow as a percent of Tier I capital and (ii) loans compared to stable deposits and borrowings

This research will adopt different method on selecting the best factors that will be used to predict the distress of Sudanese banking system. Answers from subject matter experts will be analyzed to determine which in their opinion are the most crucial CAMELS ratio, because they experienced many distressed banks and they had to know the best factors on the banks financial and accounting data. Also use the dimensionality reduction method in all available CAMELS ratios to produce a new set of CAMELS ratios it might have different prediction power than the former ones.

2.3. Bank's Distress Prediction Methodologies

Generally the process of predicting the financial distress of a bank composed from three important steps firstly, factors selection that represent the input elements for the second step which is the classification model that predict the future behavior of the bank due to the pattern identified on the input vector and eventually the evaluation step which basically determine the reliability and credibility of the selected classification model.

On next sections are organized to follow the design process of bank distress prediction model, section 2.3.1 will cover the feature selection techniques, and section 2.3.2 will cover classification techniques that have been found on the literature of bank's distress prediction, section 2.3.3 will cover evaluation methods and finally on section 2.3.4 we will have a discussion.

2.3.1. Dimensionality Reduction and Feature Selection

When dealing with huge volumes of data, a problem naturally arises. How can we cut down a dataset of hundreds of variables into an optimal model and how can we visualize data through numerous dimensions. Fortunately, a series of techniques called dimensionality reduction aim to help ease these issues. These techniques help to identify sets of uncorrelated predictive variables that can be fed into subsequent analyses.

Reducing a high-dimensional data set, i.e. a data set with many predictive variables, to one with fewer dimensions improves conceptualization. Above three dimensions, visualizing the data becomes difficult or impossible. Additionally, reducing dimensions helps to reduce noise by removing irrelevant or highly correlated variables.

There are two types of dimensions reduction techniques that are found in the literature depends on the data used: Linear or non-linear dimensions reduction.

2.3.1.1. Linear Dimensions Reduction

Linear dimensionality reduction methods are a cornerstone of analyzing high dimensional data, due to their simple geometric interpretations and typically attractive computational properties. These methods capture many data features of interest, such as covariance, dynamical structure, and correlation between data sets, input-output relationships, and margin between data classes. Methods have been developed with a variety of names and motivations in many fields.

Linear dimensionality reduction methods have been developed throughout statistics, machine learning, and applied fields for over a century, and these

methods have become indispensable tools for analyzing high dimensional, noisy data. These methods produce a low-dimensional linear mapping of the original high-dimensional data that preserves some feature of interest in the data. Accordingly, linear dimensionality reduction can be used for visualizing or exploring structure in data, de-noising or compressing data, extracting meaningful feature spaces, and more. This abundance of methods, across a variety of data types and fields, suggests a great complexity to the space of linear dimensionality reduction techniques. As such, there has been little effort to consolidate our understanding. Here we survey a host of methods and investigate when a more general optimization framework can improve performance and extend the generality of these techniques.

Principal component analysis (PCA) is a classical method that provides a sequence of best linear approximations to a given high-dimensional observation. It is one of the most popular techniques for dimensionality reduction.

Given a set of data on n dimensions, PCA aims to find a linear subspace of dimension d lower than *n* such that the data points lie mainly on this linear subspace such a reduced subspace attempts to maintain most of the variability of the data.

The linear subspace can be specified by *d* orthogonal vectors that form a new coordinate system, called the `principal components'. The principal components are orthogonal, linear transformations of the original data points, so there can be no more than *n* of them.

PCA has major two constraints. Firstly, PCA can only identify linear combinations of variables; that is to say, it can only determine linear interdependencies between components of a sample of data vectors. In a word, it can only give a linear mapping from some data space to some low dimensional code space. Secondly, since we obtain some principal axes by solving the characteristic equation of a covariance matrix, PCA can only capture the second-order correlation information between components of the data but ignores the higher-order correlation information among components of the data.

The principal components are also sensitive to the scale of measurement, now to fix this issue we should always standardize variables before applying PCA.

PCA can be defined mathematically as follow

The transformation is defined by a set of p-dimensional vectors of weights or loadings $W_{(k)} = (w_1, w_2 \dots, w_p)_{(k)}$ that map each row vector $x_{(i)}$ of x to a new vector of principal component scores $t_{(i)} = (t_1, t_2 \dots, t_m)_{(i)}$, given by $t_{k(i)} = x_{(i)}.w_{(k)}$ for $i = 1, \dots . n \quad k = 1, \dots . m$ in such a way that the individual variables of t considered over the data set successively inherit the maximum possible variance from x, with each loading vector w constrained to be a unit vector.

2.3.1.2. Nonlinear Dimensions Reduction

Some models for dimensionality reduction have been introduced in chapter 1. However, these models can only generate linear mappings from the high dimensional space to the low dimensional space and cannot find non-linear

structure in the data as discussed in previous section. Research on non-linear dimensionality reduction methods has been explored extensively in the last few years. In the following, a brief introduction to several non-linear dimensionality reduction techniques will be given.

Locally Linear Embedding

Locally linear embedding (LLE) is approach which addresses the problem of nonlinear dimensionality reduction by computing low-dimensional, neighborhood preserving embedding of high-dimensional data. A data set of dimensionality n, which is assumed to lie on or near a smooth nonlinear manifold of dimensionality $d < n$, is mapped into a single global coordinate system of lower dimensionality, d. The global nonlinear structure is recovered by locally linear fits.

Through this algorithm, every data point is mapped into the low dimensional subspace, in which the high dimensional dot products between the edges in every data point's neighborhood are preserved as well as possible by the low dimensional dot products. By keeping information from overlapping local neighborhoods, the global structure of the whole data is maintained, provided that the local neighborhoods are sufficiently connected. As has been mentioned above, LLE focuses on generating low dimensional codes preserving local linear geometry in the data. The reconstructions here are different from the ones in PCA.

The algorithm does not try to use the low dimensional codes to reconstruct the original data. LLE doesn't care how well each particular low dimensional code represents the original data point.

LLE computes the barycentric coordinates of a point x_i based on its neighbor x_j The original point is reconstructed by a linear combination, given by the weight matrix w_{ij} of its neighbors. The reconstruction error is given by the cost function E(W).

$$E(w) = \sum_i \left| x_i - \sum_j w_{ij} x_j \right|^2 \qquad (2.8)$$

ISOMAP

The idea of ISOMAP (Tenenbaum, 2000) is to map some high dimensional data to a non-linear low dimensional subspace in a way that preserves a particular kind of structure in the data. It is assumed that the data lies on, or near, a lower dimensional manifold that is embedded in the high dimensional space. The Geodesic Distance between two data points is defined as the shortest distance along the manifold and the aim is to find a low-dimensional representation that preserves the Geodesic distances as near as possible. The Geodesic distances can be estimated by finding shortest paths in a neighborhood graph derived from the data. The neighborhood graph can be constructed by connecting each data point to its k nearest neighbors.

Here is a sketch of the ISOMAP algorithm: (1) Construct the neighborhood graph; (2) Calculate Geodesic Distances between every two data points using a shortest path algorithm; (3) Given these pairwise distances, use MDS (Linear Approach) to find low dimensional codes that preserve the pairwise Geodesic distances as well as possible. ISOMAP generates low dimensional codes that preserve the non-linear geometry of the data by preserving the Geodesic Distances between every two data points. It does not try to find codes that are optimal for reconstructing the individual data points.

Kernel Principle Component Analysis

Kernel PCA (Scholkopf , 1998) is a kernel version of standard PCA. When used for dimensionality reduction, standard PCA generates low dimensional codes preserving most of the variance in the original data. But sometimes the data is not separable even in the original high dimensional space no matter how we redefine the axes of the data. We hope that the data can be separable in a higher dimensional space and then we can perform standard PCA in that space. Kernel PCA can help us to achieve this purpose while doing the calculations in a lower dimensional space by means of the kernel trick.

Factor Analysis

The broad purpose of factor analysis (FA) is to summarize data so that relationships and patterns can be easily interpreted and understood. It is normally used to regroup variables into a limited set of clusters based on shared variance. Hence, it helps to isolate constructs and concepts. Factor analysis uses mathematical procedures for the simplification of interrelated measures to discover patterns in a set of variables (Child, 2006). Attempting to discover the simplest method of interpretation of observed data is known as parsimony, and this is essentially the aim of factor analysis (Harman, 1976).The two main factor analysis techniques are Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). CFA attempts to confirm hypotheses and uses path analysis diagrams to represent variables and factors, whereas EFA tries to uncover complex patterns by exploring the dataset and testing predictions (Child, 2006). The following table summarizes the dimensionality reduction techniques that we covered so far along with their characteristics.

Method	Linear(L) /Non-Linear(NL)	Remarks
PCA	L	Preserves most variance of the data; Can only capture second-order correlations between components of the data vectors
Kernel PCA	NL	Improvement for PCA
ISOMAP	NL	Preserves the global manifold structure of the data by preserving the Geodesic distances between every two data points
LLE	NL	Preserves the locally linear structure in the data
FA	NL	summarize data so that relationships and patterns can be easily interpreted

Table 2.3: Dimensionality reduction techniques.

After studying all dimension non-linear dimensionality reduction methods, PCA outperformed all other techniques in numerous related studies Fodor and (Kamath, 2002), (Orsenigo, 2013). However PCA has three key disadvantages (1) The covariance matrix is difficult to be evaluated in an accurate manner, (2) Even the simplest invariance could not be captured by the PCA unless the training data explicitly provides this information, (3) dimensionality of the feature vectors is minimized by minimizing the reconstruction error between the original vector and the reconstructed vector.

This research will try to study different types of filter-based and wrapper based features selection techniques and proposes a hybrid filter-wrapper technique.

2.3.2. Core Feature Extraction Techniques of Proposed Method

2.3.2.1 Discriminant Analysis

Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier or, more commonly, for dimensionality reduction before later classification.

This technique was originally introduced in the biological science by (Fisher, 1936), who distinguished three species of Iris flowers based on group characteristics.

The LDA is one of the most useful technique to discriminate and predict corporate bankruptcies, in particular when there are solely quantitative predictors. Compared to other techniques such as logistic regression, classification trees, neural networks, and others, LDA has the advantages to be robust against a certain degree in the loss of assumptions (Janer, 2011), to relatively resist to temporal changes and to offer judicious possibilities of interpretation.

Once data are gathered, two statistical analysis steps will be carried out: descriptive and elaborative. The first step determines the separation representation between the preexisting groups based on the training data. The

elaborative step consists in elaborating the decision rule to classify new objects (banks).

The LDA can be implemented according to two decision rule approaches: (i) geometrical and (ii) probabilistic.

Geometrical approach

The geometrical approach exemplifies on a metrical rule to separate at best preexisting groups (distressed and not-distressed banks) in a Cartesian coordinate system. The separation points will bring closer the representative points of the banks of the same group and set apart the representative points of the banks from different groups. Therefore, the separator hyperplane maximizes inter-groups variances (between) and minimizes intra-groups variances (within).

In the case of two groups i (distressed) and j (not-distressed), the optimal separator hyperplane under the metric criterion represented as follows:

$$(u_i - u_j)' M(x - \frac{(u_i + u_j)}{2} = 0 \ , \forall i \neq j \qquad (2.9)$$

Where u_i and u_j are the means of groups *i* (distressed) and *j* (not distressed). *M* is a metric used to measure the relative distance. *M* is often used as the inverse total covariance matrix or as the inverse intra-class covariance matrix. *x* is the vector of the *k* financial ratios of the bank.

This illustration implies an understanding of distance to assess the relative closeness and distance of the point cloud. This distance is indispensable in classifying new banks. For example, if a new object (bank) for which the

descriptive variables as financial ratios are known but the classification group is unknown, the geometrical rule of classification affects this bank into the group whose average point is the closest to the representative point of the bank. The new bank characterized by *x* is affected into the group *i* if and only if the distance d(*x*, u_i) is severely lower to the distance *d(x*, u_j*)*, such as:

$$d(x, u_i) < d(x, u_j), \forall i, j \in \{1,2, \dots g\} \text{ and } i \neq j \qquad (2.10)$$

However, this approach, purely geometrical, does not consider the a priori probabilities of the different groups and their potential cost of misclassification, while the probabilistic classification approach offers such possibilities which will be addressed in the next section

Probabilistic approach

In probabilistic models, each observation *x* of the training data is no longer considered as a Cartesian coordinate but as the realization of the object's description. Each different group of object is considered by its a priori probability of appearance, which limits the possibility of appearance of different objects to be classified.

Knowing the group membership and description of objects *x*, it is possible to estimate the probability that a particular description is realized as the group *i* of the object, such as: *p(x\i)* with $i \in \{1,2, \dots g\}$. Therefore, it is assumed that certain descriptions have more chances to be realized for some groups than others based on their distributional differentiation. It is equivalently assumed that both groups have proper characteristics and that each object (bankrupt or

non-bankrupt bank) presenting these characteristics are affected in the same classification group.

The decision criterion delimits the separation between the studied groups. Therefore, with the probability *p(x\i)* it becomes possible to classify banks according to their financial ratios or descriptions *x* in the group for which the probability that this description is achieved is maximum; leading to the rule that the bank of ratio *x* is affected to group *i* if and only if the probability *p(x\i)* is strictly superior than the probability *p(x\j)* for all $j \in \{1,2, \dots g\}$ and with $i \neq j$, such as:

$$p(x\backslash i) > p(x\backslash j) \ , \forall i, j \in \{1,2, \dots g\} \text{ and } i \neq j \qquad (2.11)$$

However, it would be desirable to get the probability of belonging to a group undergoing the description of interest, leading to the rule: $p(x\backslash i) > p(x\backslash j) \ , \forall i, j \in \{1,2, \dots g\}$ and $i \neq j$, rather than that a particular description is recognized in a given group such as in the last equation, Bayes' theorem is used to facilitate this interchange as follows:

$$\rho(i|x) = \frac{\rho(x|i).\rho(i)}{\sum_{n=1}^{g} \rho(x|n).\rho(n)} \qquad (2.12)$$

$$\rho(j|x) = \frac{\rho(x|j).\rho(j)}{\sum_{n=1}^{g} \rho(x|n).\rho(n)} \qquad (2.13)$$

Thus equation 2.12 can be re written like

$$\rho(i|x) = \frac{\rho(x|i).\rho(i)}{\sum_{n=1}^{g}\rho(x|n).\rho(n)} > \rho(j|x) = \frac{\rho(x|j).\rho(j)}{\sum_{n=1}^{g}\rho(x|n).\rho(n)} \qquad (2.14)$$

This leads to

$$\rho(x|i).\rho(i) > \rho(x|j).\rho(j)\ \forall \mathrm{i} \neq \mathrm{j} \qquad (2.15)$$

Accordingly, the classification rule consists of make the most of the probability that an object belongs to a group according to its descriptions.

Quadratic Discriminant Analysis

There is no assumption with quadratic discriminant analysis that the groups have equal covariance matrices. As with linear discriminant analysis, an observation is classified into the group that has the smallest squared distance. However, the squared distance does not simplify into a linear function, thus the name quadratic discriminant analysis.

Unlike linear distance, quadratic distance is not symmetric. In other words, the quadratic discriminant function of group i assessed with the mean of group j is not equal to the quadratic discriminant function of group j assessed with the mean of group i. On the results, quadratic distance is called the generalized squared distance. If the determinant of the sample group covariance matrix is less than one, the generalized squared distance can be negative.

2.3.2.2 Genetic Algorithm (GA)

GAs are computational models of evolution. They work on the heart of a set of candidate solutions. Each candidate solution is called a “chromosome”, and

the whole set of solutions is called a “population”. The algorithm allows movement from one population of chromosomes to a new population in an iterative fashion. Each iteration is called a “generation”. There are various forms of GAs, a simple version, which is called static population model (Whitley, 1989) the population is ranked according to the fitness value of each chromosome. At each generation, two (and only two) chromosomes are selected as parents for reproduction.

Performing feature selection with GAs requires conceptualizing the process of feature selection as an optimization problem and then mapping it to the genetic framework of random variation and natural selection. Individuals from a given generation of a population mate to produce offspring who inherit genes (chromosomes) from both parents. Random mutation alters a small part of child’s genetic material. The children of this new generation who are genetically most fit produce the next generation. In the feature selection context, individuals become solutions to a prediction problem. Chromosomes (sequences of genes) are modeled as vectors of 1’s and 0’s with a 1 indicating the presence of a feature and a 0 its absence. The simulated genetic algorithm then does the following: it selects two individuals, randomly chooses a split point for their chromosomes, maps the front of one chromosome to the back of the other (and vice versa) and then randomly mutates the resulting chromosomes according to some predetermined probability.

High level algorithm for GA work is shown as follows:

1. Define stopping criteria, population size, P, for each generation, and mutation probability, p_m
2. Randomly generate an initial population of chromosomes
3. repeat:
4. | for each chromosome do
5. | Tune and train a model and compute each chromosome's fitness
6. | end
7. | for each reproduction 1 ... P/2 do
8. | Select 2 chromosomes based on fitness
9. | Crossover: randomly select a locus and exchange genes on either side of locus
10. | (head of one chromosome applied to tail of the other and vice versa)
11. | to produce 2 child chromosomes with mixed genes
12. | Mutate the child chromosomes with probability p_m
13. | end
14. until stopping criterion are met

Algorithm 2.1: GA overview (Max Kuhn, 2013)

2.3.3. Classification Models

The strong relationship between bank failures and economic growth makes the predictability of bank failures even more important. Several statistical prediction models exist in the literature. Besides, artificial intelligence techniques began to gain a greater importance in the literature due to their

predictive success. The models used in the prediction of bank failures are divided into two main classes of methods, namely statistical methods and intelligent methods. The studies focusing on the prediction of bank failures by the assistance of statistical methods date back to 1970s, whereas studies employing intelligent methods originated in 1990s.

The statistical methods comprise of (linear, multivariate and quadratic) discriminant analysis, factor analysis and logistic regression methods. On the other hand, intelligent methods comprise of artificial neural networks, evolutionary approaches, operations research, hybrid intelligent methods, fuzzy logic and SVMs etc.

2.3.3.1. Statistical Models

The early studies concerning ratio analysis for bankruptcy prediction are known as the univariate studies. These studies consisted mostly of analyzing individual ratios, and sometimes, of comparing ratios of failed companies to those of successful firms. However, few studies were published up to the mid-60s. This period is known as a relatively rich in published studies of corporate failures, in which academics advanced further in the field.

Meta-Analysis for statistical models reviewed in the literature is given in Table 2.4.

In particular, (Beaver, 1966) studied the predictive ability of accounting data as predictors of major events. His work was intended to be a benchmark for future investigations into alternative predictors of failure.

Technique	Usage Description	Reference (Authors)	Period
Logit	macroeconomic and structural level	Männasoo and Mayes (2005)	1996-2003
MDA	institutional characteristic variables	Jordan, Rice, Sanchez, Walker, and Wort (2010)	2007-2012
DEA (LOGIT)	Economic and financial variables	Tatom and Houston (2011)	2006-2011
HAZARD	macroeconomic and structural indictors	Männasoo and Mayes (2009) Jin,	1995-2004
LOGIT	accounting, audit quality and financial variables	Kanagaretn , and Lobo (2011)	2007-2010
DEA	financial market information such as credit ratings	Avkiran and Cai (2012)	2004-2009

Table 2.4: Statistical models meta-analysis.

Beaver found that a number of indicators could discriminate between matched samples of bankrupt and non-bankrupt firms for as long as five years prior to failure. In a real sense, his univariate analysis of a number of bankruptcy predictors set the stage for the development of multivariate analysis models.

Two years later, the first multivariate study was published by (Altman, 1968). With the well-known "Z-score", which is a multiple discriminant analysis (MDA) model, Altman demonstrated the advantage of considering the entire profile of characteristics common to the relevant firms, as well as the interactions of these properties. Specifically, the usefulness of a multivariate model taking combinations of ratios that can be analyzed together in order to consider the context or the whole set of information at a time compared to univariate analysis that study variables one at a time and tries to gather most information at once. Consequently to this discriminatory technique, Altman was able to classify data into two distinguished groups: bankrupt and non-bankrupt firms. He also demonstrated a second advantage: if two groups were studied this analysis reduces the analyst's space dimensionality to one dimension.

(Martin, 1977) also presented a logistic regression model (logit) to predict probabilities of failure of banks based on the data obtained from the Federal Reserve System data sample. Martin was then followed by (Ohlson, 1980) who developed a logistic regression model, logit model or logit analysis (LA), to predict bankruptcies. He principally argued the MDA approach in regards of three following points: (i) the MDA technique relies too much on assumptions, (ii) the MDA outputs score do not provide intuitive interpretation, but however agreed that if a priori probabilities are known, then it becomes possible to derive a posteriori probabilities of bankruptcy, and (iii) Ohlson pinpointed the

discriminant selection process for its relative subjectivity. On the other hand, according to Ohlson, the use of conditional logit analysis avoids all of the problems discussed above. In particular, he underlines as major advantages that in logit model there is no need for assumptions to be made regarding a priori probabilities of bankruptcy, and for the distribution properties of the predictors. This approach model is in particular interesting because it allows the practitioner to test the significance of the predictors as it is presented in the assumptions' test part of this study.

(Altman, 2011) estimated the likelihood of default inferred from equity prices, using accounting-based measures, firm characteristics and industry-level expectations of distress conditions. This approximately enables timely modeling of distress risk in the absence of equity prices or sufficient historical records of default. Model's results are comparable to that of default likelihood inferred from equity prices using the Black-Scholes-Merton structure. Finally, Altman et al. emphasized the importance of treating equity-implied default probabilities and fundamental variables as complementary rather than competing sources of predictive information. (Sun, 2011) tested the feasibility and effectiveness of dynamic modeling for financial distress prediction (FDP) based on the Fisher discriminant analysis model. They designed framework of dynamic FDP based on various instance selection methods, such as full memory window, no memory window, window with fixed size, window with adaptable size, and batch selection. They also utilized initial features set composed of seven aspects of financial ratios and proposed a wrapper integrating forward and backward selection for the dynamic modeling of FDP. Findings indicated that dynamic models can perform better than static models and should be further developed to other classification techniques. (Betz, 2014) tested the logit model in

European banks, their results show that an early-warning model based on publicly available data yields useful out-of-sample predictions of bank distress during the global financial crisis.

Based on similar recent studies which focused on statistical approaches that advise about efficiency of DA (Janer, 2011), this research will use discriminant analysis (DA) method to develop the statistical model of Sudanese bank's distress prediction. Various related studies (Janer, 2011) support the developed model and study as well as the rate of good classification is equal to 86.36% of the holdout sample. Type I and II errors are in equivalent proportions after being rebalanced with a cut-off modification achieved by nonlinear programming optimization. Various testing of the model robustness are performed, such as logistic regression, which confirms the significance of the most of the explanatory variables.

2.3.3.2. Intelligent Models

In the following table, the summary of intelligent techniques that implemented on predicting the banks financial distress will be presented along with their advantages and disadvantages.

Technology	Basic Idea	Advantages	Disadvantages	Reference
NN	Learn from examples using several constructs and algorithms just like a human being learns new things	Good at function approximation, forecasting, classification, clustering and optimization tasks depending on	Good at function approximation, forecasting, classification, clustering and optimization tasks depending on the neural network architecture	(Angelini,2008)

		the neural network architecture		
GA	Mimics Darwinian principles of evolution to solve highly nonlinear, non-convex global optimization problems	Good at finding global optimum of a highly nonlinear, non-convex function without getting trapped in local minima	Does take long time to converge; May not yield global optimal solution always unless it is augmented by a suitable direct search method	(Varetto, 1998)
CBR	Learns from examples using the euclidean distance and k-nearest neighbor method	Good for small data sets and when the data appears as cases; similar to the human like decision making	Cannot be applied to large data sets; poor in generalization	(Angela, 2003)
Rough sets	They use lower and upper approximation of a concept to model uncertainty in the data	They yield 'if–then' rules involving ordinal values to perform classification tasks	It can be (a) sometimes impractical to apply as it may lead to an empty set (b) sensitive to changes in data and (c) inaccurate	(Nursel, 2008)
SVM	It uses statistical learning theory to perform classification and regression tasks	It yields global optimal solution as the problem gets converted to a quadratic programming problem; It canwork	It is abysmally slow in test phase. It has high algorithmic complexity and requires extensive memory	(Ribeiro, 2012)

		wellwith few samples		
Decision Trees	They use recursive partitioning technique and measures like entropy to induce decision trees on a data set	Many of them can solve only classification problems while CART solves both classification and regression problems. They yield human comprehensibl e binary 'if–then' rules	Over fitting can be a problem. Like neural networks, they too require a lot of data samples in order to get reliable predictions	(Adrian, 2015)

Table 2.5: Sample of Intelligent Techniques

In the 1990s, advances in both computer speed and power and new developments in artificial intelligence and expert systems (AI&ES) software programming gave rise to a new family of failure prediction models and methods. In general, the AI&ES software programs are designed to learn from both their input data and their previous experience in solving a particular problem. In the case of bankruptcy or default prediction applications, the problem is the proper classification of a set of firm data into the proper category. The software systems actually learn and improve their prediction and categorization abilities through an iterative process as they explore the nonlinear relationships between the input data as provided by (Adnan Aziz, 2006).

Several classification methods have been applied in order to predict bankruptcy. These methods include Neural Networks (NNs), Decision Trees (DTs), Support Vector Machines (SVMs), k Nearest Neighbor (kNN), Genetic Algorithms (GAs) etc. Research studies compare the capabilities of these methods in predicting bankruptcy and often focus on the development of new, more elaborated methods of bankruptcy prediction for specific data sets pertaining to different economic system

Intelligent methods are increasingly preferred instead of statistical methods as a virtue of their superior predictive success with respect to the prediction of bank failures. One of the first studies in which the prediction of bank failures was performed via various statistical methods and then compared against ANNs was conducted by (Tam, 1991). Tam attempted to predict the bank failures in the state of Texas a year or two in advance by the assistance of various methods. Consequently, Tam concluded that the predictive power of back-propagation artificial neural networks (BPANN) model proved to be superior to any of the discriminant analysis, factor analysis, logistic analysis, and k-nearest neighbor algorithm also used in his study.

(Tam, 1992) compared the predictive powers of linear discriminant analysis, logistic regression, K-nearest neighbor analysis, Interactive Dichotomizer 3 (ID3), forward-feed artificial neural networks (FFANN), and BPANN models. Accordingly, they similarly concluded that the BPANN model generated the most favorable results but similar to some extent for LDA accuracy.

The most common (AI&ES) approach is that of a neural network (NN) that is designed to replicate the learning and classification processes carried out by the human brain. The NN approach is essentially nonparametric and nonlinear in

the model building process. While some of the earliest applications of a NN system for failure prediction in the finance literature appeared in the early 1990s as explained by (Coats, 1993) and (Altman, 1994).

(Zhang, 1999) provide a clear discussion of the general framework of the NN approach and the underlying statistical theory is related to the theory supporting the use of MDA. They provide an excellent literature review of NN applications to bankruptcy prediction in the early literature, and they develop their own NN model using COMPUSTAT data - Compustat is a database of financial, statistical and market information on active and inactive global companies throughout the world - They also compare the performance of their NN model to a LOGIT model developed on the same dataset and find that the NN model outperforms the LOGIT approach.

(Adrian, 2015) use a semi-parametric Cox survival analysis model and non-parametric CART decision trees to financial distress prediction and compared with each other as well as the most popular approaches. The analysis is done over a variety of cost ratios (Type I Error cost: Type II Error cost) and prediction intervals as these differ depending on the situation. The results show that decision trees and survival analysis models have good prediction accuracy that justifies their use and supports further investigation.

(Atiya, 2001) provides another excellent survey of the literature that is more recent. He focuses his discussion on the most common (AI&ES) approach, that of a multilayered NN, to the prediction of bankruptcy and financial distress. He also provides a discussion on the reasons an NN approach is superior to more traditional statistical approaches based on the nonlinear characteristics of the input data typically used in prediction studies and he presents results of his own

NN models that incorporate both financial and equity market information into the analysis.

(Benli, 2005) comparatively used both logistic regression and ANNs methods in the prediction of bank failures. The author recorded that with respect to general classification success, the accurate classification rates for the ANNs model and logistic regression model are 87 and 84.2%, respectively. Additionally, the prediction accuracy for the ANNs model concerning the failed banks is 82.4%, while the corresponding rate is 76.5% for the logistic regression model. As a result, it was inferred that the ANNs model possesses a superior predictive power than logistic regression model with respect to financial failures.

(Angelini, 2008) also provide an excellent discussion of the basics of NN design and operation (for the non-programmer) and they also develop a NN system and apply it to a sample of small firm data obtained from an Italian bank. The data represent 76 small firm clients of the bank, and the focus of the study is the prediction of bank loan defaults. Their final NN model produces a very low classification error rate that is much lower than is typical with classical statistical approaches.

(AVCI, 2008) investigated the applicability of CAMELS rating system in the supervision of Turkish Commercial Banking system by studying those banks devolved to savings deposit insurance fund (SDIF) as well as those that remained solvent for the period of 1996–2001. They employed discriminant analysis, logistic regression and ANNs models in their study. The authors then recorded that in light of the obtained results, it could be stated that ANNs models yielded better results relative to the discriminant analysis and logistic

regression models; however, the findings were still far from being satisfactory considering the low level of correct classification rates.

(Angela, 2003) reviews different approaches and presents a framework of a case-based reasoning (CBR) approach to business failure prediction by integrating two techniques, namely nearest neighbor and induction. It is unrealistic to assume that all attributes are equally important in the similarity function of nearest neighbor assessment

Although the concept of fuzzy sets was introduced by Zadeh as early as 1965, the use of fuzzy logic in predicting business bankruptcies was practically unknown until 2006. Since 2007 few papers were published describing the possibility of implementing such fuzzy system in forecasting this negative phenomenon in firms. According to (Virbickaite, 2008) and (Balcaen, 2006), the literature does not provide a clear image regarding the applications of alternative methods (including strategies based on expert systems) used for business failure prediction and therefore further research is necessary. According to (Virbickaite, 2008) and (Huangb, 2009), failure prediction methods based on fuzzy logic are more useful to managers than methods based on neural networks that are hardly interpretable (an explanation of a specific forecast cannot be provided using neural networks). Neural networks are useful to refine the knowledge base of the expert system when it is necessary.

While neural network approaches are the most common model used, there are other (AI&ES) approaches that have been explored, (Lin, 2009) proposed a selective ensemble which combined Decision Trees, Back-propagation NN and SVM. The authors introduced the notion of expected probability which is the trend of a classifier to predict bankruptcy or non-bankruptcy. The three

classifiers were combined in a tree structure. For reaching a decision, priority was given when a classifier decided contrary to its trend.

(Nursel, 2008) focused on the Turkish banking sector, and after reviewing a number of quantitative tools, selected to apply the Rough Set Theory (RST) approach to analyze the failures of banks during the 1995-2007 period. The data for the financial ratio analysis for the 41 banks investigated from the publicly available sources. The results showed that early warning systems based on statistical models can effectively be used to predict bank failures.

(Etemadi, 2009) employed Genetic Programming for the prediction of bankruptcy. Genetic Programming is based on the application of Genetic Algorithms to a population of Computer Programs i.e. tree structures that represent mathematical expressions. The genetic operators modify terminal nodes, functions or even sub trees.

(Yeh, 2011) combined the ISOMAP algorithm for feature selection with the SVM classifier. ISOMAP is a method that computes a low-dimensional embedding of high dimensional data points. As opposed to Principal Components Analysis, ISOMAP is a non-linear method and the data are mapped to a global coordinate system in which geodesic distances are calculated. For neighboring points the geodesic distances are approximated by the Euclidean distances, while for distant points the geodesic distances are approximated by the length of the shortest path in a weighted graph with edges connecting nearby points. The top d eigenvectors of the distance matrix are the coordinates of the new d-dimensional Euclidean space.

(Ribeiro, 2012) shown that bankruptcy of small and medium size companies can be accurately predicted if a detailed training dataset is available, specifically up to 3 years prior to the analysis. Of all the models tested (LOGIT, MLP, SVM) Support Vector Machines achieved the best performance.

From the literature review of applied intelligent techniques (Ribeiro, 2012) we discovered a promising result if SVM method is implemented along with other feature selection technique, in this research, factor analysis and SVM will joint to construct the model of Sudan's banking distress prediction.

2.3.4 Core Classification Techniques of Proposed Method

2.3.4.1 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning); the algorithm outputs an optimal hyperplane which categorizes new examples.

A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value (i.e. the class labels) and several attributes (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

(Hsu, 2016) proposed a novel procedure which guarantee high prediction accuracy comparing to other processes, given a training set of instance-label pairs (x_i, y_i) , $i = 1 \ldots . l$ where $x_i \in R^n$ and $y_i \in \{1, -1\}^l$, the support vector

machines (SVM) (Boser, 1992); (Vapnik, 1995) require the solution of the following optimization problem:

$$\min_{w,b,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \varepsilon_i \qquad (2.16)$$

$$\text{Subject to } y_i(w^T \emptyset(x_i) + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0. \qquad (2.17)$$

Here training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function $\emptyset$. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. C > 0 is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) \equiv \emptyset(x_i)^T \emptyset(x_j)$ is called the kernel function.

Generally there are many kernel functions available but the most popular ones which are continually addressed by SVM solutions are four:

Linear Kernel

The Linear kernel is the simplest kernel function. It is given by the inner product x_i, x_j . Kernel algorithms using a linear kernel are often equivalent to their non-kernel counterparts, i.e. KPCA with linear kernel is the same as standard Principle Component Analysis (PCA).

$$K(x_i, x_j) = {x_i}^T x_j \qquad (2.18)$$

Polynomial Kernel

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized.

$$K(x_i, x_j) = (\gamma {x_i}^T x_j + r)^d, \gamma > 0 \qquad (2.19)$$

Adjustable parameters are the slope alpha (γ) the constant term r and the polynomial degree *d*.

Radial basis Function

The (Gaussian) radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification (Chang , 2010)

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2, \gamma > 0 \qquad (2.20)$$

Sigmoid Function

The Hyperbolic Tangent Kernel is also known as the Sigmoid Kernel and as the Multilayer Perceptron (MLP) kernel. The Sigmoid Kernel comes from the Neural Networks field, where the bipolar sigmoid function is often used as an activation function for artificial neurons.

$$K(x_i, x_j) = \tanh(\gamma {x_i}^T x_j + r) \qquad (2.21)$$

Here, γ , r, and d are kernel parameters.

SVM Procedure

The best sequence of SVM classification task as proposed by (Hsu, 2016) is commencing by transform data to the format of an SVM package, then Conduct simple scaling on the data, The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Because kernel values usually depend on the inner products of feature vectors, e.g. the linear kernel and the polynomial kernel, large attribute values might cause numerical problems. In general, the RBF kernel is a reasonable first choice. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF (Keerthi, 2003) since the linear kernel with a penalty parameter $\tilde{C}$ has the same performance as the RBF kernel with some parameters (C,γ). In addition, the sigmoid kernel behaves like RBF for certain parameters. The second reason is the number of hyperparameters which influences the complexity of model selection. The polynomial kernel has more hyperparameters than the RBF kernel. Finally, the RBF kernel has fewer numerical difficulties comparing to other kernels.

After implementing RBF kernel Cross-validation should be used to find the best parameters C and γ. then those later parameters should be used to train the whole training set and eventually Test the model.

Figure 2.1 represents a classification problem similar to this study, to illustrate this issue. Filled circles and triangles are the training data while hollow circles

and triangles are the testing data. The testing accuracy of the classifier in Figures (*a)* and (*b)* is not good since it overfits the training data. If we think of the training and testing data in Figure *a* and *b* as the training and validation sets in cross-validation, the accuracy is not good. On the other hand, the classifier in *c* and *d* does not overfit the training data and gives better cross-validation as well as testing accuracy. This research will implement 5-fold cross-validation to optimize the model and avoid overfitting dilemma.

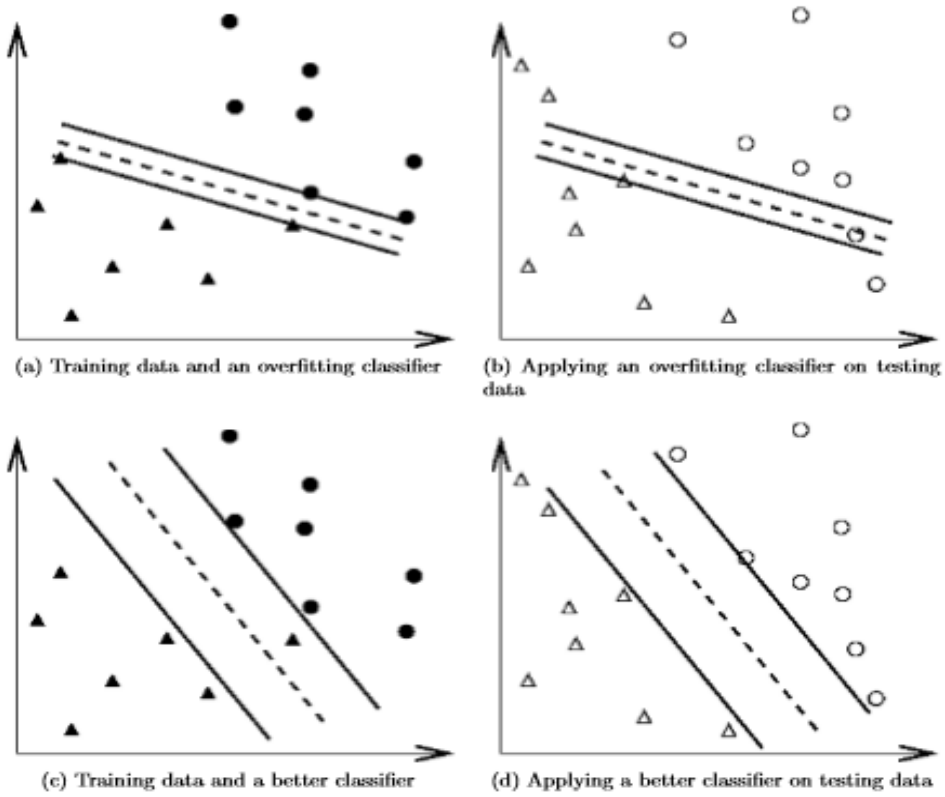

Figure 2.1: An overfitting classifier and a better classifier. (Hsu, 2016)

2.3.4.2 GA Parameters Optimization

Previous works on parameter optimization as well as results from those studies confirm that learning performances vary widely if the parameter settings

changes even on the same dataset. For instance, Eiben (2011) discuss the effect that parameters have on the performance of the Evolutionary Algorithms like the population size, the selection method, the crossover, and mutation operators.

"**Selection**" operator: Selection is performed to select excellent chromosomes to reproduce. Based on fitness function, chromosomes with higher fitness values are more likely to yield offspring in the next generation by means of the roulette wheel or tournament method to decide whether or not a chromosome can survive into the next generation. The chromosomes that survive into the next generation are then placed in a mating pool for the crossover and mutation operations. Once a pair of chromosomes has been selected for crossover, one or more randomly selected positions are assigned into the to-be-crossed chromosomes. The newly-crossed chromosomes then combine with the rest of the chromosomes to generate a new population.

"**Crossover**" operator: Crossover is performed randomly to exchange genes between two chromosomes. Suppose $S_1 = \{S_{11}, S_{12}, \dots, S_{1n}\}, S_2 = \{S_{21}, S_{22}, \dots, S_{2n}\}$ are two chromosomes, select a random integer number $0 \leq r \leq n$, S_3, S_4 are offspring of crossover (S1, S2),

$$S_3 = \{S_i | if\ i \leq r\ , S_i \in S_1, else\ S_i \in S_2\ \},$$

$$S_4 = \{S_i | if\ i \leq r\ , S_i \in S_2, else\ S_i \in S_1\ \}.$$

"**Mutation**" operator: The mutation operation follows the crossover to determine whether or not a chromosome should mutate to the next generation. Suppose a chromosome$S_1 = \{S_{11}, S_{12}, \dots, S_{1n}\}$, select a random integer number$0 \leq r \leq n$, $S_3$is a mutation of S_1,

$S_3 = \{S_i | if\ i \neq r\ , then\ S_i = S_1, else\ S_i = random(S_{1i})\}.$

Offspring replaces the old population and forms a new population in the next generation by the three operations, the evolutionary process proceeds until stop conditions are satisfied.

2.4. Bank Failure Prediction

As far as we know there are NO published researches similar to this study in Sudan in term of covering the whole Sudanese banking system. So this work is first of its nature in this country. However although there are plenty of studies available related to the selected approach (SVM) in bank distress prediction but none of them was found to hybrid factor analysis to strengthen the model in terms of variable selection methodology. Summary of related work found in the literature is provided as following:

2.4.1. CAMELS In Bank Fail Prediction

CAMELS rating are calculated in order to show financial performance of the banks in different respects. This system is a natural object of analysis, as it is not only a widespread supervisory tool, but also one of the few generally accepted quantifiers of the otherwise soft notion of bank safety (Derviz, 2008). CAMELS' ratio model is very suitable and accurate to be used as a performance evaluator banking industries and to predict the failure rate (Salhuteru, 2015). This system relies on various financial ratios obtained from periodic reports of the entities under their jurisdiction.

The ratios are also aggregated into performance indices based on various weighting or scoring schemes. The aggregation of the ratios can be a

complicated process involving subjective judgment. The changing economic conditions have made such aggregations even more difficult, increasing the need for a more reliable way to express a bank's financial condition. By concentrating on the top line and bottom line, banks across the board have improved their profit while reducing their operational costs and more number of banks has improved their financial performance by using the concept of mergers and acquisitions. CAMEL rating is used by most banks across the world as a performance evaluation technique (Raiyani, 2010).

(Duncan, 2004) indicated that all financial performance measure as interest margin, return on assets, and capital adequacy are positively correlated with customer service quality. To assess the performance of the bank is necessary to report the financial reports usually consists of a balance sheet, income statement, cash flow statement, statement of changes in equity and notes to the financial statement (Salhuteru, 2015). Some data from financial ratios are often compared to CAMELS in correctly identifying or predicting crisis events but sometimes, the relevant factors behind future failures or rating downgrades are satisfactorily captured by judiciously constructed risk sensitive summary statistics of conventional balance sheet data (Derviz, 2008).

The following table summarize the recent efforts that have been published about using CAMELS ratios on banks financial distress prediction:

Ref	Prediction Task	Capital (C)	Asset (A)	Management (M)	Earnings (E)	Liquidity (L)	Risk (S)
Dincer, et al (2011)	A Performance Evaluation of the Turkish Banking Sector after the Global Crisis via CAMELS Ratios	Equity to (Loan + Market + Principle Amount Subject to Operational Risk) / Equity to Total Assets/ Equity to (Deposit + Nondeposit Sources)	Financial Assets to Assets/ Loans and Receivables to Assets /Permanent Assets to Assets	Interest expenses to total expenses/interest incomes to total incomes/total incomes to total expenses	Net Profit to Total Assets/Net Profit to Equity	liquid assets to Assets/liquid assets to short term Liabilities/liquid assets to deposit and non-deposit sources	Total Assets to Sector Assets/ (Loans and Receivables) to (Sector Loans and Receivables)/ Deposits to Sector Deposits
Iqbal (2012).	Banking sector's performance in Bangladesh-An application	CAR	NPL				

	of selected CAMELS ratio						
Rozzani et al (2013)	Camels and performance evaluation of banks in Malaysia: conventional versus Islamic	Earning to assets	NPL	Staff costs to assets	ROA/ROE	Net loans to (deposits and short-term financing)/ Short-term liquid assets to deposits and financing	Risk sharia
Chandani et al (2014)	Impact of Gender of Leader on the Financial Performance of the Bank: A Case of ICICI Bank (India)	CAR/ proportion of debt to capital/Debt to assets/bond investments to assets	Noncurrent receivables gross to debt/ Noncurrent debt to debt/Loans	Debt to deposits/ Returns per employee	Operating profit to average capital turnover rate/ margin or net profit to assets/int	Securities to assets/Assets to deposits	-

			To assets/ Noncurrent net debt to loans		erest income to income		
Salhuteru et al (2015)	Bank Performance with CAMELS Ratios towards earnings management practices In State Banks and private banks	CAR/ Profit before tax to assets/ ROA/ Net profit margin/ Loan to Deposit					

Table 2.6: Recent efforts of using CAMELS ratios in bankruptcy prediction.

2.4.2. Support Vector Machines in Bank Fail Prediction

SVM is a quite new methodology among the group of intelligent methods. In recent years, the area of usage for SVMs increasingly expanded. For instance, image recognition, medical imaging, forecasting financial time series, electric load forecasting, electric demand forecasting, credit rating (Guo, 2012), rain forecasting, and forecasting crude oil prices are only a few examples among the countless fields of application for SVMs.

In the literature, only a few studies are present in which SVMs are used in the prediction of bank failures. Among them, (Chauhan, 2009) used a variety of intelligent methods in their study. They used 54 variables pertaining to 1,000 banks in total.

(Guo, 2012) investigated bank failures in Turkey for 1997–2003 and divided the banks into two classes as healthy–unhealthy and utilized CAMELS variables comprised of 20 financial ratios. In their study, various artificial neural analysis techniques such as support vector machines and multivariate statistical methods were used and then a comparison of their predictive powers was conducted. Besides, a third degree polynomial kernel was used in the support vector machines. According to the results of the study, it was seen that SVMs and MLPs models led to better results compared to the multivariate statistical models. On the other hand, (Erdal, 2012) compared the accuracy of SVMs and ANNs to predict the bank bankruptcies. Their study incorporates 35 privately owned commercial banks in the period between 1996 and 2000. A prediction of bank failures via the SVM was made and the results were compared using MLPs. The

study consists of three different models. In the first model, a first year data set is used while the second one uses a second year and the third uses a third year data set. A significant difference in favor of the SVMs compared with MLPs was observed in the prediction of non-failed banks as well the total accuracy. Accordingly, Model 1 (first year) and Model 2 (second year) were equally accurate as per both the SVMs and MLPs in terms of the classification of failed banks, whereas the MLPs yielded more accurate results in Model 3 (third year).

SVMs were only recently introduced in the field by (Pasiouras, 2006), who developed a nonlinear model using the RBF kernel and compared it with models developed with various other techniques. Hence, the use of the methodology in the prediction of banks financial distress was scarce, this research will be its first to study using factor analysis along with SVM and DA in predicting the financial distress of Sudan's banking system.

2.4.2.1 Ensemble SVM Classification Models

Ensemble methods have generally been used as tools to improve the accuracy of learning algorithms by constructing and combining an ensemble of weak classifiers (see Figure 2.2) each of which needs only to be moderately accurate on the training set (Perrone, 1994). Two popular methods for creating accurate ensembles are Bagging (Breiman,1996) and Boosting (Freund, 1996). Both theoretical and empirical studies have demonstrated impressive improvements in the generalization behavior.

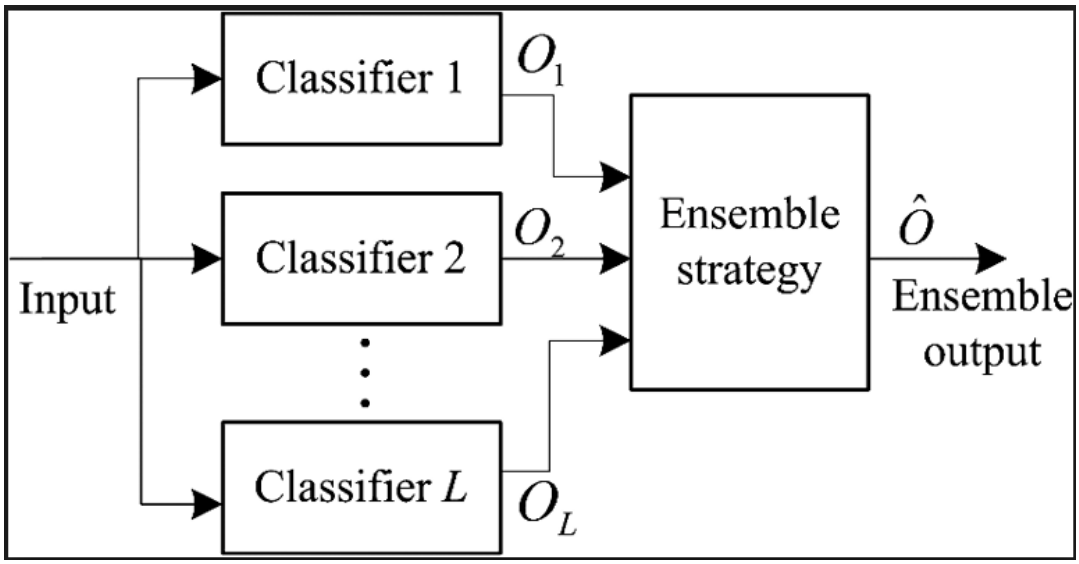

Figure 2.2: Overview of Ensemble Classification Model.

Recently, several studies on bankruptcy prediction have applied AdaBoost, which is one of popularly used Boosting algorithms, to bankruptcy classification trees. They have shown that AdaBoost decreases the generalization error and improve the accuracy Alfaro et al. (2007). An empirical comparison has shown that AdaBoost with classification tree decreases the generalization error by about 30% percent with respect to error produced with NNs. Previous studies have suggested that ensemble with classification trees is very effective for bankruptcy prediction, however, there has been little empirical testing of ensemble with SVMs in bankruptcy prediction literature. The major reason is that ensemble with decision trees provides fast training speed and well-established default parameter settings, while NNs has the difficulties for testing both in terms of the significant processing time required and in selecting training parameters. Ensemble method is expected to provide the following advantages over the traditional single SVM; First, SVM has introduced as one of prominent techniques which can show effective performance in bankruptcy prediction. Ensemble can produce even

more accurate results than any of the individual SVMs classifiers by making up the ensemble and thus intensifying discriminant capability of SVM.

Second, the classification approaches using error minimization, such as SVM, are prone to overfitting when a classifier is too closely adjusted to the training set, and the classifier's generalization error tends to increase when it is applied to previously unseen samples. Ensemble methods can make base classifiers such as SVMs to be robust to overfitting and thus reduce generalization error.

2.5. Evaluation Measures

To evaluate the results of this research, standard measures will be used to assess the classifiers performance; these measures consist of Accuracy, Sensitivity and Specificity.

Accuracy: The accuracy of a test is its ability to differentiate the non-healthy and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.22)$$

Accuracy generally represents the percent of correct classifications.

Sensitivity: The sensitivity of a test is its ability to determine the sample cases correctly. To estimate it, we should calculate the proportion of true positive in input cases. Mathematically, this can be stated as:

Sensitivity = $\frac{TP}{FN+TP}$ (2.23)

Specificity: The specificity of a test is its ability to determine the healthy bank cases correctly. To estimate it, we should calculate the proportion of true negative in healthy bank cases. Mathematically, this can be stated as:

$$\text{Specificity} = \frac{TN}{FP+TN} \quad (2.24)$$

For calculating these measures we need to take into account True Positive, True Negative, False Positive, False Negative these parameters can be obtained from confusion matrix as explained in following table

Actual **Predicted**	**Sound Bank**	**Distressed Bank**
Sound Bank	True Positive (TP)	False Positive(FP)
Distressed Bank	False Negative (FN)	True Negative (TN)

Table 2.7: Bank Distress Confusion Matrix

2.6. Discussion

The CAMEL model is very famous among controllers because of its effectiveness. This model is very good for the evaluation of the performance of the banks. CAMEL model was implemented by North

America Bank regulators to judge the financial and administrative reliability of commercial loaning organizations. This model assesses the general condition of the bank, its qualities and shortcomings (Aspal, 2014).

By studying the CAMELS ratios the research will identify the factors that can potentially be used as good predictors for Sudan's banking system distress prediction , to fulfill this objective the research will conduct an interview for the subject matter experts which representing the banks regulator asking to determine which factors from CAMELS can represent the most prediction power , also the research will consider one of the dimensionality reduction techniques (Factor Analysis) to be used in order maximize the efficiency of selected model , this research will try to utilize the benefits of factor analysis represented in its easy of pattern interpretation as well as its power to compute the maximum likelihood estimates of the parameters or factors meaning while get benefit of PCA by using it in calculation of the initial factors .

There occurs a small number of theoretical works regarding the factors that affect the probability of commercial banks failure. Diamond and (Dybvig, 1983) studied the lack of liquidity, as the reason of banking crises. The main purpose of the banks is to convert liquidity, and the situation when each depositor expects other depositors to withdraw money is the main reason for banking failure. (Chinn, 2000) developed a theoretical model of a financial crises and concluded that countries accumulation of foreign debt is one of the major reasons for the banking failures during currency devaluation. However, other theoretical studies, e.g. (Dekle, 2001), emphasized the domestic debt risk and high leverage ratio as the main factors that

influence the probability of a breakdown. The size of the credit portfolio was also shown to be positively related to the probability of bank failure by (Caminal, 2002). According to the study large banks are able to diversify individual risk, and thus will accept higher exposure to the aggregate shocks. That will make large banks more vulnerable to the country-level crisis.

The main idea of the empirical studies that examine bank-level data is to find those explanatory variables which help to predict the probability of failure. The focus of this research is to find out which factors can affect the Sudan banking system starting by all used previously factors and eliminate those are not relevant with Sudan Islamic banking system and others which have No available data to compute. Also, suggesting new factors which thought to influence the banks status.

(Khalafalla, 2013) was the only researcher who studied Sudan's banking sector from macro perspective , he designed an early warning system using logistic regression to classify commercial banks into four groups (Fair, Marginal , Satisfactory , Strong and Unsatisfactory) using the recommendation of (CAEL) which is an standard imposed by central bank of Sudan to control the capital adequacy of commercial banks. However while his study focused on only four banks in the period (2002-2009) , This research is studying all banks operated in Sudan in the period of (2006-2014) using a proposed optimized parameters ensemble SVM . Also, while he concentrated on CAEL variables, this research is proposing new factor which never used before (RainFall) as well as all CAMELS available factors. Furthermore, this research studied the Islamic finance modes and

sectors and advice about to what extent finance modes and sectors are threaten the banks performance

Based on literature review results of applied intelligent techniques (Armando, 2012), the research discovered a promising result if SVM method is implemented along with other feature selection technique, hence, newly proposed hybrid feature selection DAGA-FS and SVM will be combined to construct the model of Sudan's banking distress prediction.

(Vasileios, 2014) compares conventional and Islamic banks regarding failure risk and its sensitivity to accounting statement and macroeconomic variables. The empirical strategy was survival analysis and the sample comprises 421 banks from 20 countries over the period 1995 to 2010.

(Mohammad, 2011) focused on predicting financial distress for Kuwaiti commercial banks based on time (in years). The study aimed to predict the financial distress cycle for Kuwaiti commercial banks in the GCC region. A logistic regression was used to analyze the financial data collected for this purpose.

This research will study the factors (finance mode, sector, and payment method) and find out which of them is affecting the health of Sudan Islamic banking system. The selection of these variables was confined by the availability of data from all banks regarding financing operations in the period of study.

2.7. Summary This chapter describes in detail on the fundamental concepts and methods that have been used in prediction of bank financial distress. Literatures that conducted studies to propose

country models of bank distress prediction as well as related studies were explored in this chapter, CAMELS rating system has been discussed as an entry point for the ratios that will be used throughout this research , beside subject matter expert selected ratios Factor analysis will be implemented as dimensionality reduction techniques arguing for better classification performance, bank distress prediction methodologies were classified in two major categories : statistical and intelligent , Discriminant analysis and Support vector machine will be used as classifiers for this research according to best achieved performance in related work as well as providing novel ensemble with factor analysis to propose a classification model for Sudan’s banking system failure prediction.

Chapter Three

Research Methodology

3.1 Introduction

This chapter presents the methodology used in this research. It describes the implementation of the chosen methods in achieving the goal and objectives of the research. One of the objectives of this study is to investigate factors that represent the highly potential risk areas for the banks. This research also designs a model that can be used to predict the bank financial distress depending on the factors on the former objective. This study also considers identifying the Islamic finance factors that may affect the healthiness of Islamic credit processes. This chapter discusses about the steps taken to carry out this research. There are four sections for this chapter where section one or section 3.1 is for the introduction. It continues with the research operational framework, which is presented in section 3.2. Overview of the operational framework is depicted in Figure 3.1. This framework is separated into seven phases. Section 3.3, on the other hand, is on the evaluation and reporting of the results. Finally, section 3.4 summarized all that were presented in this chapter.

3.2 Research Operational Framework

Operational framework provides operational guidance in a structured manner. In order to help researchers achieve the objectives of the research, the operational framework should be well organized in a systematic process. Figure 3.1 illustrates the flow chart of the operational framework of this research.

- There are five phases in this operational framework, namely; Phase 1: initial study and data preparation, Phase 2: Features selection using hybrid filter-based and wrapper based technique, Phase 3: Ensemble optimized support vector

machine (SVM) model ,Phase 4: Islamic credit risk analysis, Phase 5: Research Writing. Each phase is described as follows:

3.2.1 Phase 1: Initial Study and Data Preparation

This phase consists of six main elements: problem formulation, literature review, identifying existing techniques, obtaining data set, data preprocessing and building the data warehouse. Problem formulation involves the process of identifying the issues that exist in the bank's financial distress that does not have a solution or the available solution still has chances for improvement. It is done by doing literature review to analyze the existing bankruptcy prediction methods. By reviewing previous work, the best techniques both from statistical method and machine learning method can be used in our method. This phase has the following six activities:

Literature Review: This phase reviews and studies the research work related to banks financial distress prediction approaches, factor analysis, discriminant analysis, support vector machine. The reviews of previous works have been done with a related research topic; also the optimal variables for prediction models have been studied representing in well-known CAMELS ratios. Literature review was continuously performed until the research is finished. It is important to make sure the novelty of the research and to identify useful information related to the research. Throughout the literature review, related information will be recorded.

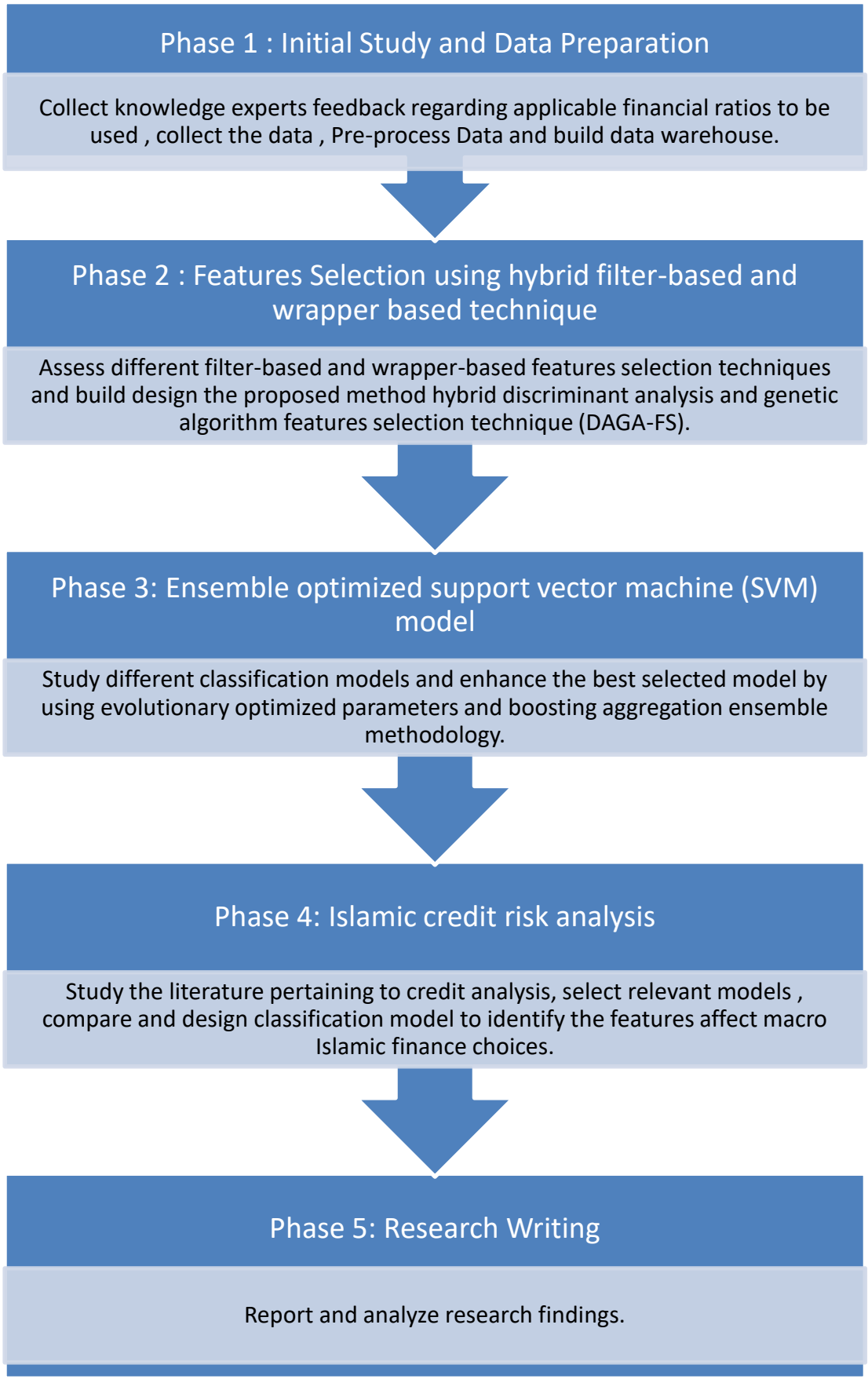

Figure 3.1: Proposed Operational Framework

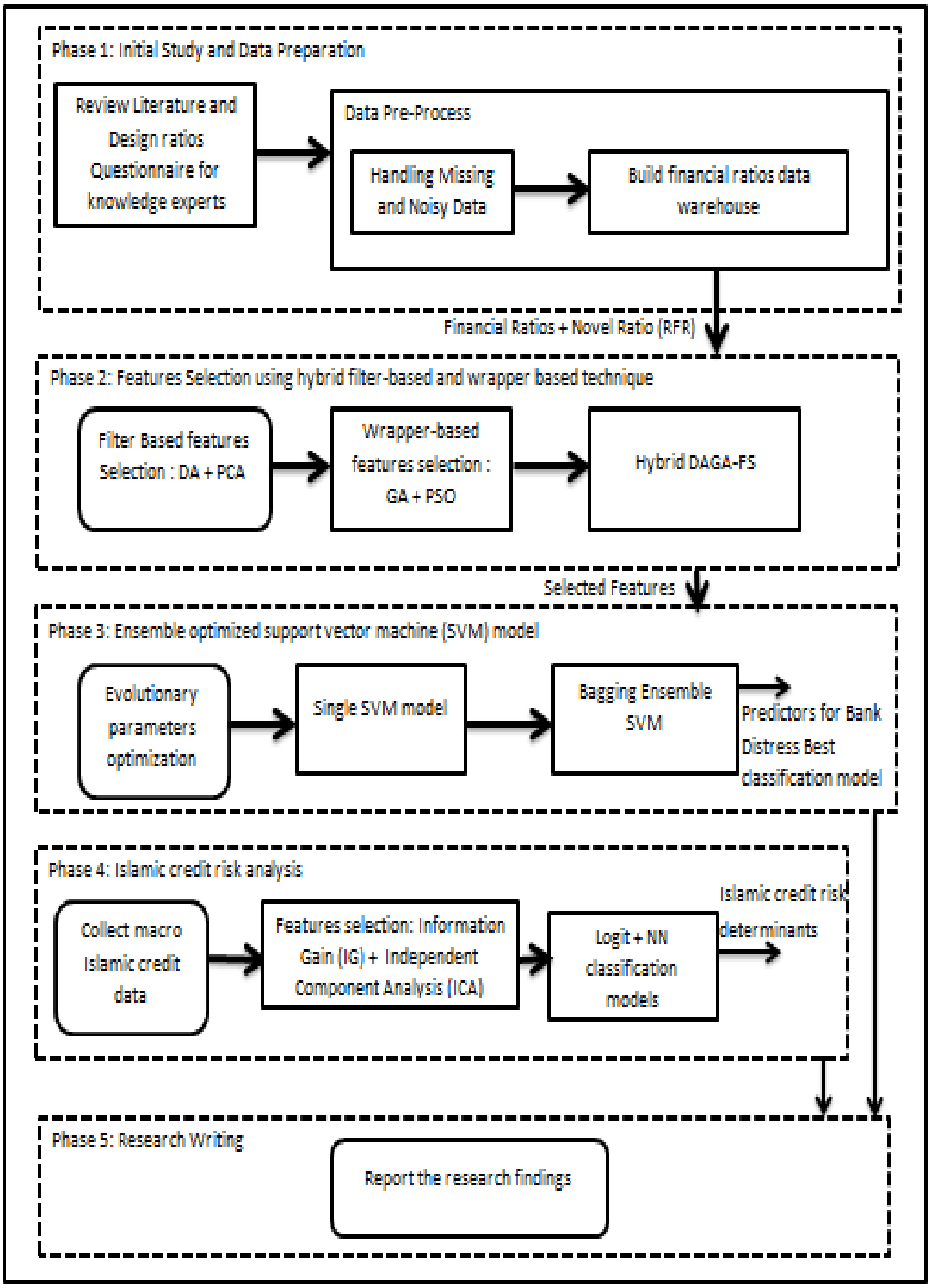

Figure 3.2: General Research Overview

Analyze Data Sets: obtaining financial data for all banks working in Sudan (37 banks) is a lengthy process starting by getting the access permission from the central bank of Sudan.

Data acquisition system (DAS) gets benefit from Microsoft excel software, the regulator authorities designed the forms and control their validity. Once bank upload excel form, DAS verify its correctness and perform the required validation, data eventually reside on structural database, and this research will access the database and analyze its content to obtain the relevant dataset.

The data will be found on financial statements sent by commercial banks via DAS system as well financial data for the sake of credit analysis.

Preprocessing: raw data extracted from DAS database contains anomalies such like: redundancy, incompleteness, null values and even logical data errors, the later anomaly is hard to discover unless check conducted by subject matter expert or data stewards, other anomalies can be rectified by means of data quality solutions.

There are different methods and techniques that central banks or countries monetary authorities used to assess their banks performance and how they could predict its distress. In this activity, we will question subject matter experts on how they could predict the probability of their banks failure, and on which factors or financial indicators they are rely on to assess the bank risks.

After specifying the expert's opinion, we will start determining which data that we need in order to carry out the prediction process. Also we need an initial explanation of the changes on the indicators along with everyone to point out whether the increase/decrease on such ratio will affect negatively

or possibly on the institution under study. There are three main activities performed in this stage: Handling missing data, Handling Noisy data, and Handling Redundant data.

Handling missing data : data is not always available , and this can be result for many reasons such like: equipment malfunction, inconsistent with other recorded data and thus deleted, data not entered due to misunderstanding, certain data may not be considered important at the time of entry and not register history or changes of the data , Imputation is going to be used to tackle this issue by Use the attribute mean to fill in the missing value, or use the attribute mean for all samples belonging to the same class to fill in the missing value.

Handling Noisy data: Noise is a random error or variance in a measured variable appears as a result for many possible reasons: faulty data collection instruments, data entry problems, data transmission problems, technology limitation, inconsistency in naming convention. This research will detect suspicious values and check by data stewards either to fix or remove the affected records also SVM requires that each data instance is represented as a vector of real numbers. Hence, if there are categorical attributes, we first have to convert them into numeric data.

Handling Redundant data: redundancy often occurred as a result of uploading the same excel file to DAS database many times because there is no such validation in the existing system, Microsoft SQL data quality can be used to detect and process the redundant data.

Figure 3.2 shows raw extracted data which contain some missing values in General Reserve amounts as well as duplicated records, such kind of anomalies might affect the quality of results generated by classification model.

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Paidupcapital	year	LegalRese	GeneralRe	Retainede	MonthLos	RWA	Revaluati	GeneralPr	EquityPar	TotalAsse	Loan	ProvsionF	NonPerfo	NetIncom	Cash	CurrentDe	ShortTerm
5067346131	2006	3.85E+08	0	0	7.82E+08	436743.5	5.69E+08	16781090	7418310	1.62E+10	1	8.52E+08	22828011	1	2.15E+09	3.27E+10	0
13550990	2006	0	0	0	7651690	334236.5	493969.5	2196193	884016.8	5186138	1	671986	10221843	1	2126396	19258139	0
6775495	2006	0	0	0	7651690	334236.5	493969.5	2196193	884016.8	5186138	1	671986	10221843	1	2126396	19258139	0
1531815	2006	543710	0	1100281	1272814	1.46E+08	0	1434931	401926	4647652	1	283279	5546317	1	654374	10253089	0
5067346131	2006	3.85E+08	0	0	7.82E+08	436743.5	5.69E+08	16781090	7418310	1.62E+10	1	8.52E+08	22828011	1	2.15E+09	3.27E+10	0
6775495	2006	0	0	0	7651690	334236.5	493969.5	2196193	884016.8	5186138	1	671986	10221843	1	2126396	19258139	0
3353624	2006	794687	0	2	1654567	106226.5	3043069	1599777	4954091	1	1	677014	55115455	1	0	0	0
3353624	2006	794687	0	2	1654567	106226.5	3043069	1599777	4954091	1	1	677014	55115455	1	0	0	0
5067346131	2006	3.85E+08	0	0	7.82E+08	436743.5	5.69E+08	16781090	7418310	1.62E+10	1	8.52E+08	22828011	1	2.15E+09	3.27E+10	0
3091678	2006	0	0	0	2072480	108300.5	2924845	796659	1248825	6816138	1	574870	71124398	1	1074699	20777414	0
0	2006	0	0	0	16501	12010	0	0	1	25642	1	0	2315513	1	40240	186688	0
900	2006	900	6770	1921660	1776017	151629.7	522920	0	1	670001	1	621273	91725439	1	305612	13816712	0
25400000000	2006	59576000	1.81E+08	6.48E+08	4.28E+09	NULL	0	1173203	6147712	5.48E+09	1	4.38E+08	56783383	126335	0	4.27E+09	0
3611945	2007	0	294821	1952	31649	3.96E+08	5241203	495588	4158840	27989860	1	844200	84984140	1	1347101	16589685	0
1531815	2007	543710	0	1100281	1272814	1.46E+08	0	1434931	401926	4647652	1	283279	5546317	1	654374	10253089	0
22121556	2007	1454003	0	1602857	-138306	1.31E+09	4514592	14824888	22465654	48761837	1	3216410	65789362	1	4063502	47289714	23582954
3353624	2007	794687	0	2	1654567	106226.5	3043069	1599777	4954091	1	1	677014	55115455	1	0	0	0
3679992	2007	0	38757	-624083	24180	NULL	0	764671	1569034	11892607	1	629930	1.32E+08	1	2302808	0	0
6000000	2007	611167	154803	6264	495006	735271.5	0	4049933	14463918	52472728	1	782491	12939635	1	3928724	24462493	0
6350920	2007	0	60587	2676361	166826	62731	0	0	41550133	6309649	1	55338	763816.4	1	121441	0	0
5067346131	2007	3.85E+08	0	0	7.82E+08	436743.5	5.69E+08	16781090	7418310	1.62E+10	1	8.52E+08	22828011	1	2.15E+09	3.27E+10	0
3091677	2007	0	0	0	2072480	108300.5	2924845	796659	1248825	6816138	1	574870	71124398	1	1074699	20777414	0
4273345	2007	0	0	871853	-6760	92279.41	0	103364	8155456	2065063	1	107702	10945270	1	739181	6444052	0

Figure 3.3: Sample of Raw Data

Prepare Data Warehouse: this research tries to propose solution to be exploited later by regulatory authorities to predict the likelihood of banking system collapse, the process of maintaining the data quality and automatic calculation of financial ratios from DAS database is lengthy and costly. The data warehouse can be built and used as permanent source of upcoming data analysis projects. More details are described in next section.

3.2.2 Phase 2: Features Selection using hybrid filter-based and wrapper based technique

In order to perform features selection task to features used in bank distress prediction classification, four techniques were selected to represent filter and wrapper based techniques that to facilitate proposing new hybrid method which combine the strengths and eliminate the weaknesses of each type. Discriminant analysis (DA), principal component analysis (PCA), particle swarm optimization (PSO) and genetic algorithm (GA) based on eleven financial ratios stored in data warehouse will be assessed according

to following high-level process: After subject matter experts identified the optimal set of financial ratios as well as rain fall ratio that proposed by researcher, these factors will be passed as parameter to next component. The next component performs feature selection techniques which will reduce the number of selected predictors to a set of strong predictability power that can be used in subsequent classification task. The process will be divided in two tasks: first the best filter based feature selection will be identified and similarly for wrapper based. The better of two methods will be combined to propose the hybrid feature selection that can better select the strong ratios to be used in subsequent tasks. The output of this component will be passed to next one which is consist of two classifier (support vector machines (SVM) and feed forward neural network (NN)) to evaluate the result set of reduced predictors. More details regarding all steps are presented in chapter4.

3.2.3 Phase 3: Ensemble optimized support vector machine (SVM) model

The output of phase two will turned as input for the classification model which is going to be designed in phase 3, after assessing different single classification models including neural network and support vector machine, the later has been selected to be enhanced due to its performance comparing to other in bank distress literature, the first enhancement is optimizing the parameter of SVM kernel.

In order to optimize SVM parameters namely (C and γ) GA is used to optimize the training parameters. GA has strong global search capability, which can get optimal solution in short time. So GA is used to search for better combinations of the parameters in SVM. After a series of iterative computations, GA can obtain the optimal solution. The methods and

process of optimizing the SVM parameters with genetic algorithm is described as follows:

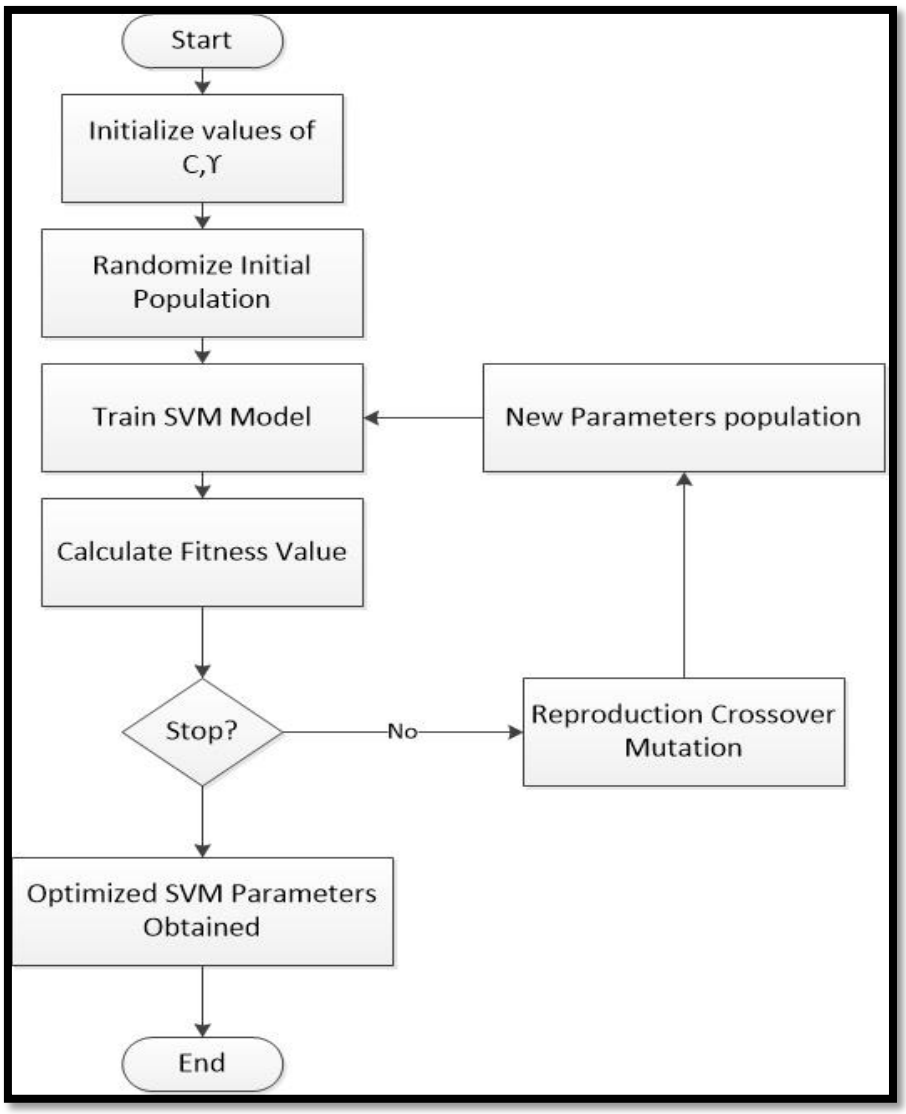

Figure 3.4: SVM parameters optimization with genetic algorithm.

Then bagging ensemble method is followed to design local ensemble SVM aiming to achieve higher accurate results.

Instead of using the same training set to fit the individual classifiers in the ensemble, we draw bootstrap samples (random samples with replacement) from the initial training set, which is why bagging is also known as *bootstrap aggregating*. Many models (versionsC_i) will be generated as explained in figure 3.10 from the same SVM machine but with different settings which lead to generate different predictions (P_i) be combined and

aggregated using majority voting aggregation strategy. More details are given in chapter 5.

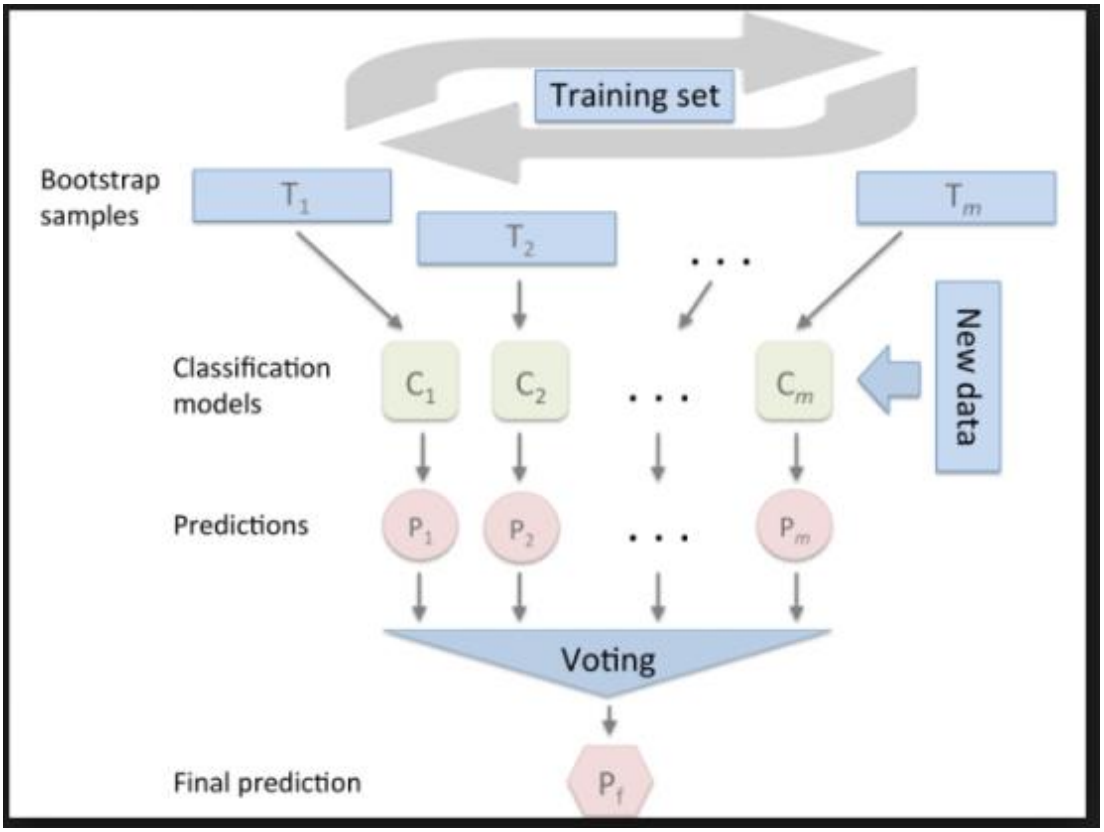

Figure 3.5: Ensemble SVM using Bagging methodology.

3.2.4 Phase 4: Islamic Credit Risk Analysis

After Islamic finance in Sudan including the basic definitions and differences from the earlier conventional system has been investigated and reported, then applied Islamic finance modes in Sudan's banking system and the relevant regulations and risk management policies will be evaluated, by this study we can be able to conduct further analysis of credit risk analysis following steps will be performed.

Financial data of all commercial banks will be gathered, due to limited access to data only few dimensions are available to researchers which are (Finance Mode, Sector and Payment Method).

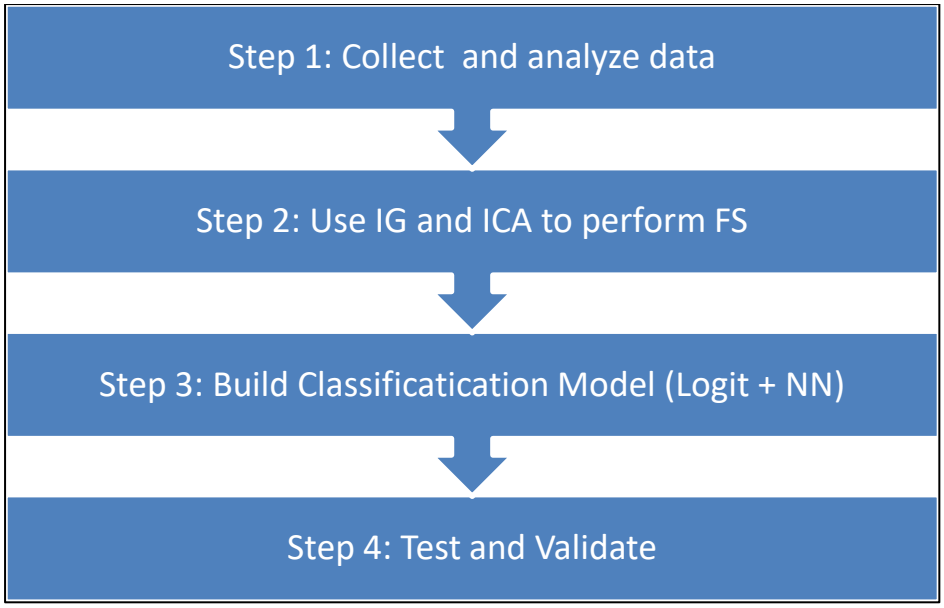

Figure 3.6: Credit risk analysis steps.

These dimensions are the core base of formulating the monetary policies regarding financing controls. After data is analyzed and cleansed then two features selection method will be applied to determine the best namely information gain (IG) and independent component analysis (ICA). The selection of both methods came after reviewing the existing researches Alhaj (2016), the selected features will form input to designed classification model. Again neural network and logistic regression has been tested and the best performance was selected. Further details will be presented in chapter 6.

3.2.5 Phase 5: Writing

The last phase of this research is writing the research report as thesis chapters. The report includes the problem statements, objectives, scopes of this study, as well as the literature reviews of the previous researches in bank financial distress prediction, and the implicated methods are included in the report. The report also explains clearly on the methodology used in this research. The analysis of the results will clear up the lesson learned

from this research and also the future directions of the research are included in the report.

3.3 Factor's Data Warehouse

The reason for such an implementation would give financial analysts at all levels of the monetary authority an integrated, secure and consistent data source from which they could report on and set their business needs more efficiently than possible without one and in the same time serve this research objectives by consolidating all the data required by predictive models in single unified data source.

To gain the desired results, 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 (Fig. 4.2) 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.

The design of a data warehouse, in this case in a regulatory authority, looks to solve the problem of integrating multiple systems into one common data source. With the diverse roles that a central bank of Sudan has both on the local and external sides of the financial sector i.e. the internal sources for

managing internal staff procedure as well as external monitoring and regulating commercial banks, data more likely to reside in different transactional databases across the central bank dependent agencies such like stock market , electronic banking system company , credit bureau office moreover some information brought manually using different storage media from different government sectors and agencies.

Much of the information exists in a silo in and of itself. The information does not translate across the spectrum to help process of monitoring and tracking commercial banks. An example of this may be that the polices and researches department has information about the exchange rates, information that may be important to the prudential supervision department since the exchange rate fluctuations may affect the overall banks performance and put them in danger zone; however, this useful information is not shared between these two departments.

Certainly, having a data warehouse that shares this kind of information with the masses could cause internal strife or possible breaches of security. Therefore, devising a plan that restricts data, as appropriate, makes reasonable sense.

Different sources that used to calculate identified ratios will be presented in table 3.1

Source	Description	Owner
Banks balance sheets data	Excel sheets from all commercial banks	Every single sheet owned by different department
Banks liquidity data	Provided by external system called SIRAG	Central operations department
Banks reserves	Provided by external system called SYMBOLS	Financial administration department
Economic Statistics	Provided by external agency called : Central Statistics Corporation	Central Statistics Corporation
Rainfall data	Provided by external agency called : Meteorological Authority	Meteorological Authority

Table 3.1: Ratios Sources Properties.

All data in table 3.1 are found to be generated in spreadsheets format and except banks reserve which can be available in many formats.

The most significant motivation to implement a data warehouse is to have a better platform on which to report data. Combining the data from all the other databases and sources in the environment, the data warehouse becomes the single source for users to obtain data. All reporting and information would be based on a single database, rather than on individual

repositories of data. Using a data warehouse, a report writer does not need to learn multiple databases or attempt to try to join data between more than one database, a task which may prove difficult.

Besides having one single database to report from, there are other benefits to not calculating from transactional databases. Namely, calculating and reporting from the transactional database can slow down the general database performance by utilizing resources already reserved to the system. Running such a financial factor calculation process against certain parts of a database could potentially cause slower experience for the user, because it could render a database or the application using it unresponsive. Finally, data can be in the process of an update or delete transaction in the database, which could produce wrong factors calculations while it is running.

3.4. Architecture

Two different types are commonly used for data warehouse architectures. The first classification is a structure-oriented one that depends on the number of layers used by the architecture. The second classification depends on how the different layers are employed to create enterprise-oriented or department-oriented views of data warehouses.

3.4.1. Single-Layer Architecture

Single-layer architecture is not regularly adopted in real scenarios. Its goal is to decrease the amount of data stored; to achieve this objective, it get rid of duplication of data source layer. In this case, data warehouses are virtual. This means that a data warehouse is implemented as a multidimensional view of operational data created by specific middleware, or an intermediate processing layer (Devlin, 1997).

The drawback of this architecture represents in its inability to meet the needs for separation between analytical and transactional processing.

Analysis queries are sent to operational data after the middleware interprets them. It this method, the queries affect transaction daily performance. Moreover, in spite of this architecture can meet the need for integration and correctness of data, it cannot log more data than sources do. For these reasons, a virtual approach to data warehouses can be work only if analysis requirements are particularly restricted and the data amount to analyze is big. This was the main motivation to adopt two-layer architecture.

3.4.2. Two-Layer Architecture

The need for two different layers plays an essential role in defining the typical architecture for a data warehouse system, Although it is typically called two-layer architecture to highlight a separation between physically available sources and data warehouses, it actually consists of four subsequent data flow stages figure 4.1 (Lechtenbörger, 2001) explained in following:

1. *Source layer*: A data warehouse system uses heterogeneous sources of data. That data is basically stored to corporate relational databases or legacy databases, or it may come from information systems outside the corporate environment.
2. Data staging: The data stored to sources should be extracted, cleansed to remove anomalies and process missing data, and integrated to merge heterogeneous sources into one common schema. The so-called Extraction, Transformation, and Loading tools (ETL) can combine different sources, extract, transform, cleanse, validate, filter, and load source data into a data warehouse (Jarke et al., 2000). Technically speaking, this stage work to tackle issues pertaining to distributed systems, such as inconsistent data management and incompatible data structures.

3. Data warehouse layer: Information is located in single logically centralized one repository: a data warehouse. The data warehouse can be directly accessed, but it can also be used as a source for creating data marts, which partially replicate data warehouse contents and are designed for specific enterprise departments, result in more granular control on the data perspectives and security.
4. Users Analysis: In this layer, integrated data is efficiently and flexibly accessed to issue reports, dynamically analyze information, and simulate hypothetical business scenarios. Technologically speaking, it should feature aggregate data navigators, complex query optimizers, and user-friendly GUIs. Modeling techniques are heavily used at this point to utilize the huge amount of data available.

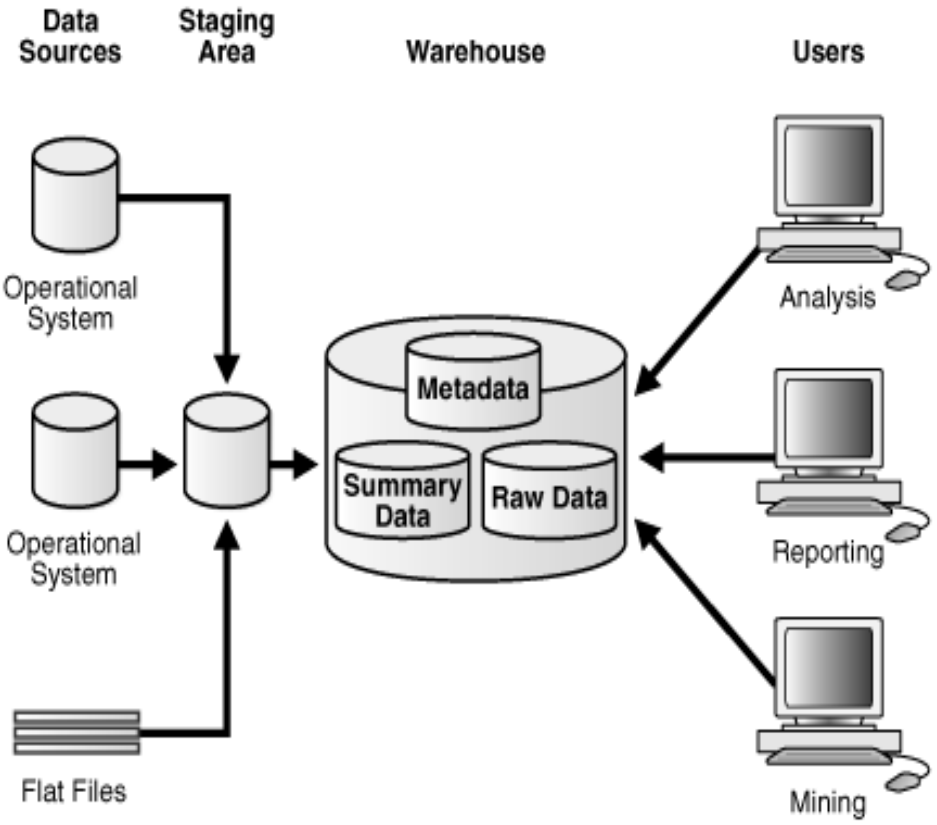

Figure 3.7: Tow-Layers Data warehouse Architecture.

3.5 ETL Process

One of the most crucial parts of the data warehouse is the extracting, transforming and loading (ETL) of data from the operational transactional databases to the data warehouse itself.

Although it is possible to build an ETL that would complete its task as the go-between a transactional database and the eventual data warehouse it is logical to extract the data and locate it temporarily before transforming. Well known as a “Data Staging” area or Operational Data Store, this area of the data warehouse exists only to pull data out of the operational system that there is no load pressure on the transactional system yet more control in the data quality tasks, then rearrange and cleanse (also known as transforming) data before being loaded into an organized data warehouse. See Figure 3.7, which shows where one would find the staging area in a typical data-warehousing environment. A data warehouse, and especially this current suggested one for financial factors, has data coming from multiple sources, and this could make an ETL run for much longer than would be feasible to have a data warehouse in the first place. Therefore, quickly pulling all source data makes sense at this scale, once the data pulled out of the transactional databases, having all the data away from business operations and in a secluded location gives the ETL a chance to stage and transform at its convenience. Once the data go through the transformations in a reasonable way, the data would next be going through the easy step of the ETL where it loads the dimension and fact tables of the data warehouse.

There are stimulating alternatives when it comes to transforming data, which in the past would not be possible or allow for any updates. Kimball says that, “there are numerous potential transformations, such as cleansing the data (correcting misspellings, resolving domain conflicts, dealing with

missing elements, or parsing into standard formats), combining data from multiple sources, DE duplicating data, and assigning warehouse keys (Kimball, Toolkit, 8)."

Having data in multiple sources that combined for one eventual data destination almost requires intervention by the ETL to make the data conform to standards. To the point of misspellings, there is a grand opportunity to do what users will still not do, as they should sometimes. With other available transformations, such as in the SQL Server Integration Services (SSIS) ETL tool, there exists the opportunity to add columns based on a derivation of data along with another popular features, known as a "Fuzzy Lookup." The Fuzzy Lookup "joins a data flow input to reference table based on column similarity; the Similarity Threshold setting determines the closeness of allowed matches" Knight et al. (2008). This lookup will take the information provided in the ETL, and based on a predetermined acceptable percentage of likelihood of a match, give the user the chance to make a very educated decision as to what to include or leave out.

3.6 Dimensional Modeling

The most important process of organizing and configuring the data is the stage known as "Dimensional Modeling," which is the step that configures and organizes the final destination for the data. Kimball states: "Dimensional modeling is widely accepted as the preferred approach for DW/BI presentation.

3.6.1 Star Schema

The star schema is the backbone of the data warehouse development, as it provides a de-normalized database structure for reporting the data that is vital to the business needs. The star schema itself is simply defined as, "a dimensional design for a relational database" Adamson (2010). The schema

gets its name because of its appearance, as it appears in the shape of a star (Figure 3.11) with the dimension tables that surround the focal point of the star schema, the fact table. The surrogate keys are the integer indexed primary key in the dimension tables that becomes the foreign key just for linking to the fact table, and this creates the basis for the star schema. Once the links are in place to the dimension tables, the star begins to take shape, and this is very common to data warehousing. The fact table possesses the many keys to the related dimension tables and makes the star schema possible. Since the fact table describes a business process, “it should provide a comprehensive set of measurements, even if some are redundant” Adamson (2010). This affords the possibility of having reporting made easier by having data already readily available within that table.

3.6.2 Snowflake Schema

There is yet another interesting schema in the data-warehousing world, known as a snowflake schema, and this is another way to bridge data with existing star schemas. The snowflake schema will appear when “the relationships between dimension attributes are made explicit in a dimensional design” Adamson (2010). With data being similar in some databases or even within the same database, it is possible for such a relationship to exist within a collection of star schemas as a need may arise within the scope of the business requirement to do so. Much like how the star schema gets its appearance from the shapes it takes as designed; the snowflake is named due to its branching off the dimension table from a star schema.

The snowflake schema is similar to the star schema. However, in the snowflake schema, dimensions are normalized into multiple related tables, whereas the star schema's dimensions are de-normalized with each dimension represented by a single table. A complex snowflake shape

emerges when the dimensions of a snowflake schema are elaborate, having multiple levels of relationships, and the child tables have multiple parent tables

3.7 Implementation

3.7.1. Pre Analysis

Because CAMELS ratios are various and fixed for different banking sectors, we interviewed the same employees again to determine which factors that can be used in selected prediction models. We have designed a questionnaire to elicit the information (Appendix A).

In the questionnaire, standard factors of BASEL have been adopted, which is specialized on evaluating the required financial stability factors to decide which best practices to adopt in order to have a healthy banking sector worldwide.

There were 27 participants with different designations on our questionnaire; all of them belong to the Department of Prudential Supervision at the Central Bank of Sudan. Most participants have nine years' experience at central bank of Sudan.

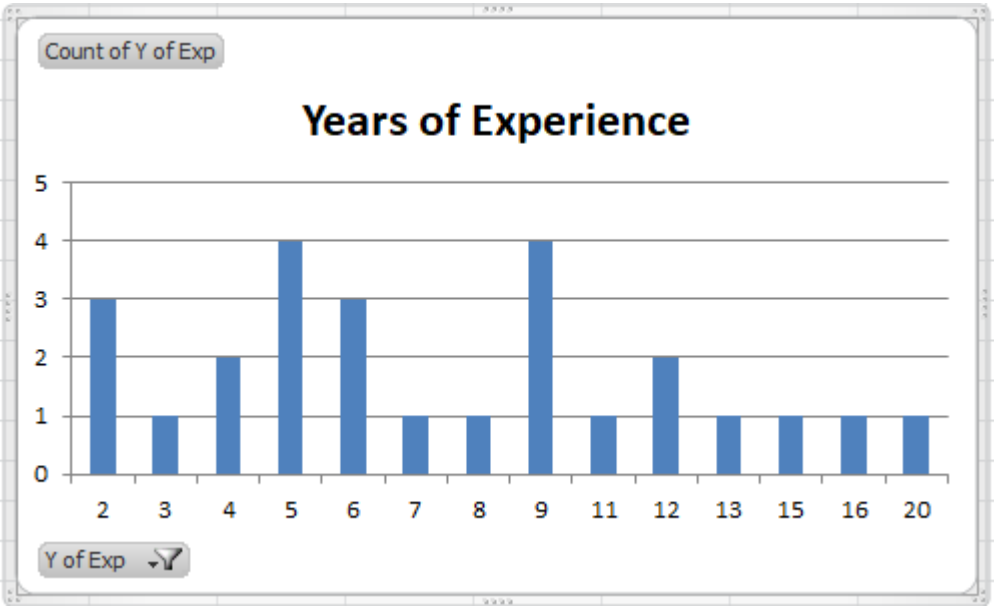

Figure 3.8: Questionnaire's Participants profile.

We have analyzed the Reponses (Appendix B) and came up with the following results, based on the highest frequencies for the 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

Table 3.2: CAMELS Ratios.

Next, methods to calculate those factors in the Sudanese banking sector will be identified. Appendix D (D.2) shows the detailed sources and full

calculation formulas for used fields. Central Bank employees have been interviewed to reach the high-level calculation formulas for selected ratios:

Ratio	Description	Calculation formula
T1C	Tier 1 capital Ratio measured as a ratio of Tier 1 capital to risk weighted assets.	Paid-up capital+ Legal reserves+ General Reserve+ Retained earnings-month loss / risk weighted assets
TCR	Total Capital ratio measured as a ratio of (Tier 1 + Tier 2 capital) to risk weighted assets	Paid-up capital+ Legal reserves+ General Reserve+ Retained earnings-month loss+ revaluation Reserve+ General provision / risk weighted assets
EAS	Ratio of equity capital to total asset.	Equity/ Total Asset
LAS	Ratio of net loans to total assets.	(Loans- Provision)/ Total Assets
LLP	Ratio of loan loss provisions to total loans.	
NPL	Ratio of non-performing loans to total loans.	Non-performing Loans/ Total Loans.
NIM	Net interest margin measured as a ratio of (interest Received - interest Paid) to total earning assets.	NA

ROE	Return on equity measured as a ratio of net Income to Capital equity.	Net Income/ Equity
ROA	Return on assets measured as a ratio of net Income to total assets.	Net Income / Total Assets.
IBR	Interbank ratio measured as a ratio of deposits due From banks to deposits due to banks.	NA
LADF	Ratio of liquid assets to deposits and short term funds.	(Cash + Current Deposits + Reserve on Central bank + Short term securities)/ Deposits and short term funds.
IDIVER	Finance related income to total income	Income from finance / Total income

Table 3.3: Ratios calculation formula.

When we take a look at the sources of information that lead us to the calculation formula, some ratios should be excluded at this stage, and that's because:

1. Data is not available by commercial banks (IBR).
2. Due to the fact that the interest system is not used in Sudan, ratios cannot be used in its banking sector (NIM).

3.7.2 Building the Data Warehouse

Having our data dispersed imposed a challenge, as well as the time and effort to collect it. So we designed a Data Warehouse (DW) that can help us and serve the Central Bank for future use. Following are the details of the performed tasks and the architecture design of the DW:

3.7.2.1 Initial analysis

Before designing our DW, we studied the current situation and came up with the following assumptions:

i) Central bank of Sudan (CBOS) has an electronic returns system that it uses to communicate with the commercial banks and other financial institutions. The system uses spreadsheets files (namely MS Excel files) that get sent back and forth.

ii) The incoming excel file is examined by the employee in-charge, who can either accept or reject the files.

iii) Banks send their financial ratios in separate files (hard copy). The Prudential Supervision Department (PSD) prepare their own ratios by examining the whole documents related to a specific ratio in a specific period of time and compare the final results to those of the bank.

The process of creating the financial ratio reports is cumbersome; It takes a 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 by automating the calculation of these ratios while loading the raw data from its sources using ETL capabilities.

3.7.2.2 DW design

DW design started by identifying the source of information involved in calculating the selected financial ratios. Then information required to calculate the ratios from the SMEs in the targeted department (PSD) was gathered 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 landing zone (LZ) 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, section 3.7.2.3 covers all the data sources incorporates in the ETL process.

A second layer has been designed (staging area) which performs all the data quality tasks which will be covered in next section as well as writing all the calculation expressions of the financial ratios.

From the staging database, DW is cloned to be the last destination of our loading process shaping the star schema (Fig. 3.8)

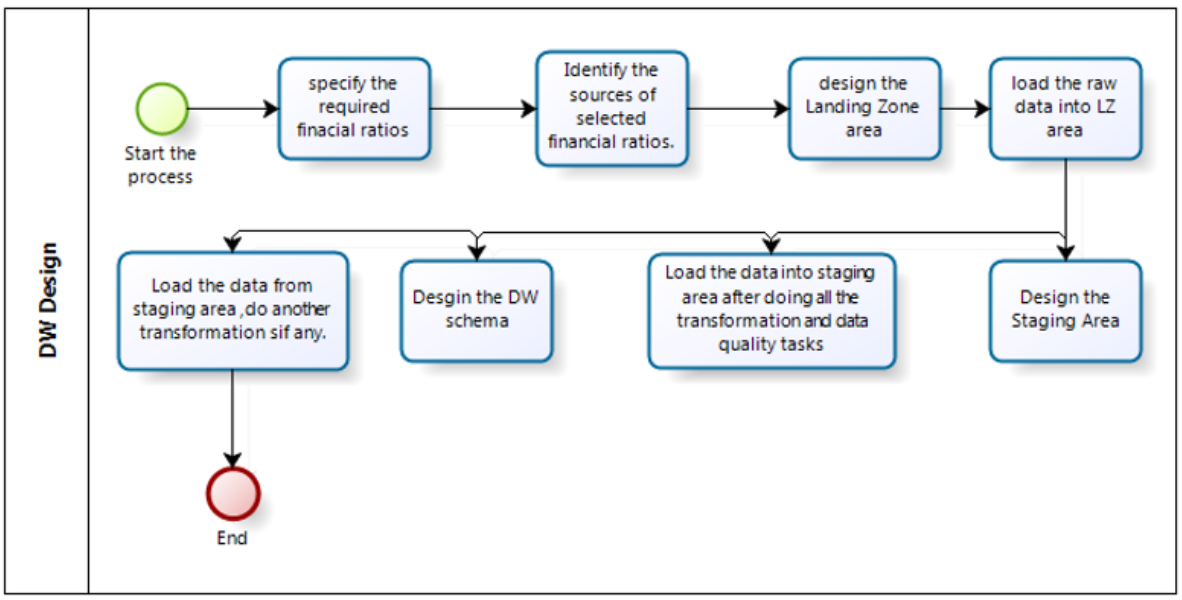

Figure 3.9: DW Design Process.

3.7.2.3 Data Quality Tasks

The main objective of a data integration solution is to combine data from one or more data sources. As you bring data together, you are likely to find a several data quality issues that require processing. For example, you may discover missing customer profile information, such as blank phone numbers or addresses. You may also uncover incorrect data, such as a customer who lives in the city of Sudan which it's a country not city. As the number of data sources that must be integrated increases, data quality issues also increase in number and complexity. Perhaps one of the most challenging data quality issues is duplication of data. Duplication arises when there are conflicting representations of the same entity across source systems. For example, the same bank data may be stored in the Bank Return System, Core Banking System, and Liquidity Management Systems and have different bank codes and names. This conflicting data makes it difficult to recognize identical banks, especially across thousands or millions of source records. The successful handling of data quality issues like these requires a flexible data quality strategy that can be applied to a broad range of issues and integration scenarios. In this research we utilized SQL Server Integration Services to accomplish these tasks; the strategy involves three key tasks of data quality solutions: profiling, cleansing, and auditing.

Profiling

When we profile data before integration, we proactively assess whether a source data extract satisfies the baseline quality standards of data integration solution. By establishing and enforcing baseline quality indicators, we determine whether it is worthwhile to execute the data integration processes using data in its current state.

The first step is to choose baseline indicators. Baseline indicators are metrics or conditions that assess the quality of the entire source data set rather than focusing on the data quality issues of a specific record.

A good rule of thumb for selecting indicators is to identify any condition or issue that would cause to stop integration processing and force to start again from the beginning.

Our rules cover the basic conditions like following:

- The record count of the source file must be greater than zero.
- NULL bank names are not allowed.
- All bank balance sheet dates must be less than or equal to the current date.
- No duplicate unique identifiers are allowed.

To satisfy these requirements, SSIS provides several functions to support the profiling of our source data sets according to custom business rules.

The Data Flow task in particular provides three useful transformations—Row Count, Multicast, and Conditional Split—that can be used together to gather data quality information.

- The *Row Count* transformation allows counting records at any point in the SSIS data flow and storing that record count as an SSIS variable that can then be used for further processing.
- The *Multicast* transformation permits using a single data set for a variety of data quality checks without having to repeatedly read data from the source.
- The *Conditional Split* transformation enables filtering the SSIS data flow based on specific conditions for each quality indicator, such as NULL data values, duplicate identifiers, or invalid data ranges.

Cleansing

After thoroughly profiling our source data quality, we can use data cleansing to ensure that integration solution processes data according to the highest quality standards. On a column-by-column and record-by-record basis, data cleansing enforces the business and schema rules of the application for each source record.

When a rule is violated, there are three choices:

- Fix the data issue by using business logic in our solution.
- Discard the record and continue processing.
- Stop processing.

Following are some examples of common data issues:

- **Missing data values** when data is missing, we may be able to retrieve the data from another data source. For example, if we are missing a bank name but have the bank code, we may be able to look up the name from the master banks data reference data source. If you do not have a data source, you may have business rules that determine how to derive the missing data.
- **Data duplicates** Data duplicates are easy to spot when a unique identifier exists. When no unique identifier exists, it becomes more difficult to spot them. To overcome this challenge, fuzzy logic can be used to perform imprecise data matches which can eliminate data duplicates.

Finding Duplicate values might be quite easily accomplished with simple queries, we did that using partition window functions, finding other inaccurate data might involve some manual work; such like checking the accuracy rules given by prudential supervision department (PSD) employees who are generally gives guidance about the data correctness:

- T1C (Tier 1 Capital) always should by less than TCR (Total Capital Ratio).
- ROA (Return on Asset) and ROE (Return on Equity) shouldn't be zero.
- LAS (Net Loan to Total Asset) and LLP (Loan Loss Provision) shouldn't be negative.

- **Inconsistent data formats** in some cases, data is not in a format that can be integrated with other sources. For example, if we have a multi-value bank address field to combine with a data source that stores normalized address data, we need to develop logic to extract the street address, city, state and other information.

To satisfy these requirements, SSIS provides a wide range of data cleansing functions to cleanse our source data sets according to general or specific business rules. In particular, the SSIS Data Flow task provides these key capabilities:

- **Reassigning column values** to detect NULL, missing, or incorrect data values, SSIS provides the ability to compare incoming data to a validated reference data set by using a Lookup transformation. SSIS also provides the ability to reassign values by using custom expressions in a Derived Column transformation.
- **Handling data duplicates** With the Fuzzy Lookup and Fuzzy Grouping transformations, SSIS provides the ability to perform imprecise data matches. The Fuzzy Lookup transformation in particular is great for matching dirty source data to a known set of cleansed, standardized data in a reference table.

Auditing

Auditing provides proof that our data integration solution satisfies necessary business, technical, regulatory standards. More specifically, auditing serves the following purposes:

- **Data lineage trail** on a record-by-record and column-by-column basis, we can track all data integration operations such as inserts, updates, and deletes. We can also track any data quality issues that encountered while executing our solution along with the action taken to resolve the issue.
- **Data validation** to ensure that we have successfully processed all data; we can use auditing to perform data validation comparisons between sources and destinations.
- **Data execution statistics** Data execution statistics help to track the overall data quality of our integration solution. We can track the success, failure, and execution duration of every component of the integration solution.

Data Warehouses can be architected in many different ways, depending on the needs of the business model and due to limitations of single layer architecture as explained in section 4.3 the tow-layer architecture has been selected as appeared in figure 4.4

In short, data was moved from databases used in operational systems into a data warehouse staging area, then into a data warehouse and finally into a set of conformed data marts. The data was copied from one database to another using a technology called ETL (Extract, Transform, and Load).

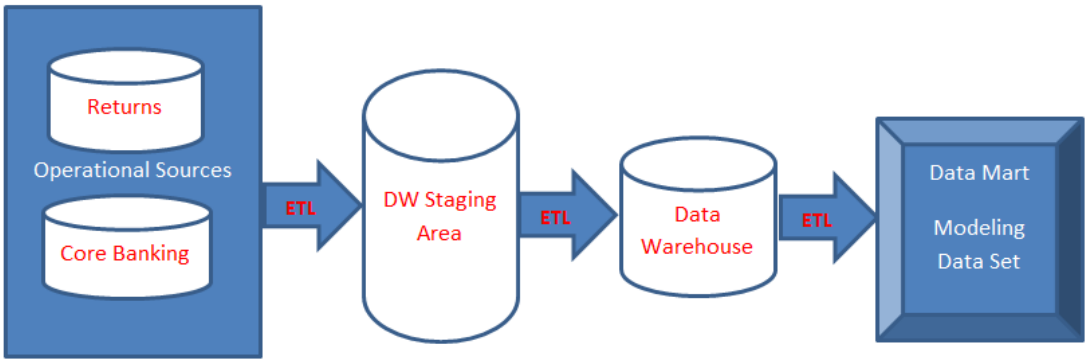

Figure 3.10: High –Level Data warehouse architecture (Two-Layer).

The Following explains the proposed DW architecture

3.7.2.4 Operational Sources

Operational databases are normally "relational" not "dimensional". They are designed for operational, data entry purposes and are not well suited for online queries and analytics.

We have two main operational sources for the DW, the first is the returns, which contains the commercial banks information, and the second is CBOS core banking system.

The challenge for Data Warehousing is to be able to quickly consolidate, cleanse and integrate data from multiple, disparate databases that run on different technical platforms. Below we can find a summary of our DW sources:

Source	Description
BBS_A	Commercial bank's primary balance sheet information
BBS_B	Commercial bank's secondary balance sheet information
FinData	All banks finance information such like credit and instalments
LKP_Bank	Used to load the banks dimension
A_CAP_Sum	Include the information required to calculate risk weighted asset.

Table 3.4: DW Sources.

The following explains the details about each source

- **Data Sources Meta Data**

Before describing the details of each data source, differences between BBS_A and BBS_B are explained in following

Balance Sheet (BBS_A)

The legend of a balance sheet lists assets and liabilities. Assets are resources owned by a bank that have value. Some assets examples include cash, real states, equipment and products. Liabilities are the obligations of a bank. One simple method to memorize the difference between assets and liabilities is to comprehend that assets are things owned while liabilities are things owed.

Income Statement (BBS_B)

The legend of an income statement lists profits and losses for a specified period of time. So, it is a log of a company's profitability for that time period. Profits may come from a business's primary activities or from secondary activities such as profits earned on running business. Losses include operating expenses as well as other losses. For example, if a bank sells one of its assets for less than the book value, the difference between the book value and the sales price is recorded as a loss.

BBS_A: this source contains the information about the primary balance sheet items for the commercial banks; many of those items are required in ratio calculation process.

Structure

Table 3.5 presents the structure and format of BBS_A data source, extract the required data with following properties: this data source is updated monthly which mean we should do the same of our incremental loading process to the ware house. (BUS_CUT_OFF_DATE) is holding the date of update by specific bank (BR_CD).

Field (ITEM_EN) contains the balance sheet items with equivalent code (CD_EN) , so we fetch the value of that item from the field (RL_CUR) which indicate the value of resident customers in local currency.

The concept of resident is very important to distinguish between the nature of work of internal banking sector monitoring (Residents) and external balance of payments monitoring (Non-residents). The customer is considered resident if he contributes in the economic for more than one year even if he's foreigner. Other fields (FL_CUR, NRL_CUR, and NRF_CUR) are considered not applicable because they present the amounts of foreign or non-resident customer operations.

Column name	Data Type	Description	Extracted?
BR_CD	varchar(10)	Bank Code	Yes
CD_EN	nvarchar(255)	Item Code	Yes
ITEM_EN	varchar(120)	Item description	Yes
RL_CUR	nvarchar(255)	Amount-Resident Local	Yes
FL_CUR	numeric(18, 0)	Amount-Resident Foreign	NO
NRL_CUR	numeric(18, 4)	Amount-Nonresident Local	NO
NRF_CUR	numeric(18, 4)	Amount-Nonresident Foreign	NO
DL_ADD_DATE	Datetime	Insert Date	Yes
BUS_CUT_OFF_ DATE	Datetime	Business Date	Yes
APPROVED	varchar(1)	Acceptance flag	Yes

Table 3.5: Balance Sheet Structure

ITEM_EN contains our required items for ratio calculation as described in table 3.3, so we are going to extract just those items from this source based on its code (CD_EN).

APPROVED field it's an acceptance indicator, which means this record has been accepted by the employee of central bank or not. To control the quality of our extracted data we just fetch the records that indicate the acceptance of commercial banks data (APPROVED='y'). Ultimately, we will extract the facts (Paid-up Capital, Legal Reserve, General Reserve , Retained Earnings , Month Loss , Revaluation Reserve , General Provision , Equity Participation , Total Asset , Reserves on Central Bank , Loans , Provision for Bad Loans , Cash , Current Deposits , Short-Term Securities , Deposits) for the sake of calculate the ratios (T1C, TCR , EAS , LAS , LLP).

Each fact is extracted based on certain code number (CD_EN). The Actual values are presented in table 4.12 DW facts.

- **BBS_B**: contains the information of secondary balance sheet items opposite to BBS_A which contains the primary one, secondary items are mainly the income statement and expenses of each commercial bank.

Structure

Table 3.6 presents the structure and format of BBS_B data source; we extracted the required data with following properties: data is updated monthly indicated by field (BUS_CUT_OFF_DATE) which means we have to do the same as incremental process for data warehouse. This data is uploaded by specific bank indicated by field (BR_CD) , containing the Income/Expenses related items described in field (PARTICULARS_EN) with equivalent code (CD_EN) providing the value of that fact in (AMT_EN).

As performed with BBS_A source, to guarantee the minimum data quality we have to extract just the rows approved by CBOS users. Ultimately, we

will extract the facts (Income from finance, TotalIncome, NetIncome) for the sake of calculating the ratios (IDIVER, ROA, ROE)

field name	Data Type	Description	Extracted?
BR_CD	varchar(10)	Bank Code	Yes
CD_EN	nvarchar(255)	Item Code	Yes
PARTICULARS_EN	varchar(120)	Item description	Yes
AMT_EN	Numeric(18, 4)	Amount-Resident Local	Yes
DL_ADD_DATE	Datetime	Insert Date	Yes
BUS_CUT_OFF_DATE	Datetime	Business Date	Yes
APPROVED	varchar(1)	Acceptance flag	Yes

Table 3.6: Income Statement Structure.

FinData: expose all the information related to bank finance operations such as credits, customer names, installments frequency, finance mode and non-performing loans.

Structure

Table 3.7 presents the structure and format of FinData data source. We extracted the required data which is the non-performing loan that required by the ratio NPL. The source contains useful information which can be used to deeply analyze the relationship between bank credit operation, the used finance modes and payment method with credit risk and bankrupt factors.

The data from this source is updated monthly, which means the same for the incremental loading to data warehouse using the approved column monthly to control the data quality.

Field name	Data Type	Description	Extracted?
BR_CD	varchar(10)	Bank Code	Yes
CSTMR_NAME	nvarchar(255)	Customer Name	No
CURRENCY	varchar(50)	Currency type	Yes
INSTALLMENT	numeric(20, 4)	Installment Amount	No
PAYMT_METHOD	nvarchar(20)	Credit payment method	Yes
MODE_OF_FIN	nvarchar(30)	Credit Finance mode	Yes
SCTR	nvarchar(120)	Sector of finance	Yes
STATUS	nvarchar(80)	Preforming/Non-Performing	No
OS_BAL	numeric(20, 4)	Outstanding balance	No
TOTAL_NON_PERF	numeric(20, 4)	Amount of Non-Performing	Yes
DL_ADD_DATE	Datetime	Insert Date	Yes
BUS_CUT_	Datetime	Business Date	Yes

OFF_DATE			
APPROVED	varchar(1)	Acceptance flag	Yes

Table 3.7: FinData Structure

LKP_Bank: this source is simply used to load the Bank's dimension in data ware house

Structure

Table 3.8 presents the structure and format of LKP_Bank data source; only required data has been extracted.

Field name	Data Type	Description	Extracted?
CODE	Bigint	Bank Code (Number)	Yes
DESC_EN	nvarchar(255)	English Description	Yes
DESC_AR	nvarchar(255)	Arabic Description	Yes
Bank_CD	Varchar(4)	3 letter bank code	Yes

Table 3.8: LKP Bank Structure.

Bank dimension is considered one of the slowly changing dimensions (SCD) with type 2 SCD which means preserving the history while adding new rows. When you implement Type 2 SCD, for the sake of simpler querying, you typically also add a flag to denote which row is current for a dimension member. Alternatively, you could add two columns showing the interval of validity of a value. The data type of the two columns should be Date, and the columns should show the values Valid From and Valid To. For the current value, the Valid To column should be NULL.

A_CAP_Sum: Risk Weighted Asset is a main component to calculate the ratios (T1C, TCR). All the information required to calculate this component is available in this source.

Structure

The values (VALS) represent specific item (Item_Funding) which indicate the assets information for the bank (BR_CD) having specific item code (SEQ_NO).

Field name	Data Type	Description	Extracted?
BR_CD	varchar(10)	Bank Code	Yes
SEQ_NO	bigint	Item code	Yes
Item_Funding	nvarchar(255)	Items description	Yes
VALS	numeric(20, 4)	Amount of item funding	Yes
DL_ADD_DATE	Datetime	Insert Date	Yes
BUS_CUT_OFF_DATE	Datetime	Business Date	Yes
APPROVED	varchar(1)	Acceptance flag	Yes

Table 3.9: A_CAP_Sum Structure

3.8 Data warehouse

The purpose of the data warehouse is to integrate all facts required to calculate the CAMELS ratios which are considered the ultimate data sets.

The data is stored at a lowest level of detail available. For example, instead of calculating the ratio directly during the load of data, we can extract the relevant facts of such ratios, store them separately, and join them at the end of load operation. This will allow the data to be sliced and diced, then summed and grouped in unimaginable ways for any other reason besides calculations of the CAMELS ratios; which is now convenient by using the joined table with relevant transformations.

Simply data warehouse is composed from set of dimensions and fact tables, the following explains those components

3.8.1 Dimensions

According to requirements, two dimensions are needed in the data warehouse which are:

- Date
- Banks

Implementing a dimension involves creating a table that contains all the needed columns. In addition to business keys, we should add a surrogate key to all dimensions that need Type 2 Slowly Changing Dimension (SCD) management. We should also add a column that flags the current row or two date columns that mark the validity period of a row when you implement Type 2 SCD management for a dimension.

Those dimensions are containing following information

DimDate		**DimBank**	
Field	Description	Field	Description
DateSk	Surrogate key	BankID	Surrogate key
FullDateAltKey	Standard	Code	Bank

	Date		Code
DayNumberOfWeek	Weak Day Number	Desc_EN	English name
DayNumberOfMonth	Month Day Number	Desc_Ar	Arabic name
DayNumberOfYear	Year Day Number	Bank_CD	3 letters bank code
WeakNumberOfYear	Year Weak Number		
MonthNumberOfYear	Year Month Number		
CalendarQuarter	Quarter		
CalendarYear	Year		

Table 3.10: Dimensions Details.

Facts

After dimensions have been implemented, fact tables should be designed. Fact tables should always be built after dimensions. A fact table is on the "many" side of a relationship with a dimension, so the parent side must exist if a foreign key constraint might be applied.

Basically our fact table representing the ratios which calculating as a result of the merging process explained in the previous section, the two dimensions (DimDate , DimBank) they related to our ratio fact table through their surrogate keys which are generated during the staging area loading process.

A ratio fact table is exposes the following properties:

Field	Data Type	Description
BR_CD	Bigint	foreign key - DimBank
Year	Bigint	foreign key - DimDate
T1C	numeric(20, 4)	Tier 1 Capital Ratio measured as a ratio of Tier 1 Capital to Risk Weighted Assets.
TCR	numeric(20, 4)	Total Capital Ratio measured as a ratio of (Tier 1 + Tier 2 capital) to Risk Weighted Assets
EAS	numeric(20, 4)	Ratio of Equity Capital to Total Asset.
LAS	numeric(20, 4)	Ratio of Net Loans to Total Assets.
LLP	numeric(20, 4)	Ratio of Loan Loss Provisions to Total Loans.
NPL	numeric(20, 4)	Ratio of Non-Performing Loans to Total Loans.
ROE	numeric(20, 4)	Return on Equity Measured as a ratio of Net Income to Capital Equity.
ROA	numeric(20, 4)	Return on Assets measured as a ratio of Net Income to Total Assets.
LADF	numeric(20, 4)	Ratio of Liquid Assets to Deposits and Short Term Funds.
IDIVER	numeric(20, 4)	Finance-Related Income to Total Income

. Table 3.11: Ratios Fact Table.

As a result of our ETL task we have identified new facts which store the required values used later in defining our DW using the transformation solutions.

Following is the list of identified facts:

Class	Ratio	Facts	Source
Capital	T1C / TCR	Paid-up Capital	BBS_A(RL_CUR) ; CD_EN=4110
	T1C / TCR	Legal Reserve	BBS_A(RL_CUR) ; CD_EN=4130
	T1C / TCR	General	BBS_A(RL_CUR) ;

		Reserve	CD_EN=4140
	T1C / TCR	Retained Earning	BBS_A(RL_CUR) ; CD_EN=4150
	T1C / TCR	MonthLoss	BBS_A(RL_CUR(-)) ; CD_EN=4440
	T1C / TCR	Risk Weighted Asset*	A_CAP_Sum-Calculated Item
	TCR	Revaluation Reserve	BBS_A(RL_CUR) ; CD_EN=4440
	TCR	General Provision	BBS_A(RL_CUR) ; CD_EN=4120*.45
	EAS/LAS	Equity Participation	BBS_A(RL_CUR) ; CD_EN=1960
	EAS/LAS/ROA	TotalAsset	BBS_A(RL_CUR) ; CD_EN=100
Asset	LADF	Reserves on Central Bank	BBS_A(RL_CUR) ; CD_EN=1110
	LAS/NPL/LLP	Loans*	BBS_A - Calculated Item
	LAS	Provision for Bad Loans	BBS_A(RL_CUR) ; CD_EN=4010
	NPL	Non-Performing Loans	FinData(TOTAL_NON_PERF) ;Status='Non-Performing'
Profitability	ROE/ROA	Net Income *	BBS_B - Calculated Item
Liquidity	LADF	Cash	BBS_A(RL_CUR) ; CD_EN=1000
	LADF	Current Deposits *	BBS_A - Calculated Item
	LADF	Short-Term Securities*	BBS_A - Calculated Item
	LADF	Deposits*	BBS_A - Calculated Item
Other	IDIVER	Income from finance	BBS_B(AMT_EN) ; CD_EN=5040
	IDIVER	Total	BBS_B(AMT_EN) ;

		Income	CD_EN=590
* Calculated Item			

Table 3.12: DW facts.

Star schema was used to design the final data warehouse. Dimensions surround the fact table to enable the aggregations in every fact table values. In the designed warehouse we rely on two dimensions: DimDate and DimBank as it's shown in Figure 4.5. To tackle the obstacle of data unavailability in other dimensions, the data warehouse can be extended to include those new ones such like branches and sectors.

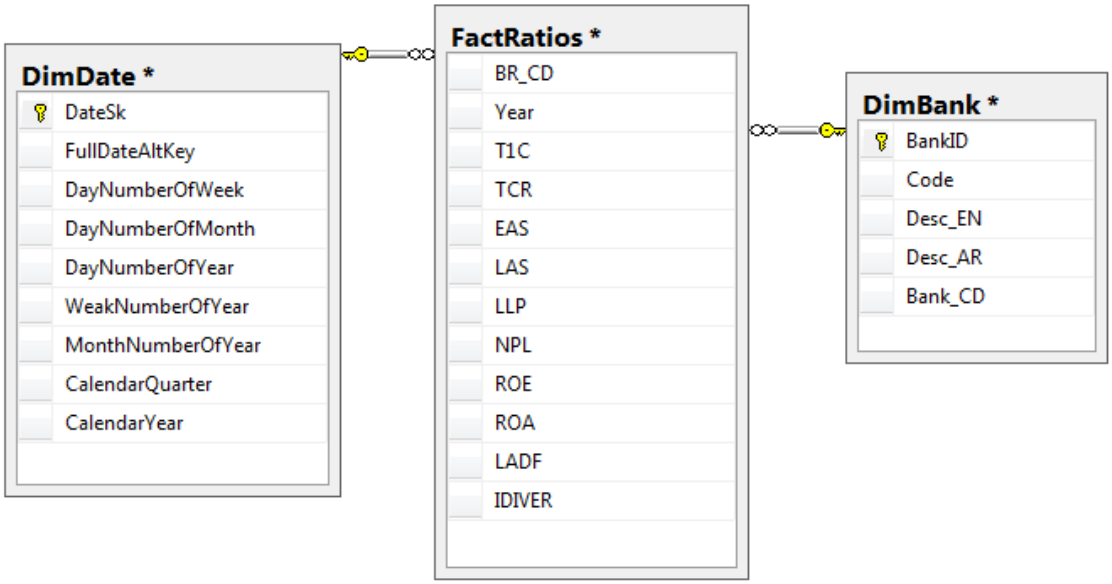

Figure 3.11: Star Schema Design for DW.

Now, after data has been ready and stored on final fact table (FactRatios) we have used MS Excel to connect to DW and analyze our final output using Pivot Tables features. Although lot of data quality transformation tasks has been performed as explained in section 3.7.2.3, the final data was found containing some anomalies and wrong values (Logical errors) , so we had to revise the whole cycle again, only to discover following issues:

1. Sometimes banks intentionally send incorrect information to obscure some facts from CBOS. Those incorrect values weren't discovered by double-checking each fact such as minimizing the volume of their total deposits to have more flexibility on liquidity ratios.
2. Some records have been approved by CBOS, after realizing that they are wrong, they didn't reject those records.

To solve the above-mentioned issues we have to design and integrate a solution to extract the data directly from data stores of commercial banks without any manual intervention. When we began to execute this solution we realized that there is no law that authorizes the Central Bank to directly access the commercial banks data stores. So we had to exclude all doubtful records from final data set.

There were 6 calculated facts that are not directly provided from their sources through the item code (CD_EN). In the following, we provide their calculation formulas:

Risk Weighted Asset =

$$(\sum(\mathrm{VALS}) \ , \forall \mathrm{SEQ} - \mathrm{NO} = 7) - (\sum(\mathrm{VALS}) \ , \forall \mathrm{SEQ} - \mathrm{NO} = 9) - .5 * (\sum(\mathrm{VALS}) \ , \forall \mathrm{SEQ} - \mathrm{NO} = 9) - \ .5 * \big((\sum(\mathrm{VALS}) \ , \forall \mathrm{SEQ} - \mathrm{NO} = 11) + (\sum(\mathrm{VALS}) \ , \forall \mathrm{SEQ} - \mathrm{NO} = 12)\big) \quad (3.1)$$

Loans =

$$\sum(\mathrm{RL\_Cur}) \ , \forall \mathrm{CD\_EN} \ \mathrm{in}(1140 \ ,1145, 1241,1250,1340,1350,1400,1420, 1500,1600,1670,1800,1870,1890,1898) \quad (3.2)$$

Net Income

$$=(\sum(\mathrm{RL_{Cur}}) \ , \forall \mathrm{CD_{EN}} = 590) - (\sum(\mathrm{RL_{Cur}}) \ , \forall \mathrm{CD_{EN}} \ \mathrm{in} \ (6010,6020,6030,6040,6050,6060)) \quad (3.3)$$

Current Deposits

$=\sum(\mathrm{RL}-\mathrm{Cur})\ ,\forall \mathrm{CD}_{\mathrm{EN}} \mathrm{in}\ (1120,1210,1310)$ (3.4)

Short-Term Securities

$=\sum(\mathrm{RL}-$

$\mathrm{Cur})\ ,\forall \mathrm{CD}_{\mathrm{EN}} \mathrm{in}\ (1145,,1250,1350,1421,1422,1452,1520,1601)$ (3.5)

Deposits = $=\sum(\mathrm{RL}-\mathrm{Cur})\ ,\forall \mathrm{CD\_EN}$ in $(3000,3100,3300,3400)$ (3.6)

In the following we explain the used symbols

Symbol	Description	Source
VALS	The values (VALS) represent specific item (Item_Funding) which indicate the assets information for the bank (BR_CD) having specific item code (SEQ_NO).	A_CAPS_ Sum
SEQ_NO	Item code	A_CAPS_ Sum
RL_Cur	Amount-Resident Local	BBS_A
CD_EN	Item Code	BBS_A

Table 3.13: Equations Description Symbols

Each fact in table 3.12 has been exported in separate table along with bank code and business date, the table will take the fact name then all facts will be merged using those two attributes (Bank, date) to create the unified fact source containing all the required arguments for the ratios calculation.

3.8.2 Staging Area

The Data Warehouse Staging Area is a temporary location where data from source systems is copied. It is mainly required in a Data Warehousing Architecture to assist in committing by time window given by the data

sources administrator. In short, all required data must be available so that can be integrated into the Data Warehouse.

However, Due to varying business cycles, data processing cycles, hardware and network resource limitations factors, it is not feasible to extract all the data from all operational databases at exactly the same time.

We also utilized the staging area also to cleanse and match data before we load it into the data warehouse.

We can say that staging area it's simply our final data warehouse but with two different aspects:

1. Staging area is used to apply all the rules of data quality issues introduced in the previous section to not overwhelm the data warehouse loading performance.
2. We generate all the dimension surrogate keys while we load our dimensions into staging area to be just cloned in the warehouse.

Since most of our data sources are refreshed monthly we designed the schedule to stage our data by the 8th of every month to guarantee that all the commercial banks have uploaded the required data.

3.8.4 DW Storage Model

We used MS SQL Server 2012 as the DW back-end designing software for its compatibility with the selected DW designing tool. It also offers many easy-to-use features such as the import and export to excel files, which was useful while performing data exploration tasks.

Proposed Solution

The design followed the best practices on data warehouse approaches with some customization depend on the sole format found on the raw date of Sudan's banking sector.

The solution will offer automatic ratio preparation customized to Sudan monetary authority while loading the raw data from its sources which can ease the building of various predictive models.

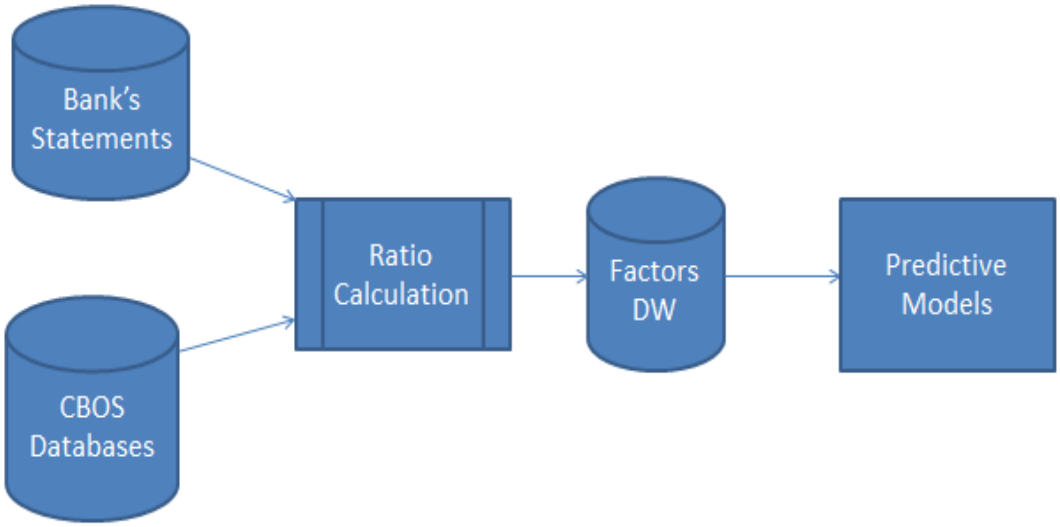

Figure 3.12: Factors DW Conceptual Design

3.9 Evaluation and Reporting Results

The reporting of the result in phase 1, 3 will be discussed, evaluated by following:

For the Initial study of the Sudan's banking sector we will rely on the expert's judgment and the procedure heritage existed on how the monetary authorities are perceived the factors that might be the major source of financial distress for the banks operated in Sudan.

To evaluate the classification results, standard measures will be used to assess the classifiers performance; these measures consist of Accuracy, Sensitivity and Specificity

3.10 Summary

As a conclusion, this report has described a research methodology and operational framework used in completing prediction of Sudan's distressed banks research using different features selection techniques as well as classification models. There are fifth phases in this operational framework, namely; Phase 1: initial study and data preparation, Phase 2: Features Selection using hybrid filter-based and wrapper based technique, Phase 3: Ensemble optimized support vector machine (SVM) model, Phase 4: Islamic credit risk analysis And Phase 5: Research Writing. Each of these stages and phases play an essential role in achieving the main objectives of this research.

Chapter 4

Features Selection Techniques of Bank Distress Classification

4.1 Introduction

Feature selection methods can be used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. Less attributes is desirable because it reduces the complexity of the model, and a simpler model is simpler to understand and explain.

The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data.

Though many novel sophisticated techniques have been proposed for effective prediction, very few have examined the effect of feature selection on financial distress prediction. Feature selection is an important data pre-processing step of knowledge discovery in databases (KDD). The aim is to filter out unrepresentative features from a given dataset (Guyon, 2003). As there are no generally agreed financial ratios for bankruptcy prediction and credit scoring, collected variables must be examined for their representativeness, i.e., importance and explanatory power, in the chosen dataset [29]. Therefore, the performance of classifiers after performing feature selection could be enhanced over that of classifiers without feature selection.

Choosing a subset of variables from an initial set is essential to the development of a model. First of all, it is essential to the parsimony of the model because, as we mentioned above, generalization requires parsimony, and parsimony is directly related to the number of variables included in a model. It is also essential for the accuracy of the model because, generally, not all variables contribute equally to its performance. Some may be less informative, others noisy, meaningless, correlated and thus redundant, or irrelevant. The aim of a selection process is therefore to find a subset of relevant variables for a given problem, composed of elements as independent as possible, and sufficiently numerous to account for the problem (Deron, 2015).

There are two main models that deal with feature selection: filter methods and wrapper methods (Kohavi, 1997). While wrapper models involve optimizing a predictor as part of the selection process, filter models rely on the general characteristics of the training data to select features with independence of any predictor. Wrapper models tend to give better results but filter methods are usually computationally less expensive than wrappers.

In this chapter, proposed hybrid feature selection method will be discussed. The best filter based feature selection will be identified and will be combined with the best wrapper based one. The selection of these methods will be depending on existing researches that have shown satisfying performance (Tsai, 2009). Filter features selection techniques are proposed which is discriminant analysis (DA) and principal component analysis (PCA). Wrapper based feature selection proposed are genetic algorithm (GA) and particle swarm optimization (PSO). Experiments have been performed using data warehouse factors of Sudan's banking sector. Standard statistical measurements of specificity, sensitivity and accuracy have been used to evaluate the classification results. . This chapter is organized as follows:

Section 4.2 explains the overview of the suggested methodology. The experimental results and discussion are shown in Section 4.3. Finally, the chapter summary is presented in Section 4.4.

4.2 Suggested Methodology

This section purposes to provide an overview of the suggested approach; in order to carry out dimensionality reduction task to features used in bank distress prediction classification existing studies have been reviewed and came up with techniques such like DA, PCA, PSO and GA based on eleven financial ratios stored in data warehouse. Figure 4.1 displays an overall diagram of the proposed approach. After subject matter experts identified the optimal set of financial ratios as well as rain fall ratio that proposed by researcher, these factors will be passed as parameter to next component. The next component performs feature selection techniques which will reduce the number of selected predictors to a set of strong predictability power that can be used in subsequent classification task. The process will be divided in two tasks: first the best filter based feature selection will be identified and similarly for wrapper based. The better of two methods will be combined to propose the hybrid feature selection that can better select the strong ratios to be used in subsequent tasks. The output of this component will be passed to next one which is consist of two classifier (support vector machines and neural network) to evaluate the result set of reduced predictors. More details on feature selection techniques used will be given on the following sections.

4.2.1 Filter based feature selection techniques

The filter based feature selection methods usually contain the following procedures. Given a dataset, the method based on a particular search strategy

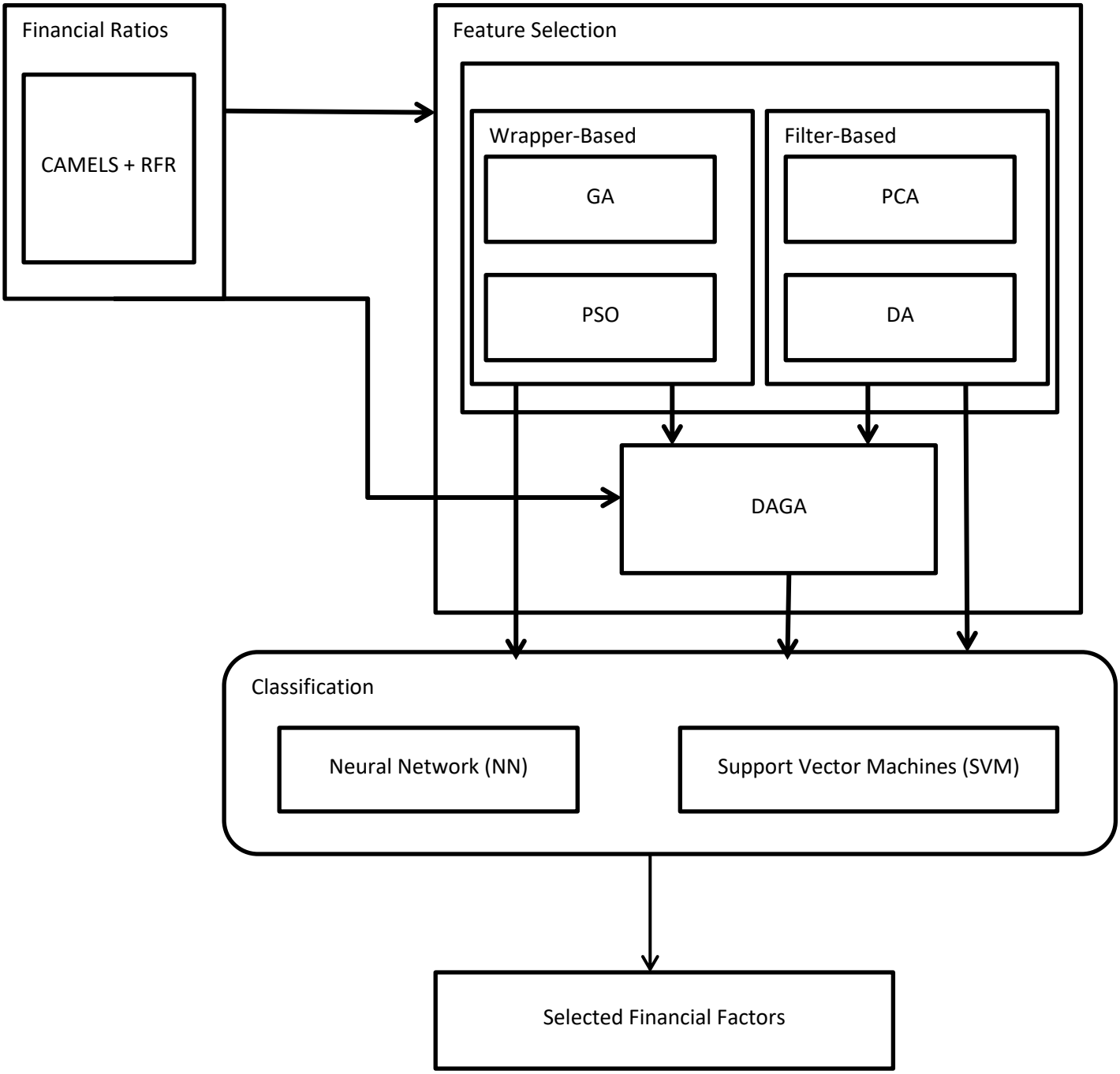

Figure 4.1: Suggested Methodology overview.

Initially searches from a given subset, which may be an empty set, a full set, or any randomly selected subset. Then, each generated subset is evaluated by a specific measure and compared with the previous best one. This search process iterates until the pre-defined stopping criterion is met. Consequently, the final

output of this method is the last current best subset. More specifically, the search strategy and evaluation measure can be different depending on the algorithms used. In addition, filter based methods do not involve any mining algorithm during the search and evaluation steps, they are computationally efficient. Some examples of filter based methods that are used in financial distress prediction are based on statistical techniques, such as t-testing, principal component analysis, discriminant analysis, and regression (Chandra, 2009).

4.2.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) also known as the Karhunen-Loeve Transform is a classical statistical method. It identifies the axes for a set of data vectors along which the correlation between components of the data vectors can be most clearly shown. It is one of the popular methods used in financial distress prediction (Canbas, 2005).

Suppose there is a data set $M = (x_i | i = 1,2, \dots, N)$ where X is an n-dimensional column vector and $X = (x_1, x_2 \dots, x_N)^T$. The mean of the data vector is $\mu = < x >$, here < > stands for the average over the data set. The data set can be represented by a matrix $D = (x_1, x_2 \dots, x_N)$ .The covariance matrix of D is C with its element C_{ij} which can be calculated as shown below (Deco, 1996) .

$$C_{ij} = < (x_i - \mu_i) - (x_j - \mu_j) > \quad (4.1)$$

By solving the characteristic equation of the covariance matrix C, we can obtain the eigenvectors that specify the axes having the properties described above and the corresponding eigenvalues that are respectively indicative of the variance of the dataset along these axes. Therefore, just by looking at the eigenvalues, we can easily find out along which axes the dataset has little or no spread. Hence, the principal axes and the eigenvalues give a good reflection of

the linear interdependence between the components of the data vectors. By choosing some eigenvectors that have the largest eigenvalues, we can form a subspace A in which the data set has the most significant amounts of variance. Thus, the dimensionality of the data can be reduced by means of this property of PCA. The logical steps of PCA is shown in algorithm 4.1

1. **Function** PCA(V)
2. **Input** Ratio v;
3. Calculate the mean of data v
4. Compute the covariance of the data v.
5. Calculate the orthogonal eigenvectors $\emptyset_j$.
6. Sort the $\emptyset_j$ in descending order.
7. Choose r from the top eigenvectors $\emptyset_j$.
8. **End function**

Algorithm 4.1: PCA algorithm.

4.2.1.2 Discriminant Analysis (DA)

Linear Discriminant analysis (LDA) is used to find a linear combination of features which characterizes or separates two or more classes of objects. The resulting combination can be used for dimensionality reduction. LDA can also be used to express one dependent variable as a linear combination of other features. In other words, LDA looks for the linear combination of features which best explains the given data [34]. LDA involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is

$$D = v_1 x_1 + v_2 x_2 + \cdots v_i x_i + a \qquad (4.1)$$

Where D is the discriminant function, v is the discriminant coefficient or weight for that feature, x is the respondent's score for that feature, a is a constant, and i is the number of predictor features.

LDA steps can be summarized in following algorithm

1. **Function** LDA(X)
2. **Input** Ratios Data set X ;
3. Compute the d-dimensional mean vectors for the different classes from the dataset.
4. Compute the scatter matrices (in-between-class and within-class scatter matrix).
5. Compute the eigenvectors $(e_1, e_2, \ldots . e_d)$ and corresponding eigenvalues $(\gamma_1, \gamma_2, \ldots . \gamma_d)$ for the scatter matrices.
6. Sort the eigenvectors by decreasing eigenvalues and choose *k* eigenvectors with the largest eigenvalues to form a *d X k* dimensional matrix *W (* where every column represents an eigenvector).
7. Use this d X k eigenvector matrix to transform the samples onto the new subspace.
8. **End function**

Algorithm 4.2: LDA algorithm

4.2.2 Wrapper based feature selection techniques

Wrapper methods search through the space of feature subsets using a learning algorithm to inform the search. They calculate the estimated accuracy of the learning algorithm for each feature that can be added to or removed from the feature subset. Accuracy is estimated using cross validation on the training set. In forward selection, a wrapper estimates the accuracy of adding each unselected feature to the feature subset and chooses the best feature to add according to this criterion. These methods typically terminate when the estimated accuracy of adding any feature is less than the estimated accuracy of the feature set already selected. The wrapper approaches of feature selection aim to find the minimum discriminative features to reach the high classification accuracy, while the filer approaches are to compute the 'best' subset of features in terms of some criteria (John, 1994). Some examples of wrapper based methods that are used in financial distress prediction are sequential forward selection (Kittler, 1978), sequential backward selection,

particle swarm optimization, randomized hill climbing and genetic algorithms (GA) (Fengyi Lin, 2014).

4.2.2.1 Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is a population based method that inspired from the behavior (information exchange) of the birds in a swarm (Kennedy, 2001).

In PSO the population is called a swarm and the individuals are called particles. In the search space, each particle moves with a velocity. The particle adapts this velocity due to the information exchange between it and other neighbors. At each iteration, the particle uses a memory in order to save its best position and the overall best particle positions. The best particle position is saved as a best local position, which was assigned to neighborhood particles, while the overall best particle position is saved as a best global position, which was assigned to all particles in the swarm. Each particle is represented by D dimensional vectors, $x_i = (x_{i1}, x_{i2}, \dots\ x_{iD}) \in S$

The velocity of the initial population is randomly generated and each particle has the following initial velocity: The velocity of the initial population is randomly generated and each particle has the following initial velocity: $\mathrm{v_i} = (\mathrm{v_{i1}}, \mathrm{v_{i2}}, \dots\ \mathrm{v_{iD}})$.

The best local and global positions are assigned, where the best local position encounter by each particle is defined as $\mathrm{p_i} = (\mathrm{p_{i1}}, \mathrm{p_{i2}}, \dots\ \mathrm{p_{iD}}) \in \mathrm{S}$.

At each iteration, the particle adjusts its personal position according to the best local position (Pbest) and the overall (global) best position (gbest) among particles in its neighborhood as follows:

$$x_i^{t+1} = x_i^{t} + v_i^{t+1}, i = 1 \dots p \qquad 4.2$$

$$v_i^{t+1} = v_i^{t} + C_1 r_{i1} \times \left(\mathrm{Pbest}_i^{t} - x_i^{t}\right) + C_2 r_{i2} \times (\mathrm{gbest} - x_i^{t}), i = 1 \dots p \qquad 4.3$$

Where C_1, C_2 two acceleration constants are called cognitive and social parameters, r_1, r_2 are random vector$\in$ [0,1].

We can summarize the main steps of the PSO algorithm as follows.

1. **Function** PSO(P,C1,C2)
2. **Input** Swarm size (p) , acceleration constants (C_1, C_2) ;
3. Randomly generate initial position and velocity of each solution (particle) in the population.
4. Evaluate each solution by calculate its fitness value$f(x_i)$.
5. Assign best personal and global solutions (Pbest) and (gbest) respectively.
6. **Loop** (Until termination criterion is satisfied)
7. Justify position of each particle using equation 4.2.
8. Justify velocity of each particle using equation 4.3.
9. Evaluate each solution and assign new (Pbest) and (gbest).
10. **End Loop**
11. Produce the best found solution so far.
12. **End function**

Algorithm 4.2: PSO algorithm

4.2.2.2 Genetic Algorithm (GA)

The Genetic Algorithm is a heuristic optimization method inspired by those procedures of natural evolution that the genes of organisms tend to evolve over successive generations to better adapt to the environment (Fogel, 1998). Genetic algorithms operate on a population of individuals to produce better and better approximations. At each generation, a new population is created by the process of selecting individuals according to their level of fitness in the problem domain, and recombining them together using operators borrowed from natural genetics. The offspring might also undergo mutation. This process leads to the evolution of populations of individuals that are better suited to their environment than the individuals that they were created from, just as in natural adaptation. Several issues shall be considered when it comes to the design of an effective GA; this includes the population size, genetic operators

(selection, crossover and mutation), and the stopping criteria. The size of the population has impacts on both the performance as well as the efficiency of the GA, which is usually set from 30 to 200 (Srinivas, 1994).

The reproduction of the current population to the next generation starts with the chromosome selection. Roulette wheel method and tournament method are two standard methods to select those chromosomes that can survive to the next generation from the current population (Back, 1996). All chromosomes that survive to the next generation are placed in a matting pool for crossover and mutation. The chromosomes are randomly selected in pairs from the matting pool for crossover. This probability is referred to as the crossover rate, which typically ranges from 0.5 to 1.0. Commonly used crossover methods are single-point, two-point and uniform crossover (Srinivas, 1994). The newly crossed chromosomes are then combined with the rest of the chromosomes to generate a new population. Following the crossover, the mutation operator produces small changes to the bit string by choosing a single bit at random, then changing its value. The probability that a chromosome is mutated, or the mutation rate, ranges typically from 0.001 to 0.05. Commonly used mutation methods are point mutation, polynomial mutation and uniform mutation. The condition with which the evolution process stops is called the stopping criteria. Commonly applied criteria can be either the convergence to a good solution or a preset number of the evolution rounds.

We can encapsulate the high level steps of the GA algorithm as follows

4.2.3 Classification

In order to carryout experiments, two machine learning algorithms were applied: Polynomial Support vector machines (SVM) and Neural network (NN).

1. **Function** GA(X)
2. **Input** Individuals (x) ;
3. **Loop** (until a stopping criterion is satisfied)
4. Assign the fitness to each individual.
5. Select the individuals according to their fitness level.
6. Recombine the selected individuals to generate a new population using *crossover* operation.
7. Change the value of some features in the offspring at random using *mutation* operation
8. **End Loop**
9. Produce the best subset of Individuals.
10. **End function**

Algorithm 4.2: GA algorithm.

4.3 Experimental design and performance evaluation

RM v.7.3 were used to perform all computations regarding feature selection (PCA,DA,PSO and GA) as well as classification techniques (SVM and NN) with machine setup of Intel Core™ i5-5200 , 2.2 GHz (4 CPUS) processor and 16 GB of RAM.

Selected subset of financial ratios has fed to Polynomial SVM and NN. SVM uses polynomial kernel to represent the ratios vector, the kernel degree was equal 2, C parameter that sets the tolerance for misclassification, where higher C values allow for 'softer' boundaries and lower values create 'harder' boundaries. A complexity constant that is too large can lead to over-fitting, while values that are too small may result in over-generalization so it was determined to be equal to 0.1.

The model by means of a feed-forward neural network trained by a back propagation algorithm (multi-layer perceptron) was configured by 15 hidden layers , the weights has been changed using learning rate equals to .3 and momentum equals to .2 . All these values have chosen by the guidance of trial and best accuracy results. The methodology is validated on the benchmark of dull dataset of financial ratios. In the evaluation, the entire dataset contains 506 records of financial ratios. Classification results are obtained using 5-fold cross-validation and 60% split (i.e. 60% training, 40% testing). To evaluate the performance of the classification, the standard statistical indices of sensitivity, specificity and accuracy were applied.

4.3.1 Filter based feature selection for banks distress prediction

Discriminant analysis (DA) outperformed principal component analysis (PCA) in term of accuracy. However they were close in number of features selected, in best classification accuracy discriminant analysis selected 9 ratios whereas PCA select subset of 8 ratios as explained in Table 4.3 and Table 4.4

4.3.2 Wrapper based feature selection for banks distress prediction

PSO in its best classification accuracy has selected 6 ratios and genetic algorithm has been able to select seven financial ratios. Features are ranked differently in decreasing order based on their importance. Table 4.1 shows the rank of financial ratios selected by PSO feature selection technique in descending order.

All ratios with weight less than 0.5 have been excluded as indication of poor prediction ability , the selected ratios are clearly marked asset and capital CAMELS dimension as good predictors with ratios (ROE,ROA) and (T1C , TCR) respectively , along with liquidity ratio LADF and Non-CAMELS ratio RFR.

Feature	Weight
ROE	**0.96849**
LADF	**0.93664**
T1C	**0.86336**
RFR	**0.69259**
TCR	**0.56462**
ROA	**0.56207**
LLP	0.48466
LAS	0.2205
NPL	0.18245
EAS	0.06876
IDIVER	0.06161

Table 4.1: Feature ranking using PSO feature selection.

4.3.3 Proposed feature selection technique (DAGA-FS)

After examining the results of both filter based and wrapper based feature selection methods, which identified DA and GA as best feature selection methods. However both filter and wrapper based feature selection methods have advantages and disadvantages can be minimized if their strengths has combined together. Table 4.2 shows comparative analysis for both techniques

Advantages		Disadvantages	
Filter	**Wrapper**	**Filter**	**Wrapper**
Fast , Scalable and independent of classifiers	Interacts with classifier, Models feature dependencies	Ignores feature Dependencies, Ignores interaction with the classifier	Computationally intensive , Classifier dependent selection , Risk of overfitting

Table 4.2: Filter and Wrapper based feature selection comparison.

By taking the superior filter and wrapper methods which are DA and GA, new hybrid feature selection techniques is proposed and has been evaluated usingm

the same classification machine algorithms and validated using the same full dataset of financial ratios.

4.4 Results and discussion

In order to distinguish between results of studied feature selection techniques, accuracy and number of selected ratios have been used to differentiate and select the one with superior performance. The equivalent sensitivity, specificity, accuracy and number of financial ratios of feature selection techniques using Neural Network (NN) classifier were: 95.60%, 89.13%,92.52% and (9) respectively for the **DA**; 81.32%, 86.36%, 85.13% and (8) respectively for the **PCA**; 83.15%, 81.33%, 79.69% and (6) respectively for the **PSO**; 94.4%,90.07%, 94.87% and (7) respectively for **GA**; 97.64%,98.59%,99.33% and (8) respectively for **DAGA-FS** proposed hybrid feature selection technique. Table 4.3 clarifies the sensitivity, specificity and accuracy along with number of selected ratios of DA, PCA, PSO, GA and proposed DAGA-FS using feed-forward neural network (NN) classifier.

Classifier (NN)	Feature Selection	Number of Features	Sensitivity %	Specificity%	Accuracy%
Filter based	**DA**	9	95.60%	89.13%	92.52%
	PCA	8	81.32%	86.36%	85.13%
Wrapper based	PSO	6	83.15%	81.33%	79.69%
	GA	7	94.4%	90.07%	94.87%
Hybrid	**DAGA-FS**	8	97.64%	98.59%	99.33%

Table 4.3: Evaluation results of DA, PCA, PSO, GA and DAGA-FS using feed-forward neural network.

In addition to that, The corresponding sensitivity, specificity, accuracy and number of financial ratios of feature selection techniques using Polynomial support vector machine (SVM) classifier were: 94.65% , 87.79%, 90.82% and (9) respectively for the **DA**; 78.37%, 84.81%, 83.83% and (8) respectively for

the **PCA**; 78.66%, 80.83%, 81.88% and (6) respectively for the **PSO**; 92.98%, 88.37%, 93.57% and (7) respectively for **GA**; 90.12%, 89.12%, 91.78% and (8) respectively for **DAGA-FS** proposed hybrid feature selection technique. Table 4.4 clarifies the sensitivity, specificity and accuracy along with number of selected ratios of DA, PCA, PSO, GA and proposed DAGA-FS using polynomial support vector machine classifier.

Classifier (SVM)	**Feature Selection**	**Number of Features**	**Sensitivity %**	**Specificity%**	**Accuracy%**
Filter based	**DA**	9	94.65%	87.79%	90.82%
	PCA	8	78.37%	84.81%	83.83%
Wrapper based	PSO	6	78.66%	80.83%	81.88%
	GA	7	92.98%	88.37%	93.57%
Hybrid	**DAGA-FS**	8	90.12%	89.12%	91.78%

Table 4.4: Evaluation results of DA, PCA, PSO, GA and DAGA-FS using support vector (SVM) classifier.

Overall, feed-forward neural network exhibit better performance comparing to support vector machine. Wrapper based genetic algorithm feature selection achieved the best accuracy rate however it shows lower sensitivity comparing to filter based discriminant analysis which shows the best sensitivity performance but with two more redundant ratios comparing to genetic algorithm , PSO and PCA achieved the worst performance. While PCA shows high accuracy PSO has a fewer financial ratios with higher sensitivity.

On the other hand figure 4.2 demonstrates the proposed DAGA-FS achieves better accuracy with only 8 dimensions although DA has better sensitivity performance but still DAGA-FS select less redundant features with higher accuracy followed by DA and GA respectively.

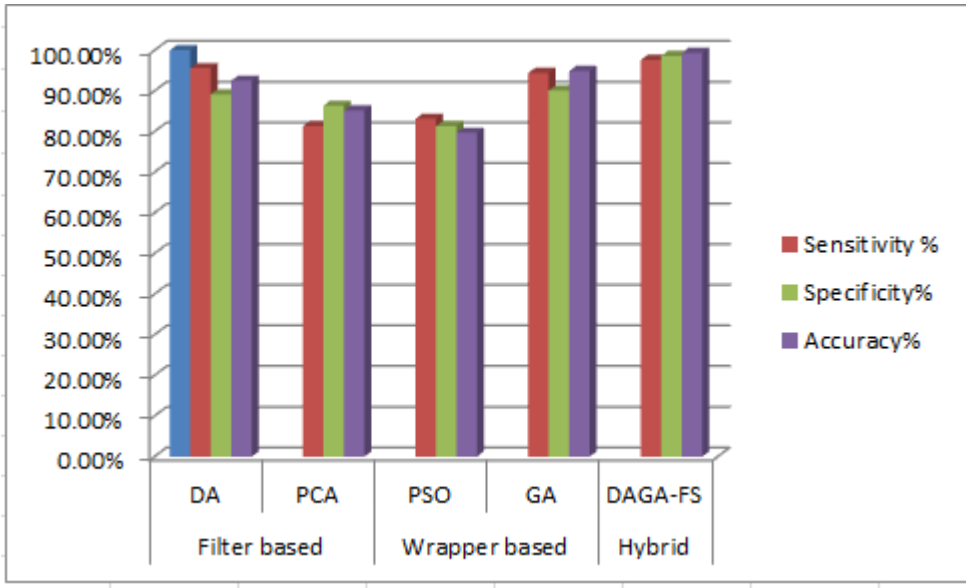

Figure 4.2: Comparison of DA, PCA, PSO, GA and DAGA-FS using the NN classifier.

Polynomial support vector machines explain that filter based feature selection techniques show higher sensitivity and accuracy average but in contrast wrapper based were able to maintain fewer ratios with less sensitivity and specificity as well as higher accuracy.

Figure 4.3 clarifies that DA shows best performance with 9 features but with slight difference comparing to genetic algorithm (7 features) and proposed hybrid method DAGA-FS (8 features).

4.4.1 Discussion

Two types of features selection have been studied and compared. Also two machine algorithms namely feed forward neural network and polynomial support vector machines were utilized to evaluate the performance of all studied methods. Wrapper based feature selection techniques shows its distinctive better performance comparing to filter based,

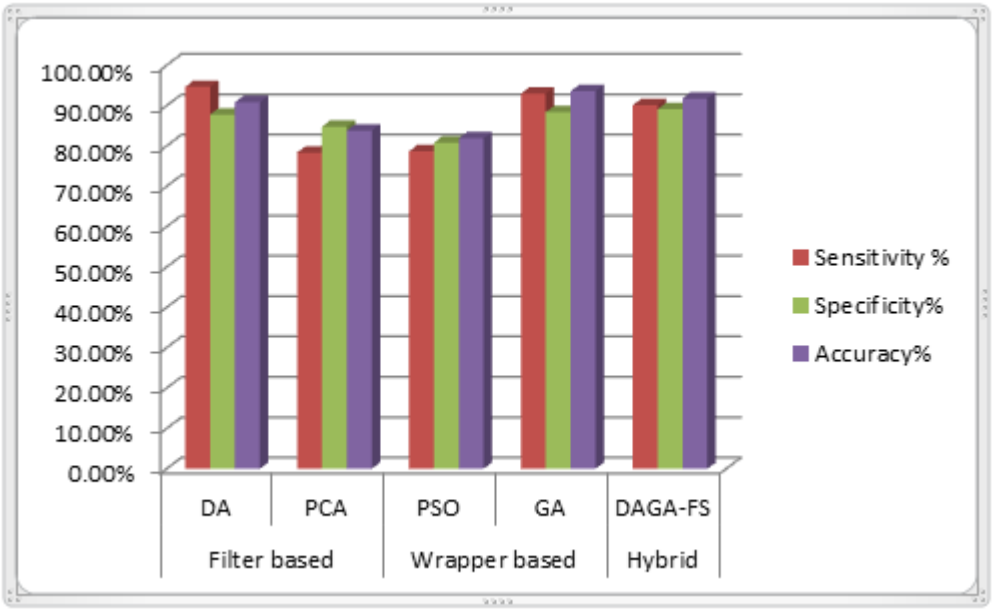

Figure 4.3: Comparison of DA, PCA, PSO, GA and DAGA-FS using the SVM classifier.

Two types of each feature selection generic scheme have were studied : DA and PCA for filter based, PSO and GA for wrapper based. The results are shown in Tables 4.3 and 4.5 and clarify that GA and DA outperformed others in term of accuracy, sensitivity and specificity. PSO shows the worst performance but it shows better ability to remove unnecessary information comparing to PCA.

The results showed that proposed DAGA-FS hybrid feature selection method can provide satisfactory and better performance comparing to the four wrapper and filter based methods using both NN and SVM classification. DAGA-FS shows that Sudan's bank distress prediction can be will represent using 8 financial ratios shown in the Table 4.5. DAGA-FS exluded 3 features which are EAS : Ratio of equity capital to total asset , LAS: Ratio of net loans to total assets and IDIVER: Finance related income to total income. This support the claims advantage of hybrid feature selection methods comparing to wrapper of filter based ones

Feature (Ratio)	Description
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
LLP	Ratio of loan loss provisions to total loans.
NPL	Ratio of non-performing loans to total loans
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.
LADF	Ratio of liquid assets to deposits and short term funds.
RFR	Rain Fall Ratio.

Table 4.5: DAGA-FS selected features.

4.5 Summary

In this chapter, wrapper based and filter based features selection methods have been evaluated on banks distress prediction, new hybrid DAGA-FS is proposed and evaluated using two classification techniques (NN and SVM). DAGA-FS outperformed other methods with NN and showed satisfying results with SVM behind DA method. A reduced set of eight financial features have been identified using the new proposed method which will be utilized in next classification tasks.

Chapter 5

Hybrid Classification of Evolutionary Algorithms and Bootstrap Aggregation Support Vector Machine for Bank Distress Classification

5.1 Introduction

This Chapter aims to build a classification technique to accurately classify the banks financial ratios by combining evolutionary algorithms namely (genetic algorithm and bootstrap aggregation) for better support vector machine using the result set of DAGA-FS feature selection method proposed in chapter 4. Hence the research target of this chapter is to design an EA-SVM method that can present high accuracy capabilities by eliminating overfitting for the banks distress prediction task. The Support Vector Machine method has a good learning and generalization ability but unfortunately it lacks the determination of the parameters for a given value of the regularization and kernel parameters and choice of kernel. In a way the SVM moves the problem of over-fitting from optimizing the parameters to model selection. Sadly kernel models can be quite sensitive to over-fitting the model selection criterion (Cawley, 2010).

In order to get the optimal parameters automatically, researchers have tried a variety of methods. Using evolutionary algorithms to optimize parameters of an SVM Classifier has become one of the favorite methods in recent years. In this chapter, GA is used to optimize the SVM classifier's parameters as well as bagging ensemble SVM classifiers.

The rest of the chapter continues as follows: Section 5.2 provides an overview of the proposed approach. Section 5.3 describes the parameters optimization and SVM methods. Experimental settings, results and discussions are contributed in Section 5.4. Benchmarking of the proposed model against

existing work is presented in Section 5.5. The chapter concludes with a summary in Section 5.6.

5.2 Proposed Approach

This section describes the general overview of suggested approach which starts by the features generated using DAGA-FS discussed in chapter 4; these features are then fed to the hybrid classification model. The hybrid method consists of evolutionary algorithms represented in genetic algorithms and support vector machine (GA -SVM), finally the bootstrap aggregation or bagging will be used to enhance the performance of single SVM classification model. Figure 5.1illustares the flowchart of proposed model. The proposed model tries to find an optimal set of feature weights and the classifier configuration using EA algorithms. Unlike statistical methods, the EA-SVM model needs no information about the weights of features; it receives the feedback of the SVM classifier to determine the searching directions. The experiments were conducted with the RBF kernel. However, due to flexible design of EA algorithms, the proposed model can adapt to other kernel parameters as well as classifiers such as kNN and naive Bayes.

5.3 Parameters Optimization Techniques and SVM

This section presents information about evolutionary algorithm techniques and the selected classifier support vector machine

5.3.1 Evolutionary algorithms for parameters optimization

Evolutionary algorithms form a class of heuristic search methods based on a particular algorithmic framework whose main components are the variation operators (mutation and recombination or crossover) and the selection operators (parent selection and survivor selection) (Eiben, 2003).

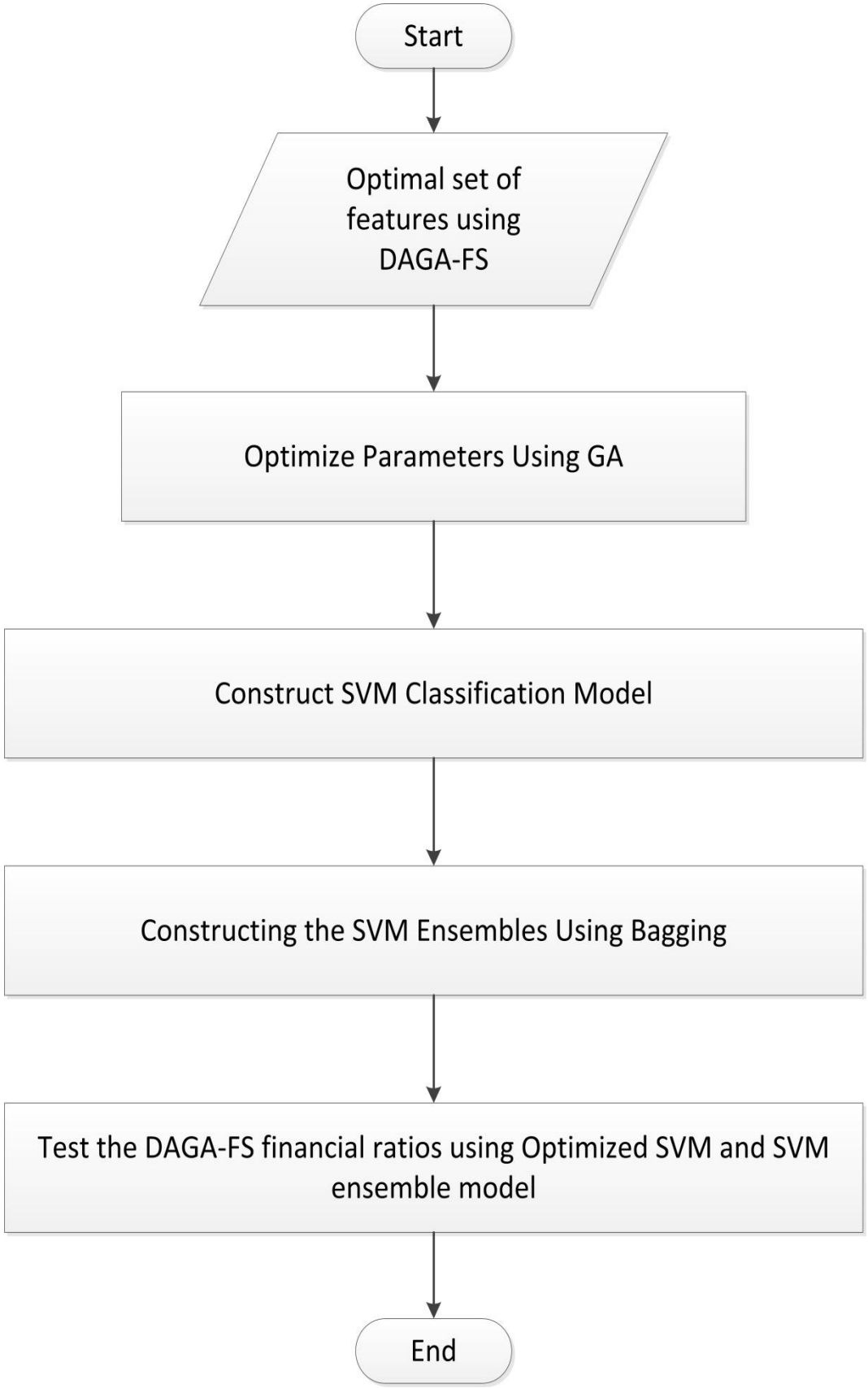

Figure (5.1)The flowchart of proposed model.

The general evolutionary algorithm framework is shown in Figure 5.2.

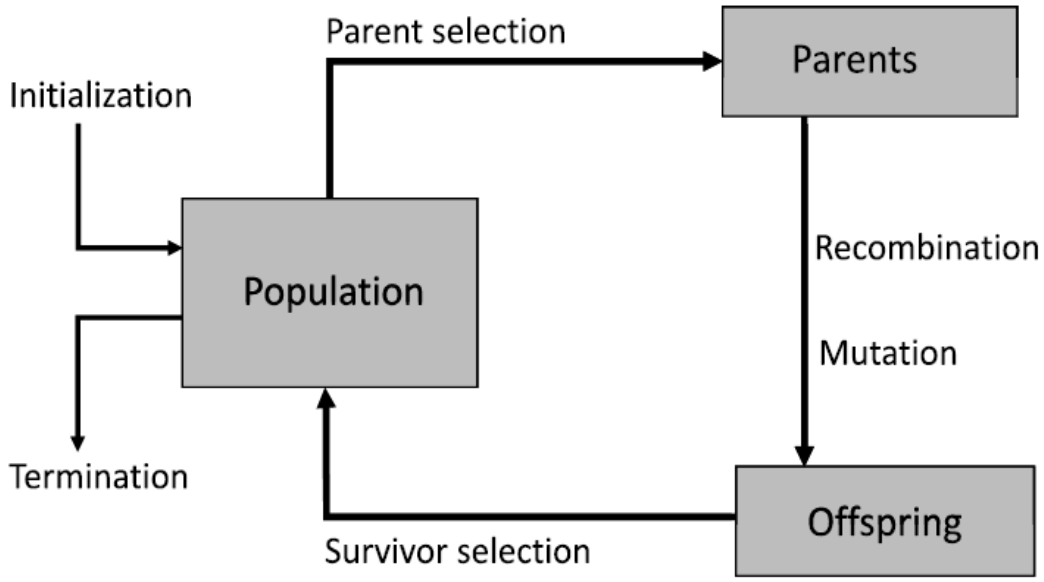

Figure 5.2: General framework of an evolutionary algorithm.

5.3.1.1 Genetic Algorithm (GA)

The genetic algorithm is a popular optimization method that attempts to incorporate ideas of natural evolution. Its procedure improves the search results by constantly trying various possible solutions with some kinds of genetic operations. In general, the process of GA proceeds as follows. First of all, GA generates a set of solutions randomly that is called an initial population. Each solution is called a chromosome and it is usually in the form of a binary string. After the generation of the initial population, a new population is formed that consists of the fittest chromosomes as well as offspring of these chromosomes based on the notion of survival of the fittest. The value of the fitness for each chromosome is calculated from a user-defined function. Typically, classification accuracy (performance) is used as a fitness function for classification problems. In general, offspring are generated by

applying genetic operators. Among various genetic operators, selection, crossover and mutation are the most fundamental and popular operators. The selection operator determines which chromosome will survive. In crossover, substrings from pairs of chromosomes are exchanged to form new pairs of chromosomes. In mutation, with a very small mutation rate, arbitrarily selected bits in a chromosome are inverted. These steps of evolution continue until the stopping conditions are satisfied

5.3.2 Support Vector Machines

Support vector machines (SVMs) are a set of 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.

5.3.2.1 Model development

As stated in literature review, 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 (Shin KS, 2005) Second, the SVM guarantees the existence of a unique, optimal, and global solution because SVM training is equivalent to solving a linearly constrained QP (Vapnik, 1998) . Third, the SVM implements the SRM principle that is known to have good generalization performance Burges (CJC, 1998), Finally, the SVM can be constructed with small training datasets to obtain prediction performance (Chen, 2011) 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 $T = \{x_i, d_i\}, i = 1,2, \ldots . n, \text{where } x_i \in R^m$ is the input information for the training object i on a set of m independent variables and $d_i \in \{0,1\}$ corresponding outcome (dependent variable). Formally, the aim of the analysis is the development of a function $f(x) \rightarrow d$ that distinguishes between the two classes of healthy and non-healthy banks. In the simplest case, f(x) is defined by the hyperplane $XW = \gamma$ as follows:

$$f(x) = \text{sgn}(xw - \gamma), \qquad (5.13)$$

Where w is a normal vector to the hyperplane and γ 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 $2/\|W\|$ (Vapnik, 1998) and (Fotios, 2008).

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:

$$\min \frac{1}{2} W^T W + Ce^T y, \qquad (5.14)$$

$$s.t.\ D(Xw - e\gamma) + y \geq e,$$

$$y \geq 0, w, \gamma \in R,$$

Where D is an n x n matrix such that $D_{ii} = d_i$ and $d_{ij} = 0 \; \forall \; i \neq j \; j$, X is an *n* x *m* 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, instead of solving the primal problem (5.14), it is more convenient to consider its dual Lagrangian formulation

$$\max e^T u - \frac{1}{2} u^T DX X^T Du , \qquad (5.15)$$

$$\text{s.t. } e^T Du = 0 ,$$

$$0 \leq u \leq Ce .$$

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

$$f(x) = sgn(x X^T Du - \gamma). \qquad (5.16)$$

(Burges, 1998) highlighted two reasons for using the Lagrangian formulation of the problem. The first is that the inequality constraints will be replaced by constraints on the Lagrange multipliers themselves, which will be easier to handle. The second is that in this reformulation of the problem, the training data will only appear (in the actual training and test algorithms) in the form of dot products between vectors. The latter is a crucial issue allowing generalizing of the procedure to the nonlinear case. Therefore, to generalize a linear SVMs model to a nonlinear one, the problem data are mapped to a higher dimensional space *H* (feature space) through a transformation of the form.

$$x_i\, x_j{}^T \rightarrow\ \theta(x_i)\, \theta^T\, (x_j) \qquad (5.17)$$

The mapping function θ is implicitly defined through a symmetric positive definite kernel function

$$K(x_i, x_j) = \theta(x_i)\, \theta^T\, (x_j)$$

Various kernel functions exist, such as the polynomial kernel, the radial basis function (RBF) kernel, the sigmoid kernel, etc. (Scholkopf,2002). 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 trade-off constant C. All the data used during model development were 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.

5.3.3 Ensemble Support Vector Machine

An ensemble of classifiers is a collection of several classifiers whose individual decisions are combined in some way to classify the test examples (Dietterich, 1998). It is known that an ensemble often shows much better performance than the individual classifiers that make it up.

The SVM has been known to show a good generalization performance and is easy to learn exact parameters for the global. Because of these advantages,

their ensemble may not be considered as a method for improving the classification performance greatly. However, since the practical SVM has been implemented using the approximated algorithms in order to reduce the computation complexity of time and space, a single SVM may not learn exact parameters for the global optimum. Sometimes, the support vectors obtained from the learning is not sufficient to classify all unknown test examples completely. So, there is no guarantee that a single SVM always provides the global optimal classification performance over all test examples. To overcome this limitation, this research to use an ensemble of support vector machines to accurately predict the Sudan's banking failure. Figure 5.3 shows a general architecture of the proposed SVM ensemble. During the training phase, each individual SVM is trained independently by its own replicated training data set via a bootstrap method explained in the following as well as aggregated SVM combination strategies.

5.3.3.1 Constructing the SVM Ensembles Using Bagging

In this research, bagging technique of (Breiman, 1996) will be adopted to construct the SVM ensemble. In bagging, several SVMs are trained independently via a bootstrap method and then they are aggregated via an appropriate combination technique. Usually, a single training set is available $TR = \{(x_i, y_i)|i = 1,2, \dots l\}$ But K training samples sets are needed to construct the SVM ensemble with K independent SVMs. From the statistical fact, the training sample sets should be different as much as possible in order to obtain higher improvement of the aggregation result. For doing this, we often use the bootstrap technique as follows.

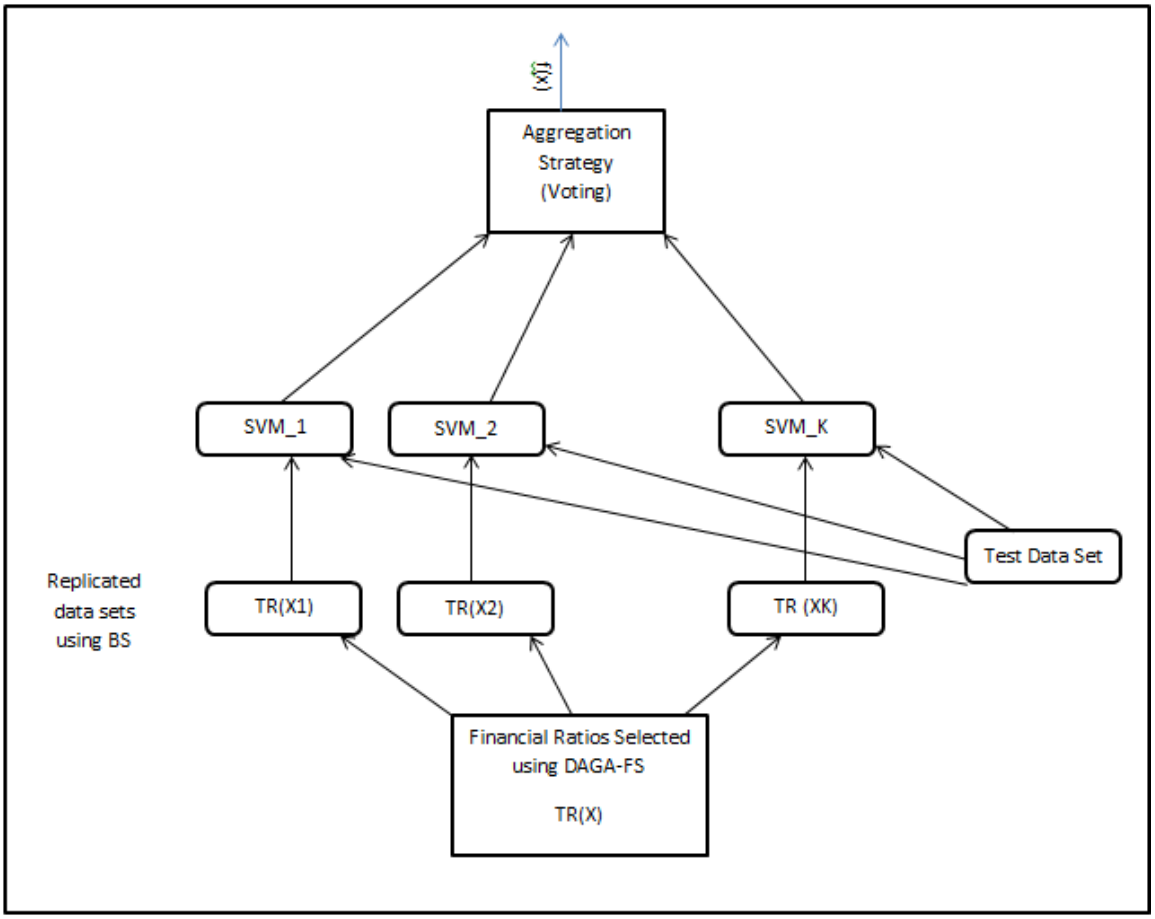

Figure 5.3: A general architecture of the SVM ensemble.

Bootstrapping builds K replicate training data sets $\{TR_K^B$ |k =1, 2... K} by randomly resampling, but with replacement, from the given training data set TR repeatedly. Each example x_i in the given training set TR may appear repeated many times or not at all in any particular replicate training data set. Each replicate training set will be used to train a certain SVM.

5.3.3.2 Aggregation Strategies for SVM Ensembles

After training, several independently trained SVMs will be aggregated in an appropriate combination manner. Linear combination method namely majority voting will be utilized.

Majority voting is the simplest method for combining several SVMs. Let $f_k(k = 1,2, \dots, k)$ be a decision function of kth SVM in SVM ensemble and $C_j(k = 1,2, \dots, j)$ refer to a label of j-th class. Then let $N_j = \#\{k | f_k(x) = C_j\}$,

i.e. the number of SVMs whose decisions are known to the jth class Then, the final decision of the SVM ensemble $f_{mv}(x)$ for a given test vector x due to the majority voting is determined by $f_{mv}(x) = argmax_i N_j$.

5.4. Experimental Settings

The experiments were conducted using the full dataset of financial ratios; but features that selected using DAGA-FS that was implemented in Chapter 5 form the main input vector. A total of 8 optimum features were selected. To carry out the computations, RM v.7.3 was used with the hardware configuration of Intel Core™ i5-5200, 2.2 GHz (4 CPUS) processor and 16 GB of RAM. The parameters optimization and SVM ensemble classification tasks was developed in RM.

The methodology is authenticated on the benchmark of the full financial ratios data set. In the evaluation, the entire dataset contains 506 records of financial ratios. Classification results are obtained using 5-fold cross-validation and 60% split (i.e. 60% training, 40% testing). To evaluate the performance of the classification, the standard statistical indices of sensitivity, specificity and accuracy were applied.

In the experiments, the GA-based parameter optimization best attained results used *Gaussian* mutation with tournament selection. On the other hand SVM classifier employed radial based kernel (SVM-RBF) and bagged to ratio .9 (10 samples of training data sets will be created).

5.5 Experimental results and discussions

The proposed optimized ensemble SVM classifier achieved the best accuracy (94.68 %) comparing to other classifiers as shown in Table 5.1. However, it is computational time is far higher than others. Neural Network (NN) classifier achieved the best computational time with just 4 seconds. Non-Optimized and Optimized parameters Single SVM have shown nearly the same

accuracy but in contrast the non-optimized single SVM reports less far computational time by less than one fifth.

Classifier	Accuracy %	Time/s
Non-Optimized Single SVM	91.78%	5
NN	89.33%	4
Optimized Single SVM	91.94%	33
Optimized Ensemble SVM	94.68%	60

Table 5.1: Classification results of the DAGA-FS features with NN, Non-Optimized Single, Optimized Single and Optimized Ensemble SVM Classifiers.

Figures 5.4 and 5.5 show the classification accuracy and computational time respectively, it is obvious that non-optimized single SVM shows the best combination of computational time and resulted accuracy.

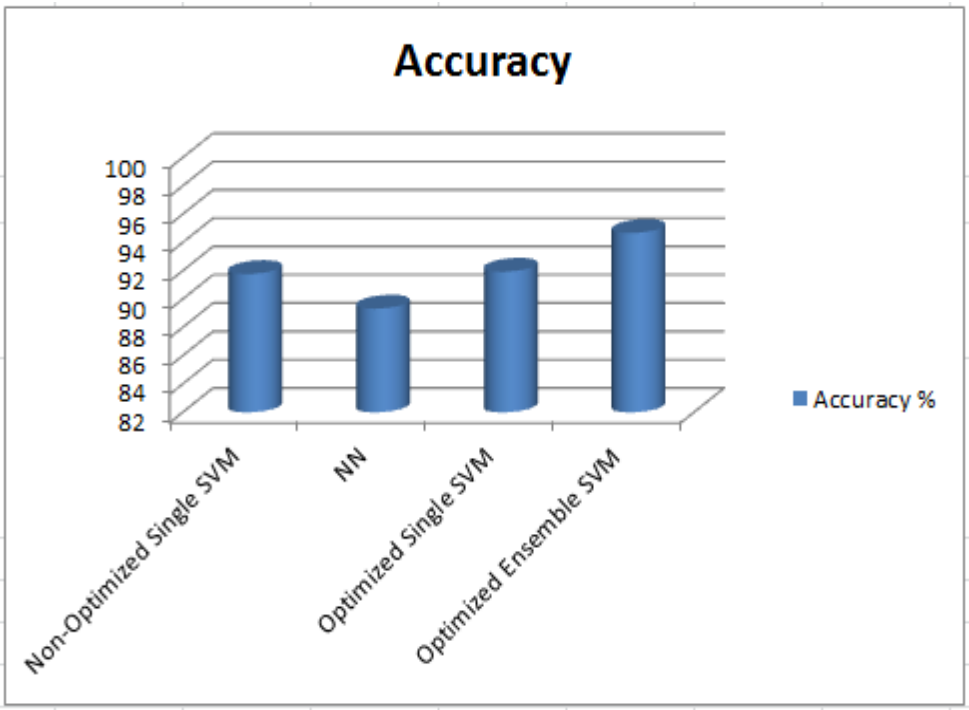

Figure 5.4: Classification Accuracy for DAGA-FS features with the non-optimized single SVM, NN, optimized single SVM and optimized ensemble SVM.

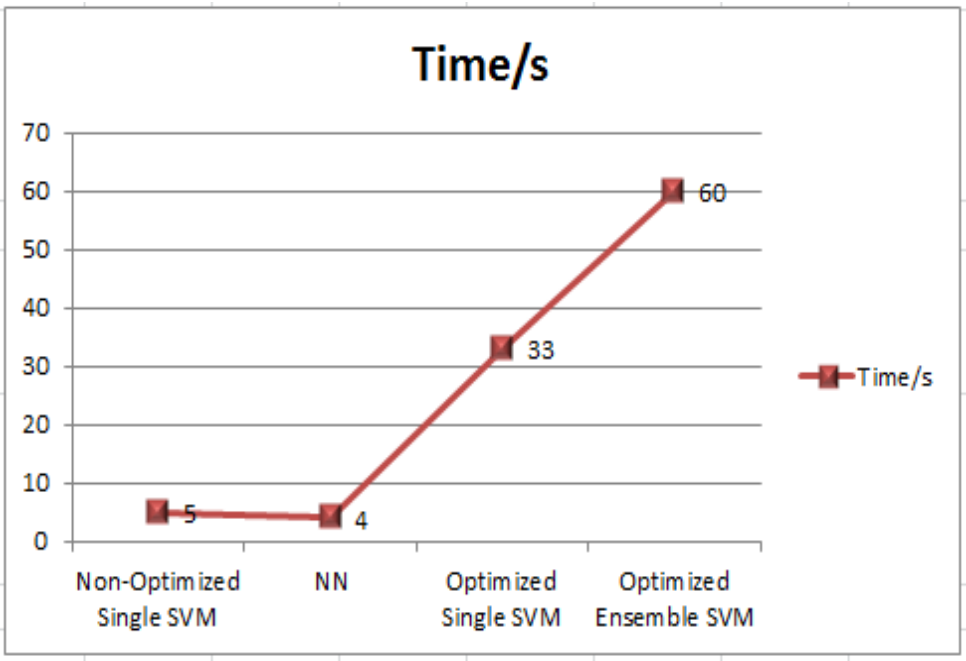

Figure 5.5: Computational time in seconds for DAGA-FS features with the non-optimized single SVM, NN, optimized single SVM and optimized ensemble SVM.

Bank distress classification method-based optimized ensemble SVM classifier uses the significant features generated by DAGA-FS proposed features selection method is benchmarked against the state-of-the-art. The details of the work and our proposed approach are shown in Table 5.2.

Author(s)	# Features	Classifier	Accuracy %	Domain
Myoung-Jong et al. (2010)	7	Ensemble NN	76.47%	Companies
Aykut et al. (2016)	35	Hybrid Ensemble RS-B-J483	83.7%	Banks
Yinhua et al. (2014)	12	Hybrid of TR4 and SVM	93.85%	Banks

[3] random subspace-bagging-j48 classifiers

[4] Trait Recognition Technique

Xiao-Feng Hui and Jie Sun (2006)	16	SVM	91.5%	Companies
Chang-Chih Chung et al. (2016)	7	CMNN[5]	97.50%	Companies
Proposed Methodology	8	Optimized Ensemble SVM	94.68%	Banks

Table 5.2: Comparison of the classification accuracy of the proposed method and some existing systems.

(Myoung-Jong, 2010) proposed an ensemble with neural network for improving the performance of traditional neural networks on bankruptcy prediction tasks. Experimental results on Korean firms indicated that the bagged and the boosted neural networks showed the improved performance over traditional neural networks with overall accuracy 76.47%

(Aykut, 2016) used three common machine learning models namely Logistic, J48 and Voted Perceptron as the base learners. In addition, an attribute-base ensemble learning method namely Random Subspaces and two instance-base ensemble learning methods namely Bagging and Multi-Boosting are employed to enhance the prediction accuracy of conventional machine learning models for bank failure prediction. The models are grouped in the following families of approaches: (i) conventional machine learning models, (ii) ensemble learning models and (iii) hybrid ensemble learning models. Experimental results indicate a clear outperformance of hybrid ensemble machine learning models over conventional base and ensemble models with average accuracy 83.7%. These results indicate that hybrid ensemble learning models can be used as a reliable predicting model for bank failures. Xiao-Feng Hui and Jie Sun (2006) applied support vector machine (SVM) to the early-warning of financial distress. Taking listed companies' three-year data before as sample data, adopting cross-validation and grid-search technique to find SVM model's good parameters, an empirical study is carried out. By comparing the experiment result of SVM with Fisher discriminant analysis, Logistic

[5] Cerebellar Model Neural Network

regression and back propagation neural networks (BP-NNs), it is concluded that financial distress early-warning model based on SVM obtains a better balance among fitting ability, generalization ability and model stability than the other models with achieved accuracy 91.5% . (Chang-Chih, 2016) proposed a novel prediction system which is based on intelligent classification to distinguish bankruptcy prediction. The method is referred to as a cerebellar model neural network (CMNN). A CMNN can be thought of as a learning mechanism imitating the cerebellum of a human being. Through training, this CMNN can be viewed like an expert of financial analyzer and then it can be applied to bankruptcy prediction. Their study uses an artificial neural network, a genetic programming, and the proposed CMNN to construct financial distress prediction models and compare the performance of above three models using some Taiwanese company data, and it confirms CMNN is better than the others with average accuracy 97.50%.

Our proposed method which is based on Ensemble SVM with evolutionary parameters optimization proves satisfactory accuracy rate (94.68%) comparing to others except CMNN model. It used 8 financial ratios selected by a proposed feature selection approach DAGA-FS and It can be reliably used as bank distress prediction system.

5.6 Summary

In this chapter, optimized parameters ensemble SVM system was built. Genetic evolutionary based algorithm was used to optimize SVM kernel parameters (C , γ). Then bootstrap aggregation technique used to build the ensemble SVM system. The proposed system will be fed by significant financial ratios generated by DAGA-FS feature selection methodology and its performance was measured and compared with other classification systems and state-of-art. It produced high classification results (94.68 %) which enable it to be safely used to classify and predict Sudan's bank financial distress.

Chapter 6

Islamic Credit Risk Analysis

6.1 Introduction

Islamic banking system has been expanding so quickly over the past few years. Moreover, it has been developing significantly around the non-Muslim territories including Middle Eastern countries, Southeast Asian countries, and European countries and even in North American countries. The existing of Islamic banks is to attract the customers who seek to avoid interest. Since interest is forbidden in Islam, Islamic banks have to avoid dealing with interest in any form. For that reason, Islamic banks came up with Profit-Loss Sharing System (PLS) and other sales contracts will be discussed throughout this research. The aim is to design a classification model which aid to identify the risk source of bank defaults in term of Islamic credit system.

6.2 Islamic Finance

During the last decade, news on Islamic finance has been dominated by the expansion and growth of the industry and the creation of Islamic finance institutions all around the globe. The industry turned 50 years old in 2012 – since the first experiments in interest-free financial institutions appeared around 1963, and the first commercial Islamic banks were created in 1975. Since the foundation of these first Islamic banks and institutions more than 35 years ago, Islamic finance has not only proved that it is possible to do finance without interest, but that it is also profitable and beneficial for economic development in the long run. With more than a trillion US$ in assets (Usmani, 1999), Islamic banking and finance represents today not only an interesting and unique development in the financial sector of Arab and Muslim countries but an opportunity for project finance and community development all around the globe. It is attracting the attention of financial markets both in Europe and in non-Muslim emerging countries where, in the aftermath of the crisis, it

represents an alternative form to conventional banking – more stable, less speculative and more related to real economic development

Conventional banks have been playing the role of financial intermediation for several decades. Interest differential was the original source of revenue in primitive models of banking which have been supported by charges, commissions and fees as new methods of income generation in modern banking system. Other than the role of intermediation, banks have been providing temporary cushions for short-term money requirements in the market as well as assisting in performing the function of price discovery in the money market.

Conventional banking is largely based on interest rates, accounting, various products and services, risk management activities, as well as long-term strategies, which are based on interest rates. Islamic scholars have raised questions about the necessity and validity of the interest in the process of financial intermediation. In the desire to provide sustainable and justified distribution of wealth and income, Islamic finance has attempted to find alternates to the conventional form of financing. Interest has been considered as a form of exploitation since it is merely a charge on money. Hence, the prohibition of giving and taking interest among the Muslim population can be considered as a prime reason for the origin of Islamic banking.

Interest, as called Riba (literally meaning 'extra'), is considered to be Haram (literally meaning 'forbidden') and hence is prohibited. Added to this was the disenchantment of Muslims for investing in economic activities involving Gharar (literally meaning 'risk', 'uncertainty', and 'hazard'). Due to the strong emphasis in Islamic economics on equitable distribution of wealth in the society, there has been insistence on sharing the revenue with the less fortunate in the form of Zakat (literally meaning 'purification').

Hence, the Islamic banking is mainly based on the absence of Riba in the transactions, avoidance of Gharar in contractual terms, payment of Zakat for the needy and poor, and avoidance of Haram activities. A common thread running across all these tenets is protection of the poor and weak from exploitation by the rich and powerful. Islamic finance has a strong root in sustainable society with focus on welfare, equality, and justice. Social implications of commercial activities cannot be neglected in Islamic finance since it has a strong emphasis on a socially responsible form of financing. The activities of Islamic finance are not purely materialist, although profit is a motive, but it is supported by strong social responsibilities and accountabilities.

The social objectives cannot be separated from commercial objectives in Islamic finance. Some of these differences between conventional financing and Islamic financing are summarized in Table 6.1

Islamic finance	Conventional finance
Interest is prohibited	Primarily based on Interest rate
Unstructured and still informal in many ways	Structured and formalized
Stress on social, ethical and financial efficiency	Stress on financial efficiency
Standards for risk management, accounting and other activities are still are still developing	Highly systematized in terms of risk management, accounting and other and other standards
Non-existence of short-term money market	existence of short-term money market

Table 6.1: Differences between conventional and Islamic financing.

6.2.1 Islamic Finance in Sudan

The banking system in Sudan has passed through six stages. The first stage, from 1903 to 1956, during the British colonial rule, was characterized by the domination of foreign banks branches in Sudan. The second stage from 1956 to 1976, following the independence of the country, witnessed the establishment of the Central Bank of Sudan (CBOS) and other national banks, which operated, hand in hand, with the then existing branches of foreign banks until their nationalization and amalgamation into national banks between 1970 and 1975. The third stage, from 1976 to 1989, was marked by the declaration of Shari'ah law in Sudan, Islamisation of financial legislations, and establishment of many Islamic banks. The fourth stage, from 1989 to 2002, witnessed the strengthening of Islamisation of financial institutions and legislation. The fifth stage, from 2002 to 2011, following the Comprehensive Peace Agreement (CPA), signed in 2002 between the Government of Sudan and the Sudan People Liberation Movement (SPLM) of South Sudan, has been embodied in the Transitional Constitution of The Republic of Sudan, and the financial system witnessed the establishment of two banking systems in Sudan. An Islamic banking system existed in the North of Sudan, whilst it was agreed in the Nevasha agreement that a conventional banking system would be implemented in the South of Sudan. The Central Bank of South Sudan (CBSS) was established as a branch of CBOS to look after the conventional banking system, while CBOS carries on its responsibilities as supervisor of the Islamic banking system operating in the North of Sudan. The sixth stage was the return to a full Shari'ah - compliant financial system following the declaration of independence of South Sudan. The emergence of Islamic banks has helped in attracting considerable funds to the banking system. Customers had previously shied away from conventional banking services. Recently, some Sudanese banks have started expanding into other markets in Africa. For instance, the

Islamic Bank of Khartoum is trying to expand its customer base to East Africa in the medium-term, beginning with Kenya. The Islamic Bank of Khartoum was privatized in 2002, and is now 60 % owned by Dubai Islamic Bank. In 1992, the state established the High Shari'ah Supervisory Board to oversee the progress of the reforms and their compliance with Shari'ah . The body comprises scholars, jurists, and economists and is subject to the terms of the law regulating the banking activities. The insurance sector is also based on Shari'ah principles and is stipulated in state legislation. The insurance sector was given a range of incentives, including tax exemption on all of its assets and profits, and the firm's assets cannot be confiscated or nationalized.

In 1994, the Khartoum Stock Exchange (KSE) was set up. The exchange trades shares of 58 Sudanese companies as of May 10, 2016, some investment funds, and a number of government ṣukūk . KSE requires full information disclosure, which ensures a high level of transparency. The stock exchange has its own Shari'ah board, which screens and approves the products prior to their trading. IMF played a significant role in supporting Sudan's endeavors. Amongst other things, the IMF specialists helped to devise government bonds, based on the mechanism of mushārakah. In 2003, KSE launched the Khartoum Index, developed with the assistance of the IMF. In 5 years, it grew from 1000 to 2500 points. Today, KSE is one of the top five African stock exchanges—it ranks fifth, with a trade volume of trading US$5 billion (not including ṣukūk trading). Government mushārakah certificates (GMCs), also known as Shahama, are short-term securities. Through Shahama, the state raises money in the domestic market instead of printing more banknotes. After 1 year, holders of GMCs can either cash or extend them. These certificates are backed by the stocks and shares portfolio of various companies owned by the Ministry of Finance, and are therefore, asset-backed. The profitability of GMCs can

reach 33 % per annum and depends on the financial results of the companies involved.

Hence, the profit of a GMC varies and is not fixed. The government issues these certificates on a quarterly basis. Government investment certificates (GICs) are medium-term securities based on various contracts financed by the Ministry of Finance of Sudan via the istisna , murabaḥa, and ijarah tools. The Ministry of Finance acts as the originator of the certificates. GICs are based on restricted mudarabah , which means that the raised money is invested solely in the projects stipulated in the original contract. Ijarah certificates of the Bank of Sudan (CICs) are backed by the buildings owned by the central bank.

According to the law, Sudanese banks must invest up to 30 % of deposits in CICs. These bonds use ijārah as the method of financing. At the end of each term, an independent surveyor evaluates the buildings. Sudan's latest issue of ṣukūk was fully subscribed and the country was able to raise the equivalent of US$160 million. More such issues are planned for the coming years. The ṣukūk issues are designed to help make up for the loss of oil revenue. (HSSB, 2015)

6.3 Islamic Finance Modes

Islamic banks use a number of non-interest-based financing modes. The use of a particular mode is dependent on the nature, purpose and size of transactions. In selecting the mode, it is very much the know-how and knowledge of the Islamic banker which comes into play. These modes could be classified as debt type instruments, quasi - debt type Instruments, profit and loss sharing instruments or hybrid instruments.

Debt type instruments include Murabaha, Salam, Istisna, Tawarruq and Qard Hasan. On the other hand, Ijarah is quasi-debt instruments; it is one of the simplest asset-based financial instruments. Profit-And-Loss-

Sharing Instruments include Musharakah, Mudarabah and other hybrid modes. The Following explains each one in detail

6.3.1 Murabaha

It is the most frequently exercised mode of Islamic financing which is practically implemented in financial institutions and in other financial transactions. It is defined as: "Murabaha is particular kind of sale where the seller expressly mentions the cost of the sold commodity he has incurred, and sells it to another person by adding some profit" (Haskafi et Muhammad Bin Ali) For financial transactions by using Murabaha, it is very important that all conditions of sale defined by the Islamic jurists should be fulfilled. For example:

I) before sale, the commodity of sale has to be in the possession of vendor.

II) If the sale is attributed to a future date or event, it will be regarded as void and if parties want to affect sale, a fresh sale contract is required.

III) Price should be certain for the validity of the sale

IV) The delivery of the commodities must be certain etc.

6.3.2 Salam

Salam is also known as 'forward sale'. Salam was originally allowed to meet the needs of small farmers who needed money during the harvesting period to meet expenses for harvesting as well as to maintain their family. Since borrowing on interest was not permitted, they were allowed to use forward sales. It was also used originally for the import and export trade in the Arab region. As discussed in the previous sections, according to Islamic jurisprudence the essential conditions of a valid sale include that the commodity should be ready at the time of the sale. However, Salam and Istisna (which is explained later), are the two exceptions which are permitted according to Islamic fiqh (legislation religion authority). Hence, Salam is a contract of forward sale, whereby a commodity is sold on a future date, for

which the complete price is paid on the spot. In this case, the price is paid in cash and the delivery is deferred. The buyer is called the *rabb-us-salam*, and the seller is called the *muslam ilaih*. Salam is permitted under certain conditions:

1) The price should be paid in full and on the spot.

2) It can be used for specific products and it can be used for commodities which can be clearly explained in quantity and quality.

3) The time and place of delivering the commodity should be specific.

4) Salam cannot be used for barter transactions.

There can be several applications of Salam, depending on the circumstances. It should be noted that Salam exposes banks to market risk, especially fluctuations in the commodity prices and instability of irrigation sources such like rain fall ratios.

Banks are not very happy to take the delivery of the commodity which is a necessary condition in the Salam. To avoid this, modern bankers are using the Parallel Salam, where a bank enters into two simultaneous agreements for the same future date, one as a buyer and the other as a seller. Parallel Salam takes care of commodity price fluctuations to a certain extent, but still requires managing the risks from the non-delivery of the commodity on the due date.

6.3.3. Istisna'a

Istisna is a mode of finance it is defined as: "Kind of sale where a commodity is transacted before it comes into existence". We can say in current era of global business a party orders to manufacture a product and for this some time he/she have to pay advance payment.

The important point in the case of Istisna that the manufacturer uses its own material for production otherwise the contract will be of Ijara rather than Istisna. Also it is important to fix the price with the approval of concerned

people and specification of product should also be settled. There are some differences between Istisna and Salam:

1) In Istisna manufacturing of commodity is necessary

2) In salam full price is paid in advance but in Istisna there is no such condition

3) Delivery time is important time of Salam and not of Istisn

6.3.4 Tawarruq

Tawarruq is a financial instrument in which a buyer purchases a commodity from a seller on a deferred payment basis, and the buyer sells the same commodity to a third party on a spot payment basis (meaning that payment is made on the spot). The buyer basically borrows the cash needed to make the initial purchase.

Later, when he secures the cash from the second transaction, the buyer pays the original seller the installment or lump sum payment he owes (which is cost plus markup, or murabaha).

Because the buyer has a contract for a murabaha transaction, and later the same transaction is reversed, this scenario is called a reverse murabaha. Both transactions involved must be sharia-compliant.

Tawarruq is a somewhat controversial product. Because the intention of the commodity purchases isn't for the buyer's use or ownership, certain scholars believe that the transactions aren't sharia-compliant. Their argument is that the absence of any real economic activities creates interest, which is prohibited in sharia.

6.3.5 Qard Hasan

Qard means a loan given for something good or to help someone in the name of Allah. Hasan or Hassan means good or acceptable; of good faith.

Qard Hassan is a contract involving a loan with two parties on the basis of social welfare. It also can fulfill a short-term need of the borrower. In modern day terms, many compare this to a payday loan. During the loan process, the repayment amount must be the same as the amount borrowed. This means no interest or riba must be applied to the loan.

However, in terms of good faith, the borrower may pay the lessor more money in the future; yet it cannot be discussed or agreed upon during the contract. This means that if they give the lessor a bonus or extra payment it is permissible, but the discussion of such an arrangement is prohibited. This is often done in a measure of good faith and as a means to thank the lessor.

Ultimately, under Islamic finance laws of Shariah, each loan should always be free of profit. In terms of Qard Hassan, qard often means a bank deposit, which in North American banking means a loan to the bank to use it until the depositor asks for its return. In general, this is considered an interest-free loan. The only permissible loan involving interest is Qardhul Hasam.

Also, a contract such as this must be extended upon in goodwill. The debtor only is required to repay the amount that was initially borrowed with no riba or interest attached to the agreement. This type of arrangement typically doesn't occur with a Muslim mortgage, however in some good faith arrangements, they can.

6.3.6 Ijarah

Ijarah is also commonly known as leasing. There are two connotations in Islamic jurisprudence for Ijarah: one is for hiring services and another is for assets and properties. In the first case, the service can be hired and the reward can be paid for labor. In the second, Ijarah refers to the sale of a definite

usufruct in return or a definite reward. Since the first type of Ijarah is not used in Islamic banking, the discussion will be restricted to the second connotation. For a bank, the second connotation is more important and applicable. Banks can use Ijarah for leasing equipment, machinery, vehicles, and other assets. The lessor in this case is called 'mu'jir', and the lessee is called 'musta'jir'. Ijarah is subject to certain rules as provided by Islamic fiqh:

1. The subject of the lease must have a useful value.
2. The ownership of the asset should remain with the lessor and only the right to use the asset should be transferred to the lessee.
3. The liabilities pertaining to the ownership remain with the lessor; whereas the liabilities pertaining to the use are with the lessee.
4. The period of lease must be pre-defined.
5. A jointly-owned property can be leased and the lease rental must be shared in the proportion of their share in the property.
6. The consideration for the lease (rental) should be determined in the beginning.
7. The lease period will begin subject to delivery of the leased asset and not the date of commencement of use of the asset by the lessee.

6.3.7 Musharakah

Musharakah is also known as 'Partnership Financing' or 'Joint Venture Financing'. Islamic Fiqh does not refer to the word Musharakah; it is a derivation from the word shirkah, which means sharing. Unequal distribution of wealth is treated as a sin in Islam and hence any partnership with unjust distribution of profits or losses is not permitted. It is believed that the interest is the root cause of unequal distribution of wealth and hence moderation is used while distributing the profits and losses of the business, where, in the case of profit, the managing partner should not take a large share and, in the case of

loss, the non-working partners should not be asked to bear all the burden of loss, which generally is the case in conventional partnerships.

To overcome these issues of conventional partnership, Musharakah is used as an alternative. Musharakah is a form of equity financing which refers to a partnership agreement between the bank and the customer where equity is contributed jointly and profits and losses are shared on agreed terms; however, it is not just lending money. The capital can be contributed in cash or in the form of goods or assets. The profit-sharing ratio can be decided at the time of the agreement but, in the absence of a loss-sharing ratio, the losses will compulsorily be shared as per the proportion of the capital. Both the parties have the right to manage, although one of them may surrender their right in favor of the other. Musharakah is seldom used due to the high degree of uncertainty over the returns. It is used in cases involving huge investments and for joint-venture projects. In Musharakah, the finance comes from both the parties. In case only one party finances the whole project; Mudarabah is used, which is explained in the next section. Musharakah can also take the shape of Diminishing Musharakah, where the entrepreneur keeps purchasing the share of the financier regularly and thus diminishes the contribution of the financier, which eventually is completely phased out. A possible use of Diminishing Musharakah can be for real estate financing. However, Diminishing Musharakah cannot be used easily for financing regular trade transactions.

The general law of partnership applies for the termination of Musharakah. The partners can terminate the Musharakah by giving a notice. Death or irrationality of one of the partners also leads to the termination of Musharakah.

6.3.8 Mudarabah

It is the most frequently exercised mode of Islamic financing which is practically implemented in financial institutions and in other financial transactions. It is defined as: “Mudarabah is particular kind of sale where the

seller expressly mentions the cost of the sold commodity he has incurred, and sells it to another person by adding some profit".

For financial transactions by using Mudarabah, it is very important that all conditions of sale defined by the Islamic jurists should be fulfilled. For example:

I) before sale, the commodity of sale has to be in the possession of vendor.

II) If the sale is attributed to a future date or event, it will be regarded as void and if parties want to affect sale, a fresh sale contract is required.

III) Price should be certain for the validity of the sale.

IV) The delivery of the commodities must be certain.

Musharakah and Mudarabah have some common features. Both should not be confused with simple financing of business. They require participation in business either in the form of contribution to the capital or management or both and prohibit one party benefiting at the cost of the other in sharing profits and losses. The partners are free to determine the ratio of profit and loss sharing, subject to the condition that, in the event of losses, they should be shared as per their capital contributions and profits for the financing partner (who is not taking part in management) cannot exceed his share in capital. Both can be securitized, particularly where the investments are vast. The investment can be divided equally in parts and Musharakah /Mudarabah certificates can be issued to each contributor like debenture certificates.

Under certain circumstances these certificates can be traded on the secondary market and hence can provide the much required liquidity for Islamic banks.

6.3.9 Deferred Credit

A deferred credit could mean money received in advance of it being earned, such as deferred revenue, unearned revenue, or customer advances. A deferred credit could also result from complicated transactions where a credit amount arises, but the amount is not revenue.

A deferred credit is reported as a liability on the balance sheet. Depending on the specifics, the deferred credit might be a current liability or a noncurrent liability. In the past, it was common to see a noncurrent liability section with the heading *Deferred Credits*.

6.3.10 installment

A payment made as part of a series of payments on the same good, service, or obligation. For example, if one buys expensive consumer goods (such as furniture), one may agree with the seller to pay in installments until the furniture is paid in full. Likewise, one also makes installment payments on loans. Installments may or may not require profit to be paid.

6.3.11 Letters of Guarantee

A letter of guarantee is a type of contract issued by a bank on behalf of a customer who has entered a contract to purchase goods from a supplier and promises to meet any financial obligations to the supplier in the event of default. A letter of guarantee may also be issued by a bank on behalf of a call writer guaranteeing that the writer owns the underlying asset and that the bank will deliver the underlying securities should the call be exercised. Call writers will often use a letter of guarantee when the underlying asset of a call option is not held in their brokerage account.

6.4 Islamic Credit Risk

The risks faced by banks in their operations are of many types: interest risk, exchange rate risk, trade risk (or market risk), political risks, and risks that represent changes in the value of tangible assets and goods, etc. Credit risk is deemed to be the most important type of risk faced by a bank in its relationship with the owners of wealth. It is related to the ability of a debtor to repay at the time appointed for repayment and in accordance with the conditions stipulated in the contract. If the debtor fails to abide by his obligations, it leads to a loss for the creditor and, therefore, becomes a risk for the bank. The existence of

credit risk is not dependent on direct financing by the bank, like bank loans. The bank also faces this type of risk in guarantees and acceptance paper when the originator of the financial instruments owned by the bank is unable to meet his obligations (as in the case of bonds). So is the case in other indirect financing operations. Therefore, prudent bank management includes strict and detailed regulations specifically for credit risk with the purpose of managing it in a suitable manner.

Conventional banks face credit risk in almost all of their operations, because the relationship between the banks and those who transact with them is that of a debtor with a creditor in all cases. Islamic banks also face this form of risk in most of the modes of financing that they use. It is well known that murabaha, Istisna, and installment sale are sales with delayed payment thus generating debts in the accounts of the banks. The fundamental form of risk in all these contracts is credit risk. Salam gives rise to a commodity debt rather than a cash debt, but it also involves credit risk. Mudarabah and Musharakah, on the other hand, are contracts of participation, and the funds given by the bank to entrepreneurs are not liabilities.

Nevertheless, these two also bear a credit risk in two ways. First, in the case of wrongful act or negligence, the entrepreneur is liable to guarantee the capital which means a debt liability. Second, when the capital of Mudarabah or Musharakah are employed in a deferred sale, which is what takes place in most Mudarabas, the owner of capital (rabb al-mal), the bank in this case, bears an indirect credit risk. This risk pertains to the ability of the counter parties to repay.

6.5 Credit Risk Analysis Model

This section highlights the difference between ordinary credit risk models and our proposed one. In order to study the risks of Islamic finance only three

features will be studied (finance modes, Finance Sector and Payment Type). The classification model tries to investigate which of these features is linked with the insolvency occurred to customers.

The Proposed methodology is explained in figure 6.1. The credit analysis data set namely (Finance Mode (FM), Finance Sector (Sec) and Payment Type (PT)) will be fed to feature selection techniques. Two features selection techniques will be studied to select the best by applying two classifications techniques which are proved its efficiency on building credit risk classification models Grier (2007). Then ordered list of significant factors can be identified which is one highly important in determining the possibility of credit risk in Sudan's banking sector.

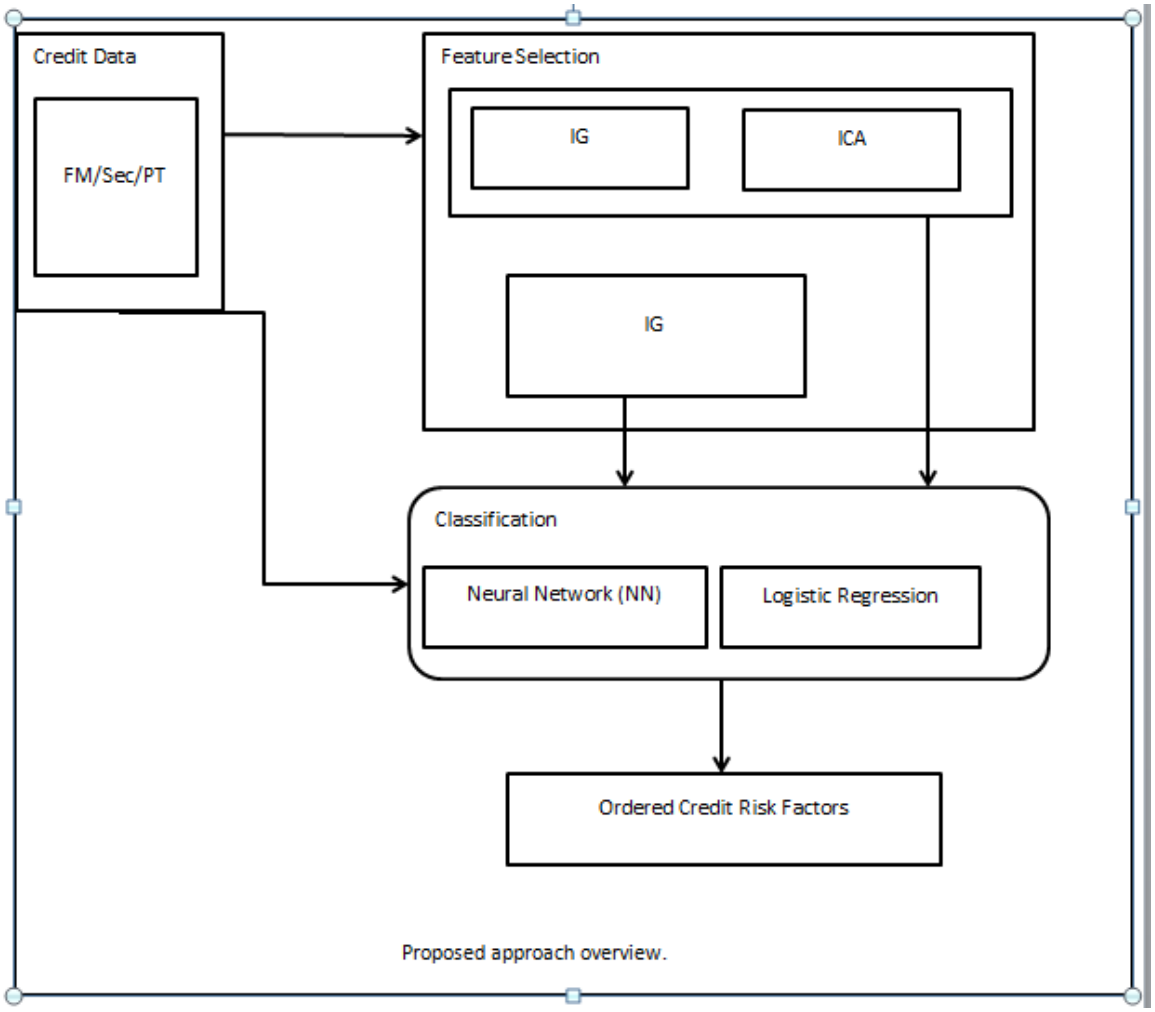

Figure 6.1: Proposed approach overview.

6.5.1 Credit Data Set

Data has taken from DAS data source as explained in data warehouse designing task, banks regularly send their finance information containing the dimensions required by central bank to study weather banks comply with finance policies and circulars or not. Those dimensions are (finance mode, sector, payment type, bank) along with customers and non-performing loan amount accumulated for each bank. Our sample consists of 15,218 bad cases and 1,078 good ones. This implies the bad rate of the whole sample equals almost to 70 %. In the following paragraphs, the variables that are going to be used in the analysis to describe the dependent variable will be introduced:

6.5.1.1 Payment Method

This variables indicates weather the payment method in case of financing the imports trade, the possible options for this variable include: nil value, which indicates imports operation will take place without actual transfer. Advance Payment, this in case of bank has no trust in his customer demanding certain percent of finance operation to be paid in advance. Deferred Payment, as the name suggests payments will scheduled at certain installments during defined period of time.

6.5.1.2 Mode of Finance

This variable define the Islamic finance mode used to permit the loan for such client, those modes have discussed in the previous sections which include following values but not limited to : Musharakah , Mudarabah , Salam.

6.5.1.3 Sector

Refers to those sectors of the economy which are targeted by finance operations .Central bank of Sudan specify the priority Sector Lending to the banks for providing a specified portion of the bank lending to few specific sectors like agriculture and allied activities, micro and small enterprises, poor

people for housing, students for education and other low income groups and weaker sections.. This is essentially meant for an all-round development of the economy as opposed to focusing only on the financial sector.

The above variables has never been used before in any published study as best for our knowledge , so this work to evaluate the credit risk as a macro perspective which is different from the bank's wise perspective where the variable such like age, level of education are being used consider is first effort for its type at least in Sudan. In addition to that, the information used in this analysis is classified confidential and without the approval of monetary authority in the republic of Sudan the results couldn't be available.

6.5.2 Features Selection

Feature Selection involves identify the significant features in order to achieve optimal classification results. This section explains the dimension reduction based on credit risk data set using: independent component analysis (ICA as well as features selection technique IG

6.5.2.1 Independent Component Analysis (ICA)

ICA is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed nongaussian and mutually independent and they are called the independent components of the observed data. These independent components, also called sources or factors, can be found by ICA.

ICA is superficially related to principal component analysis and factor analysis. ICA is a much more powerful technique, however, capable of finding

the underlying factors or sources when these classic methods fail completely. (Aapo Hyvarinen,2001)

6.1.1.1 Information Gain (IG)

Information gain (IG) measures how much "information" a feature gives us about the class, it is applied on the data to obtain the reduced D features using the concept of ranking the features according to the information provided. IG tries to measure the information obtained when making a decision based on a given feature (Salzberg, 1994). Computing information gain for a feature involves computing the entropy of the class label for the entire dataset and subtracting the conditional entropies for each possible value of that feature. The entropy calculation requires a frequency count of the class label by feature value. In more detail, all instances are selected with some feature value e ; then the numbers of occurrences of each class within those instances are counted, and the entropy for e is computed. This step is then repeated for each possible value e of the feature. The entropy of a subset can actually be computed more easily by constructing a count matrix, which tallies the class membership of the training examples by feature value.

6.5.3 Classification

In order to evaluate the performance of each feature selection technique, two classifiers have used namely: logistic regression and neural network.

6.5.3.1 Logistic Regression

The crucial limitation of linear regression is that it cannot deal with dependent variable's that are dichotomous and categorical. Many interesting variables are dichotomous: for example, consumers make a decision to buy or not buy, a product may pass or fail quality control, there are good or poor credit risks, an employee may be promoted or not.

```
1: Function  IG (C|E) feature ranking based entropy
2: Initialization :
3:   S=0;
4:   C← domain of a class label ;
5:   E← domain of an attribute values ;
6:  For each c_i ∈ C do:
7:      Calculate p(c[i]);
8:      H_c=S+ p(c[i])*log2(p(c[i]));
9:       S← H_c;
10: End For
11: For each e_i ∈ E:
12:      Calculate P(e[j]);
13:      Sum=S+ P(e[j])*log2(p(e[j]));
14:      S← Sum;
15:  End For
16: For each c_i do :
17:   For each e_j do :
18:      Calculate p(c[i] |e[j]);
19:      M=S+ p(c[i] | e[j]) * log2 p(c[i] |e[j]);
20:      S← M;
21:   End For
22:  End For
23: H(C|E) = (-1)*Sum*(-1)*M ;
24: IG= H_c- H(C|E)
25: return IG
26: End function
```

Figure 6.2: Information Gain (IG) algorithm (Alhaj, 2016).

A range of regression techniques have been developed for analyzing data with categorical dependent variables, including logistic regression and discriminant analysis.

Since the dependent variable is dichotomous we cannot predict a numerical value for it using logistic regression, so the usual regression least squares deviations criteria for best fit approach of minimizing error around the line of

best fit is inappropriate. Instead, logistic regression employs binomial probability theory in which there are only two values to predict: that probability (p) is 1 rather than 0, i.e. the event/bank belongs to one group rather than the other. Logistic regression forms a best fitting equation or function using the maximum likelihood method, which maximizes the probability of classifying the observed data into the appropriate category (defaulted or not) given the regression coefficients.

Like ordinary regression, logistic regression provides a coefficient 'b', which measures each independent variable's partial contribution to variations in the dependent variable. The goal is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model (i.e. an equation) is created that includes all predictor variables that are useful in predicting the response variable. Variables can, if necessary, be entered into the model in the order specified by the researcher in a stepwise fashion like regression.

We do not estimate the probability of 0 or 1 but the probability of the chance of "success" (or event). If P is the probability of the event, the (1 - P) is the probability that the event will not happen. The chance of the event is then P / (1 - P). The chance of event is called the Odds ratio.

The outcome of the regression is not a prediction of a Y value, as in linear regression, but a probability of belonging to one of two conditions of Y, which can take on any value between 0 and 1 rather than just 0 and 1.

Unfortunately a further mathematical transformation – a log transformation – is needed to normalize the distribution. This log transformation of the p values to a log distribution enables us to create a link with the normal regression equation. The log distribution (or logistic transformation of p) is also called the logit of p or logit (p)

Logit (p) is the log (to base e) of the odds ratio or likelihood ratio that the dependent variable is 1.In symbols it is defined as:

$$logit(P) = log\left[\frac{P}{1-P}\right] = ln\left[\frac{P}{1-P}\right] \quad (6.2)$$

Whereas p can only range from 0 to 1, logit (p) scale ranges from negative infinity to positive infinity and is symmetrical around the logit of 0.5 (which is zero). The formula below shows the relationship between the usual regression equation (a + bx ... etc.), which is a straight line formula, and the logistic regression equation.

The form of the logistic regression equation is:

$$logit(p(x)) = log\left[\frac{p(x)}{1-p(x)}\right] = a + b_1x_1 + b_2x_2 + \cdots \quad (6.3)$$

This looks just like a linear regression and although logistic regression finds a 'best fitting' equation, just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, which maximizes the probability of getting the observed results given the fitted regression coefficients. A consequence of this is that the goodness of fit and overall significance statistics used in logistic regression is different from those used in linear regression. P can be calculated with the following formula

$$P = \frac{e^{a+b_1x_1+b_2x_2+\cdots}}{1+e^{a+b_1x_1+b_2x_2+\cdots}} \quad (6.4)$$

Where:

p = the probability that a case is in a particular category,

e = the base of natural logarithms (approx. 2.72),

a = the constant of the equation and,

b = the coefficient of the predictor variables.

6.5.3.2 Neural Network (NN)

(Atiya, 2001) reviewed the applications of NN in credit risk analysis and developed an NN. He developed novel indicators for the NN, for the study he collected data from defaulted and solvent US firms. He reported a prediction accuracy of 84.52% for the in-sample set and 81.46% for the out-of-sample set. He proved that the use of the indicators in addition to financial ratios provided significant improvement.

Artificial neural networks (ANNs) are computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn (progressively improve performance) to do tasks by considering examples, generally without task-specific programming. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as "cat" or "no cat" and using the analytic results to identify cats in other images. They have found most use in applications difficult to express in a traditional computer algorithm using rule-based programming.

An ANN is based on a collection of connected units called artificial neurons, (analogous to axons in a biological brain). Each connection (synapse) between neurons can transmit a signal to another neuron. The receiving (postsynaptic) neuron can process the signal(s) and then signal downstream neurons connected to it. Neurons may have state, generally represented by real numbers, typically between 0 and 1. Neurons and synapses may also have a weight that varies as learning proceeds, which can increase or decrease the strength of the signal that it sends downstream. Further, they may have a threshold such that only if the aggregate signal is below (or above) that level is the downstream signal sent (Schmidhuber, 2015).

6.6 Experimental design and performance evaluation

RM v.7.3 were used to perform all computations regarding feature selection (ICA and IG) as well as classification techniques (Logit and NN) with machine setup of Intel Core™ i5-5200 , 2.2 GHz (4 CPUS) processor and 16 GB of RAM.

Selected subset of credit factors has fed to logit and NN. SVM uses polynomial kernel to represent the ratios vector, the kernel degree was equal 2, C parameter that sets the tolerance for misclassification, where higher C values allow for 'softer' boundaries and lower values create 'harder' boundaries. A complexity constant that is too large can lead to over-fitting, while values that are too small may result in over-generalization so it was determined to be equal to 0.1.

The model by means of a feed-forward neural network trained by a back propagation algorithm (multi-layer perceptron) was configured by 15 hidden layers , the weights has been changed using learning rate equals to .3 and momentum equals to .2 . All these values have chosen by the guidance of trial and best accuracy results. The methodology is validated on the benchmark of dull dataset of financial ratios. In the evaluation, the entire dataset contains 506 records of financial ratios. Classification results are obtained using 5-fold cross-validation and 60% split (i.e. 60% training, 40% testing). To evaluate the performance of the classification, the standard statistical indices of sensitivity, specificity and accuracy were applied.

6.7 Results and discussion

In order to distinguish between results of studied feature selection techniques, accuracy and number of selected ratios have been used to differentiate and select the one with superior performance. The equivalent sensitivity, specificity, accuracy and number of financial ratios of feature selection techniques using Neural Network (NN) classifier were: 97.18%,

51.11%, 94.95% and (2) respectively for the **ICA**; 100%, 60.19.36%, 97.53% and (2) respectively for the **IG.** Table 6.2 clarifies the sensitivity, specificity and accuracy along with number of selected ratios of ICA and IG using feed-forward neural network (NN) classifier.

Classifier	Feature Selection	Number of Features	Sensitivity %	Specificity%	Accuracy%
NN	ICA	2	97.18%	51.11%	94.95%
	IG	2	100%	60.19.36%	97.53%

Table 6.2: Evaluation results of ICA and IG using feed-forward neural network.

The equivalent sensitivity, specificity, accuracy and number of financial ratios of feature selection techniques using logistic regression (logit) classifier were: 100%, 70.57%, 95.27% and (2) respectively for the **ICA**; 99.13%, 72.42%, 98.73% and (2) respectively for the **IG.** Table 6.3 clarifies the sensitivity, specificity and accuracy along with number of selected ratios of ICA and IG using logit classifier.

Classifier	Feature Selection	Number of Features	Sensitivity %	Specificity%	Accuracy%
Logit	ICA	2	100%	70.57%	95.27%
	IG	2	99.13%	72.42%	98.73%

Table 6.3: Evaluation results of ICA and IG using Logistic Regression.

Overall, Logit classifier present better classification results compared to NN with average classification accuracy equals to 97 %. Moreover information gain (IG) shown better classification results compared to ICA, however In logit model ICA showed higher sensitivity with equivalent power of redundant information removing (2 features selected).

Information Gain (IG) produced a sorted weight of selected features which is presented in Table 6.4

Feature	Weight
Sector	.27
Mode of Finance	.01
Payment Method	0

Table 6.4: Ordered features weight using IG.

As inferred from Table 6.4 sector represent the important factor that affect the prediction power of classification system followed by mode of finance while payment method was classified as unimportant feature.

On the basis of the findings the following recommendations arose. It has showed that finance sector is the most influential Islamic banking feature on commercial banks performance. Therefore, commercial banks in Sudan should warn its clients on the need to promote partnership through financing business ideas. Also among the most recommended measures put in place is by selecting key financial and other indicators to monitor programs based on the statutory requirements on Islamic banking products. Developing systems for managing future performance based on the statutory requirements is also highly recommended to absorb the risks of conducting excessive credit operations on highly risk sectors.

Finance mode is the only way banks can guarantee marketing their financial products. Since access to finances is a vital tool for economic development there is need for the development faith based financial products which can be used by investors at different stages of investment such as seed, start-up, expansion, development or bridge finance and working capital finances. Commercial banks should propose alteration on the practices of current Islamic banking products since currently rules are known but the compliance is difficult. Also central bank and commercial banks jointly should set up education platforms geared towards making its customers aware of the benefits and costs associated with Islamic banking products.

6.8 Summary

We have demonstrated the use of risk modeling using logistic regression analysis and neural network models to identify macro finance characteristics associated with likelihood to default on a bank loan. We identified that finance sector and finance mode are important predictors for designed model by using information gain(IG) feature selection method and payment method not considered as important predictor. The combination of IG-Logit was find the best classifier can be used to reliably classify Islamic macro Islamic finance properties with classification accuracy of 98.73%.

Chapter 7

Conclusion and future work

7.1 **Introduction**

This chapter presents the summaries of the proposed techniques. There are several proposed techniques in this thesis that consist of feature selection, parameters optimization and ensemble classification techniques. This research specifically investigates the strength of Sudan's banking sector by using the data collected from all commercial banks operated in Sudan. Our proposed solution to predict the distress of Sudan's banking sector is the first study ever conducted in Sudan banking sector as well as novel approach by dividing the process on identifying the most important factors that affect the banking sector as well as design novel models to discover which is best one to be used in predicting the bank's distress on Sudan.

Moreover, we analyzed the credit process of Sudanese commercial banks concentrating on Islamic finance mode and other macro factors such as sector and payment method to discover which are the most risky factors in the bank's credit procedures.

7.2 The Proposed Method

The proposed methods to tackle the Bank Distress Prediction research addressed both the problem of selecting the most important bank factors or ratios , the problem of designing a best prediction model and the analyzing the potential threats for current practices of Islamic credit and figure out the significant predictors for loan default prediction model. The aim of this research is to propose classification model to predict the failure of Sudan's Banking sector.

The research studied various feature selection techniques and proposed a hybrid DAGA-FS as best feature selection technique, followed by studying different classification models to propose optimized parameters ensemble SVM model as best classification model for predicting banks distress.

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, optimized parameters ensemble 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 8 factors instead of 11. Proposed features selection method DAGA-FS has identified 8 ratios with highest predictive power which are: T1C, TCR, LLP, NPL, ROE, ROA, LADF and RFR, the later ratio is a novel one used for the first time by this research.

Islamic Commercial banks credit records have been analyzed and classification models built with logistic regression analysis to identify macro finance characteristics associated with likelihood to default on a bank loan. We identified that finance mode and sectors are important predictors for designed model. Payment method was not classified as important predictor for Islamic credit risk analysis. The selected model shows satisfactory performance.

7.3 Contribution of the Study

As mentioned earlier in the previous section. The main goal of this thesis is to predict the bank distress of Sudan's banking sector. The highlight of this thesis is that stand alone and hybrid as well as ensemble models were built and are used to solve the classification problem. Therefore, this study reaches a number of contributions for bank distress especially on Sudan banking sector. Major activities corresponding to contributions are summarized as follows:

i. **Identification of important factors that can be used to predict the Sudan's banking sector**

The first step in building a prediction model is the identification of important ratios to be used for training and validation. These factors are attributes that attempt to represent the data used for the task. The quality of the model depends on selection of suitable factors based on their feature scores. This study identified the important features that attempt to represent the data used for the task. This study concentrated on 11 features selected after designing and analyzing a questionnaire targeted SMEs of commercial banks prudential supervision and performing factor analysis to define new potential clusters for those variables and compare them with the genuine CAMELS clusters , we designed a novel Data warehouse to host the raw data that will be used on prediction models.

This research is first published one to use rain fall ratio which never studied before its effect on bankruptcy prediction, and classified as a good predictor with noticeable discriminant power.

ii. **Design a model that can be used to predict the bank financial distress depending on the factors on the first step (i).**

After conducting a comprehensive literature review and related work to same area of bank's distress prediction and after doing several experiments we have selected evolutionary optimized parameters ensemble SVM as a proposed model to be used in this research.

The designed model has been compared with non-optimized parameters single SVM, single NN and optimized single SVM classification models and produce higher classification results. It also benchmarked with state-of-art and promising results were found except with a CMNN model built for company's financial distress prediction.

iii. Identify most trusted Islamic finance methods that have the low risk impacts

Sudan banking sector is adopting the Islamic finance system, there is high percentage of delinquency which can lead to bank collapse unless the credit risk is mitigated in proper and well-informed way, we took the challenge to identify which are the most trusted and safe Islamic macro finance features can be applied on Sudan's banking sector and what are the most risky ones in order to be more controlled and monitored. New and different data set have been designed and reviewed then two features selection techniques were studied namely information gain (IG) and Independent component analysis (ICA), IG was determined the best and abled to weight sector and finance mode as good predictors for credit risk classification systems. Payment method was not classified as important predictor for Islamic credit risk analysis

Good credit personal loans accounted for 7.5% (1,301 people), **poor credit 92.5%** (16,035 people). This is consider very high delinquency ratio which require very extensive investigation to figure out the main causes and threatens, following we analyze these risks from the Islamic finance perspective

7.4 Future Work

The research achieved all the objectives of the study by determining the significant features affect banks performance using the proposed features selection technique DAGA-FS. These features have used to build accurate classification model using the proposed evolutionary optimized parameters ensemble SVM model and finally build end-to-end classification model to contribute on macro analysis for Islamic banking credit risk analysis. However, a number of research opportunities still exist and further research can be conducted into them. Specifically, further studies can be conducted into the following areas:

1. There is an emerging sophisticated procedure in features selection that can be used to confirm this research findings such as least absolute shrinkage and selection operation method (Gorban, 2016) could be employed to strengthen the validity of this study. Also instead of relying on knowledge expert and macro Islamic finance features a large scale of other features can be studied similar to one suggested by this research (rain fall ratio).
2. To insure the external validity of this study and take into account cultural and environmental differences (Rain Fall Ratio) between countries. It could be deployed on other national settings.
3. CMNN model showed distinctive classification accuracy in company's domain (Chang-Chih, 2016). There is opportunity to enhance this work and customize it in favor of banks distress classification system.
4. With continuous influencing of social media on economic system because Social media is transforming banking relationships in very significant ways, from improving customer service to allowing users to send money to others via online platforms. New financial technology companies are using social media to help people simply open a bank account. Social media can even impact the ability to get a loan; a text mining model can be built to study the effect of social media news and complaints on Sudan's banking sector.

The goals of future work will be achieved as the following objective:

- To study absolute shrinkage and selection operation method as a novel features selection methodology for bankruptcy prediction.
- To identify other bank factors that can also affect and contribute on Sudan's banking failure.
- 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.

- To study design a novel classification model such as CMNN and replicate it on external environment with national setting.

7.4 Summary

The goals of this research are to predict Sudan's banking sector distress, which considers the problem of selecting the most important factors affect our banking sector, problem of building the best classification model and to identify the risks surrounding Islamic credit operations in Sudan. This chapter presents the summaries of the proposed techniques that have been done in this research. This study proposes various bank factors which believed to affect the performance of Sudan's banking sector as well as discovering the features affect the credit decisions of Islamic financing from macro perspective as well as their equivalent classification models. The results show that the research goals have been achieved. Furthermore, this research discusses the plan for future work to improve the current work and how it will be applied to another data source such as texts.

References

***A**apo Hyvarinen, Juha Karhunen, Erkki,Oja (2001). Independent component analysis (1st ed.). New York: J. Wiley.* ISBN 0-471-22131-7.

Adamson, C., 2010. "*the complete reference star schema"*. New York: McGraw-Hill Osborne Media.

Adnan Aziz, M. and Dar, H.A., 2006. 'Predicting corporate bankruptcy: Where we stand?', *Corporate Governance: The international journal of business in society*, 6(1), pp. 18–33.

Albertazzi, U. and Gambacorta, L., 2009. 'Bank profitability and the business cycle', *SSRN Electronic Journal*, . doi: 10.2139/ssrn.935026.

Alfaro, E., GJmez, M., & Garca, N., 2007. Multiclass corporate failure prediction by **A**daBoost. M1. Advanced Economic Research, 13, 301–312.

Alhaj TA, Siraj MM, Zainal A, Elshoush HT, Elhaj F, 2016. Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation. PLOS ONE 11(11): e0166017.

Altman, E., Fargher, N. and Kalotay, E., 2011. 'A simple empirical model of equity-implied probabilities of default', *The Journal of Fixed Income*, 20(3), pp. 71–85.

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

Altman, E.I. and Saunders, A., 1997. 'Credit risk measurement: Developments over the last 20 years', *Journal of Banking & Finance*, 21(11-12), pp. 1721–1742.

Altman, E.I., Haldeman, R.G. and Narayanan, P., 1977. 'ZETATM analysis A new model to identify bankruptcy risk of corporations', *Journal of Banking & Finance*, 1(1), pp. 29–54.

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 & Finance,* 18(3), pp. 505–529.

Angelini, E., di Tollo, G. and Roli, A., 2008. 'A neural network approach for credit risk evaluation', *The Quarterly Review of Economics and Finance*, 48(4), pp. 733–755.

Atiya, 2001. "Bankruptcy prediction for credit risk using neural networks: A survey and new results", IEEE Transactions on Neural Networks 12 (4) 929–935.

AVCI, E. and ÇİNKO, M., 2008. 'Estimation of index returns by models of artificial neural networks: Developing European Stock Exchange', İktisat İşletme ve Finans, 23 (266).

Aykut Ekinci, Halil Ibrahim Erdal, 2016. "Forecasting Bank Failure: Base Learners, Ensembles and Hybrid Ensembles" Computer Econ DOI 10.1007/s10614-016 9623-y.

Back, T. ,1996.. "*Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*". Oxford University Press on Demand.

Bardos M. ,2001. : 'Analyse discriminante: application au ri sque et scoring financier, Dunod'.

Bardos M., Zhu W.H., 1997. 'Comparaison de l'analyse discriminante linéaire et des réseaux neuronaux : application à la détection de défaillance d'entreprises', Revue StatistiqueAppliquée, XLV (4), 65-92.

Barker, Holdsworth, D., 1993. 'The Causes of Bank Failures in the 1980s', Research *Paper No. 9325*, Federal Reserve Bank of New York.

Barr, R.S., Killgo, K.A., Siems, T.F. and Zimmel, S., 2002. 'Evaluating the productive efficiency and performance of US commercial banks', *Managerial Finance*, 28(8), pp. 3–25.

Bastien P, Vinzi V. E, Tenenhaus M., 2005. 'PLS generalized linear regression', Computational Statistics and Data Analysis, 48-1.17-46.

Beaver, W.H., 1966. 'Financial ratios as predictors of failure', *Journal of Accounting Research*, 4, p. 71.

Benli, Y. K.,2005. 'Bankalarda Mali Basarisizligin Ongorulmesi Lojistik Regresyon ve Yapay Sinir Agi Kars laştirmasi' , Gazi Universitesi Endustriyel Sanatlar Egitim Fakultesi Dergisi, 16, 31-46.

Bennett, K.P., Cristianini, N., Shawe-Taylor, J. and Wu, D., 2000. 'Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis', *Machine Learning*, 41(3), pp. 295–313.

Bernhard Sc̈holkopf ,Alex J. Smola, Klaus-Robert M̈uller , 1998. , 'Nonlinear component analysis as a kernel eigen value problem', Neural Computation, 10:1299–1319,1998.

Betz, F., Oprică, S., Peltonen, T.A. and Sarlin, P., 2014. 'Predicting distress in European banks', *Journal of Banking & Finance*, 45, pp. 225–241.

Bikker, J.A. and Haaf, K., 2002. 'Competition, concentration and their relationship: An empirical analysis of the banking industry', *Journal of Banking & Finance*, 26(11), pp. 2191–2214.

Blum, M., 1974. 'Failing company Discriminant analysis', *Journal of Accounting Research*, 12(1), p. 1.

Bolt, W., de Haan, L., Hoeberichts, M.M., van Oordt, M.R.C. and Swank, J. , 2012. 'Bank profitability during recessions', *SSRN Electronic Journal.*

Bolton, C. , 2009. 'Logistic regression and its application in credit scoring', University of Pretoria, , pp. 38–56.

Bourke, P. , 1989. "Concentration and other determinants of bank profitability in Europe, north America and Australia", *Journal of Banking & Finance*, 13(1), pp. 65–79.

Breiman , 1996. “Bagging predictors. Machine Learning” 24(2) 123–140, 402.

Brooks, S. and Stevens, J. , 2000. “Applied Multivariate statistics for the social sciences”, The Statistician, 43(1), p. 219.

C.-F. Tsai (2009) “*Feature selection in bankruptcy prediction*” Knowledge.-Based Syst. 22 (2) 120–127.

Calomiris, C.W., Himmelberg, C.P. and Wachtel, P. , 1995. ‘Commercial paper, corporate finance, and the business cycle: A microeconomic perspective’, *Carnegie-Rochester **C**onference Series on Public Policy*, 42, pp. 203–250.

Canbas, S., Cabuk, A. and Kilic, S.B. , 2005. “Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case”, *European Journal of Operational Research*, 166(2), pp. 528–546.

CBOS, 2016., http://www.cbos.gov.sd (Accessed: 20 January 2017).

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.

Chang-Chih Chung, Tsung-Shih Chen, Lee-Hsuan Lin, Yu-Chen Lin,Chih-Min Lin ,2016. “Bankruptcy Prediction Using Cerebellar Model Neural Networks” Int. J. Fuzzy Syst. 18(2):160–167.

Chauhan, N., Ravi, V. and Karthik Chandra, D., 2009. ‘Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks’, *Expert Systems with Applications*, 36(4), pp. 7659–7665.

Child, D., 2006. ‘The essentials of factor analysis. (3rd ed.)’ , New York, NY: Continuum International Publishing Group.

Coats, P.K. and Fant, L.F., 1993. ‘Recognizing financial distress patterns using a neural network tool’, *Financial Management*, 22(3), p. 142.

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

Costello, Anna B. and Jason Osborne, 2005. "Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Practical Assessment Research & Evaluation", 10(7).

Curry, T.J., Fissel, G.S. and Elmer, P.J., 2009. 'Can the equity markets help predict bank failures?', *SSRN Electronic Journal*.

D.K. Chandra, V. Ravi, I. Bose (2009) "*Failure prediction of dotcom companies using hybrid intelligent techniques*" Expert Syst. Appl. 36 (4830–4837.

Deakin, E.B. ,1972. 'A Discriminant analysis of predictors of business failure', *Journal of Accounting Research*, 10(1), p. 167.

Dean P. Foster and Robert A. Stine ,2004. "Journal Article Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy" Journal of the American Statistical Association Vol. 99, No. 466 (Apr., 2004), pp. 303-313.

***D**eco & Obradovic (1996). An Information-Theoretic Approach to Neural Computing. New York, NY: Springer.*

Demirguc-Kunt, A. and Huizinga, H. ,1999. 'Determinants of commercial bank interest margins and profitability: Some international evidence', *The World Bank Economic Review*, 13(2), pp. 379–408.

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

Deron Liang ,2015. 'The effect of feature selection on financial distress prediction' Knowledge-Based Systems v73.

Deron Liang ,2015. "The effect of feature selection on financial distress prediction " Knowledge-Based Systems 73 (2015) 289–297.

Derviz, A. and Podpiera, J., 2008. 'Predicting bank CAMELS and S&P ratings: The case of the Czech Republic', *Emerging Markets Finance and Trade*, 44(1), pp. 117–130.

Detragiache, E. and Demirgüç-Kunt, A. ,1998. 'Financial liberalization and financial fragility', *IMF Working Papers*, 98(83), p. 1.

Dietterich (1998) "Machine Learning Research: Four Current Directions". The AI Magazine, 18(4) 97–136 400.

Doumpos, M. and Zopounidis, C. ,2010. 'A multicriteria decision support system for bank rating', *Decision Support Systems*, 50(1), pp. 55–63. doi: 10.1016/j.dss.2010.07.002.

Duncan, E. and Elliott, G. ,2004. 'Efficiency, customer service and financial performance among Australian financial institutions', *International Journal of Bank Marketing*, 22(5), pp. 319–342.

Duttweiler, R. ,2009. '*Managing liquidity in banks: A top down approach*. Chichester', United Kingdom: Wiley, John & Sons.

Dzeawuni, W.A. and Tanko, D.M. ,2009. 'CAMELs and banks performance evaluation: The way forward', *SSRN Electronic Journal*.

Eccles, R.G., Ioannou, I. and Serafeim, G., 2014. 'The impact of corporate sustainability on organizational processes and performance', *Management Science*, 60(11), pp. 2835–2857.

Edmister, R.O., 1972. 'FINANCIAL RATIOS AS DISCRIMINANT PREDICTORS OF SMALL BUSINESS FAILURE*', *The Journal of Finance*, 27(1), pp. 139–140.

Eiben, A. E., and Smith, J. E, 2003. "Introduction to evolutionary computing". Springer, Berlin, Heidelberg, New York .

Eiben, A.E., Smit, S.K , 2011. "Parameter tuning for configuring and analyzing evolutionary algorithms. Swarm and Evolutionary Computation" p. 19 31.

Erdal, H.I. and Ekinci, A. ,2012. ‘A comparison of various artificial intelligence methods in the prediction of bank failures’, *Computational Economics*, 42(2), pp. 199–215. doi: 10.1007/s10614-012-9332-0.

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).

Etemadi, H., Anvary Rostamy, A.A. and Dehkordi, H.F. ,2009. ‘A genetic programming model for bankruptcy prediction: Empirical evidence from Iran’, *Expert Systems with Applications*, 36(2), pp. 3199–3207.

Fengyi Lin , Deron Liang, Ching-Chiang Yeh , Jui-Chieh Huang ,2014. "Novel feature selection methods to financial distress prediction" Expert Systems with Applications 41 2472–2483.

Flannery , M.J. and Protopapadakis, A. ,1998. ‘Macroeconomic factors DO influence aggregate stock returns’, *SSRN Electronic Journal*.

Fodor, I.K. and Kamath, C., 2002. ‘Dimension reduction techniques and the classification of bent double galaxies’, *Computational Statistics & Data Analysis*, 41(1), pp. 91–122.

***F**ogel, David B. ,1998. “Evolutionary Computation: The Fossil Record”. New York: IEEE Press.* ISBN 0-7803-3481-7.

Frost, Stephen M., 2004. ‘Chapter 20 -Corporate Failures and Problem Loans’ .*The Bank Analyst's Handbook: Money, Risk and Conjuring Tricks*. Chichester, United Kingdom: Wiley, John & Sons.

G. C. Cawley and N. L. C. Talbot ,2010. “*Over-fitting in model selection and subsequent selection bias in performance evaluation*”, Journal of Machine Learning Research, 2010. Research, vol. 11, pp. 2079-2107.

Gaytán, Alejandro; Johnson, Christian A., 2002. ‘A Review of the Literature on Early Warning Systems for Banking Crises’ , Central Bank of Chile Working Paper, No. 183, Santiago.

Gorban, A.N.; Mirkes, E.M.; Zinovyev, A. ,2016. "Piece-wise quadratic approximations of arbitrary error functions for fast and robust machine learning." Neural Networks, 84, 28-38.

Gorsuch, R.L. and L, R. ,1983. Factor analysis. 2nd edn. United States: Lawrence Erlbaum Associates.

Grier, A., 2007. 'Credit Analysis of Financial Institutions', *Euro money Institution Investor PLC*, United Kingdom.

Gungor, B. , 2007. 'Turkiye'de Faaliyet Gosteren Yerel ve Yabancı Bankaların Karlılık Seviyelerini ' Etkileyen Faktorler: Panel Veri Analizi. İktisat İsletme ve Finans, 22(258), 40-63.

Guo, X., Zhu, Z. and Shi, J. , 2012. 'A corporate credit rating model using support vector domain combined with fuzzy clustering algorithm', *Mathematical Problems in Engineering*, 2012, pp. 1–20.

Guyon, A. (2003) 'Elisseeff, An introduction to variable and feature selection, J. Mach'. Learn. Res. 3 1157–1182.

Harman, H.H. and Kockov-KratochvlovA. , 1976. 'Use of factor analysis to classify strains of yeasts: Application to genus torulopsis berlese', *Journal of Mathematical Biology*, 3(1), pp. 27–52.

Hassan, M. K., & Bashir, A.- H. M. , 2003. 'Determinants of Islamic banking profitability'. Economic Research Forum Conference 10th annual conference.

HSSB, H. ,2015. Publications. Available at: http://www.hssb.gov.sd (Accessed: 16 January 2017).

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. , 2016. 'A Practical Guide to Support Vector Classification', *National Taiwan University,* , pp. 1–16.

***J**ackson, J.E. and Bartholomew, D., 2008. 'Latent variable models and factor analysis',* Technometrics*, 31(2), p. 266.*

Janer, J., 2011. 'Bankruptcy Prediction and its Advantages Empirical E vidence from S MEs in the French Hospitality I ndustry', *Department of Economics Copenhagen Business School*.

Jarke, M., Quix, C., Calvanese, D., Lenzerini, M., Franconi, E., Ligoudistianos, S., Vassiliadis, P. and Vassiliou, Y. , 2000. 'Concept based design of data warehouses', *ACM SIGMOD Record*, 29(2), p. 591.

John, G. H., Kohavi, R., & Pfleger, K. ,1994. "*Irrelevant features and the subset selection problem*". Proceedings of ICML-94.

Kandrac, J. ,2014. 'Modelling the causes and manifestation of bank stress: An example from the financial crisis', *Applied Economics*, 46(35), pp. 4290–4301.

Karlyn, Mitchell ,1984. 'Capital Adequacy at Commercial Banks', *The Journal of Economic Review*, p. 17-30.

Kaufman.

Keerthi, S.S. and Lin, C.-J. ,2003. 'Asymptotic behaviors of support vector machines with Gaussian kernel', Neural Computation, 15(7), pp. 1667–1689.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. and Murthy, K.R.K., 2003. 'A fast iterative nearest point algorithm for support vector machine classifier design', *IEEE Transactions on Neural Networks*, 11(1), pp. 124–136.

Kennedy J, Eberhart RC ,2001. "Swarm intelligence". San Mateo: Morgan Kaufman; 2001.

Kent, C. and P. Lowe, 1997. "Property-Price Cycles and Monetary Policy', *paper presented at the Central Bank Economists* Meeting", Basle, 28 to 29 October.

Khalafalla Ahmed., 2013. "Predicting Banks' Failure: The Case of Banking Sector in Sudan for the Period" (2002-2009) , JBSQ , Volume 4.

Kline, P. and Paul, K. ,1993. "An easy guide to factor analysis". New York: Taylor & Francis.

Knight, B., Veerman, E. and Dickinson, G. ,2008. "*Professional Microsoft SQL server 2008 integration services"*. Indianapolis, IN: Wiley Pub.

Kohavi, R., John, G. ,1997. "*Wrappers for feature subset selection. Artificial Intelligence*" Journal, Special issue on relevance 97(1-2), 273–324.

Kong, F. and Wang, W. ,2009. 'The Early Warning of Financial Crisis in China: Based on the Model of Artificial Neural Network', *Information Engineering and Computer Science, 2009. ICIECS.*

Kosmidou, K., Pasiouras, F., Zopounidis, C. and Doumpos, M. ,2006. 'A multivariate analysis of the financial characteristics of foreign and domestic banks in the UK', *Omega*, 34(2), pp. 189.

Lechtenbörger, J. and Lechtenborger, J. ,2001. "*Data warehouse schema design*. Berlin": IOS Press, US.

Lin, F., Yeh, C.-C. and Lee, M.-Y. , 2011. 'The use of hybrid manifold learning and support vector machines in the prediction of business failure', *Knowledge-Based Systems*, 24(1), pp. 95–101.

Lin, G.-F., Chen, G.-R., Huang, P.-Y. and Chou, Y.-C. , 2009. 'Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods', *Journal of Hydrology*, 372(1-4), pp. 17–29.

Martin, D., 1977. 'Early warning of bank failure', *Journal of Banking & Finance*, 1(3), pp. 249–276.

Max Kuhn and Kjell Johnson, 2013."Applied Predictive Modeling", Amazon Books, ISBN-13: 978-1461468486.

Mazzillo, JA. ,1993. 'The structure of the US banking system and banking supervision'.

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.

Mohammad Ahmad Al-Saleh , Ahmad Mohammad Al-Kandari ,2011. 'Prediction of Financial Distress for Commercial Banks in Kuwait' World Review of Business Research Vol. 2. No. 6.

Molyneux, P. and Thornton, J. ,1992. 'Determinants of European bank profitability: A note', *Journal of Banking & Finance*, 16(6), pp. 1173–1178.

Mu-Yen Chen ,2011. '*Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches*', Computers and Mathematics with Applications 62.

Myoung-Jong Kim , Dae-Ki Kang ,2010. "Ensemble with neural networks for bankruptcy prediction" Expert Systems with Applications 37 - 3373–3379.

Ohlson, J.A. ,1980. 'Financial ratios and the probabilistic prediction of bankruptcy', *Journal of Accounting Research*, 18(1), p. 109.

Orsenigo, C. and Vercellis, C. , 2013. 'Linear versus nonlinear dimensionality reduction for banks' credit rating prediction', *Knowledge-Based Systems*, 47, pp. 14–22.

Paliwal, M. and Kumar, U.A. , 2009. 'Neural networks and statistical techniques: A review of applications', *Expert Systems with Applications*, 36(1), pp. 2–17.

Pam, W.B. ,2013. 'Discriminant Analysis and the Prediction of Corporate Bankruptcy in the Banking Sector of Nigeria', *International Journal of Finance and Accounting*, 2(6), pp. 319–325.

Pasiouras, F., Gaganis, C. and Zopounidis, C. ,2006. 'The impact of bank regulations, supervision, market structure, and bank characteristics on individual bank ratings: A cross-country analysis', *Review of Quantitative Finance and Accounting*, 27(4), pp. 403–438.

Perrone, M. E. ,1994. Putting it all together: Methods for combining neural networks. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.). Advances in

Neural Information Processing Systems (Vol. 6, pp. 1188–1189). San Mateo, CA: Morgan

Perry, P. ,1992. 'Do Banks Gain or Lose from Inflation', *Journal of Retail Banking* 16, 25-30.

***P**ett, M.A., Lackey, N.R. and Sullivan, J.J. ,2003.* Making sense of factor analysis: The use of factor analysis for instrument development in health care research. *Thousand Oaks, CA: SAGE Publications.*

***Q**i, X., Luo, R., Carroll, R.J. and Zhao, H. ,2015. 'Sparse regression by projection and sparse Discriminant analysis', Journal of Computational and Graphical Statistics, 24(2), pp. 416–438.*

***R**aiyani, JR. ,2010. 'Effect of mergers on efficiency and productivity of Indian banks: A CAMELS analyses ', ASIAN JOURNAL OF MANAGEMENT RESEARCH.*

Refait, C. ,2004. 'La prévision de la faillite fondée sur l'analyse financière de l'entreprise: Un état des lieux', *Économie & prévision*, 162(1), pp. 129–147.

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.

Richard Eldridge ,2016. "https://www.weforum.org/agenda/2016/04/how-social-media-is-shaping-financial-services.".

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.

Roman, A. and Şargu, A.C. ,2013. 'Analysing the financial soundness of the commercial banks in Romania: An approach based on the Camels framework', *Procedia Economics and Finance*, 6, pp. 703–712.

Rostami, M. ,2015. ‘Determination of Camels model on bank’s performance’, *International Journal of Multidisciplinary Research and Development*, 2(10), pp. 652–664.

Salhuteru, F. and Wattimena, F. ,2015. ‘Bank performance with CAMELS ratios towards earnings management practices in state banks and private banks’, *Advances in Social Sciences Research Journal*, 2(4).

Salzberg, S.L. ,1994. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. Machine Learning. 16(3), 235-240.

Sarle. Neural Network FAQ, ,1997. URL ftp://ftp.sas.com/pub/neural/ FAQ.html. Periodic posting to the Usenet newsgroup comp.ai.neural-nets.

***S**chmidhuber, J. ,2015. "Deep Learning in Neural Networks: An Overview". Neural Networks. 61: 85–117.*

Short, B.K. ,1979. ‘The relation between commercial bank profit rates and banking concentration in Canada, western Europe, and Japan’, *Journal of Banking & Finance*, 3(3), pp. 209–219.

Sinkey, J.F. ,1975.‘A Multivariate statistical analysis of the characteristics of problem banks’, *The Journal of Finance*, 30(1), p. 21. doi: 10.2307/2978429.

***S**rinivas, M., & Patnaik, L. M. ,1994. “Genetic algorithms: A survey”. Computer, 27(6), 17–26.*

Sun, J. and Li, H. ,2011. ‘Dynamic financial distress prediction using instance selection for the disposal of concept drift’, *Expert Systems with Applications*, 38(3), pp. 2566–2576.

Tabachnick, B.G. and Fidell, L.S. ,2012. Using Multivariate statistics (6th edition). 6th edn. Boston: Pearson Education.

Tam, K.Y. and Kiang, M.Y. ,1992. ‘Managerial applications of neural networks: The case of bank failure predictions’, *Management Science*, 38(7), pp. 926–947. doi: 10.1287/mnsc.38.7.926.

Tenenbaum, J.B. ,2000. ‘A global geometric framework for Nonlinear Dimensionality reduction’, *Science*, 290(5500), pp. 2319–2323.

Tsai, C. F. ,2009. “Feature selection in bankruptcy prediction. Knowledge-Based Systems", 22(2), 120–127.

Usmani, M.M.T. and Taqi, U.M.M. ,1999. An introduction to Islamic finance. New Delhi: Idara Isha’at-e-Diniyat (P).

V. Torra ,2006. “Companies’ Financial Distress Prediction”. (Eds.): MDAI, LNAI 3885, pp. 274 – 282.

Varetto, F. ,1998. ‘Genetic algorithms applications in the analysis of insolvency risk’, *Journal of Banking & Finance*, 22(10-11), pp. 1421–1439.

Vasileios Pappasa, Marwan Izzeldina, Ana-Maria Fuertesb, Steven Ongenac (2014) ‘A Survival Analysis of Islamic Bank Failure Risk’.

***V**elicer, W.F. ,2000. ‘Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components’,* Multivariate Behavioral Research, *22(2), pp. 41–71*.

Vinãls, J., Pazarbasioglu, C., Surti, J., Narain, A., Erbenova, M. and Chow, J. ,2013. ‘Creating a safer financial system: Will the Volcker, Vickers, and Liikanen structural measures help?’, *Staff Discussion Notes,* 13(4), p. 1.

Wheelock, D.C. and Wilson, P.W. ,2000. ‘Why do banks disappear? The determinants of U.S. Bank failures and acquisitions’, *Review of Economics and Statistics*, 82(1), pp. 127–138.

Wilson, J.W. and Jones, C.P. ,1987. ‘Common stock prices and inflation: 1857–1985’, *Financial Analysts Journal*, 43(4), pp. 67–71.

Xiao-Feng Hui and Jie Sun ,2006. “An Application of Support Vector Machine to Companies’ Financial Distress Prediction” V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 274.

Yap J. T. ,1998., "*Developing an Early Warning System for BoP and Financial Crises: the Case of thePhilippines*." Philippine Institute for Development Studies, Discussion Paper Series No. 98-40.

Yinhua Li , Yong Shi , Anqiang Huang, Haizhen Yang ,2014. "Failure Prediction in Commercial Banks with a Hybrid Prediction Model" Ann. Data. Sci. 1(2):209–220.

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 & Accounting*, 15(1), pp. 27–45.

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

Zhou ,2014. 'Obtaining functional topics from source code based on topic modeling and static analysis', *SCIENTIA SINICA Informationis*.

List of Publications

- **Local Journals**
  - Salim, N. and Mohammed A SirElkhatim. (2014). Prediction of bank's financial distress - review, SUST Journal of Engineering and Computer Sciences (JECS), Vol. 16, No. 1 . 2015, pp.42-55, ISSN 0128-3790

Status: Published.

- **International Journals**
  - Mohammed A. SirElkhatim, Naomie Salim (2017). Predicting Bank Financial Failures Using Discriminant Analysis, And Support Vector Machines Methods: A Comparative Analysis in Commercial Banks in Sudan (2006-2014), INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 6, ISSUE 04, APRIL 2017.

Status: Published.

  - Mohammed A. SirElkhatim, Naomie Salim (2017). Islamic Credit Risk Analysis: Case Of Sudanese Banking Sector (2006-2014), INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 6, ISSUE 06, JUNE 2017

Status: Published.

  - Mohammed A. SirElkhatim, Naomie Salim (2017). Optimized Parameters Ensemble SVM Bank Distress Classification, Journal of Machine Learning Research. Aug 2017.

Status: Submitted.

APPENDIX A

Factors Elicitation Questionnaire

Bank Distress Prediction Problem

This questionnaire is aimed to identify the most applicable indicators that can be used to predict the bank distress problem of Sudan's banking sector , the selection of given indicators is based on intensive review of the related work and existing literature.

The indicators had taken from CAMELS standard ratios groups without neglect Non-CAMELS ratios which can affect bank distress prediction.

The questionnaire will be answered by subject matter experts in central bank of sudan.

1. Department: …….. .
2. Years of experience: …….. .

3. In your opinion which the most important indicators that can be used to predict the bank distress problem

   A. Capital Adequacy Indicators :

Feature	Description	Selection Justifications

   B. Asset Quality indicators :

Feature	Description	Selection Justifications

   C. Management Quality indicators:

Feature	Description	Selection Justifications

D. Profitability indicators:

Feature	Description	Selection Justifications

E. Liquidity indicators:

Feature	Description	Selection Justifications

F. Non-CAMEL indicators:

Feature	Description	Selection Justifications

G. Macroeconomic indicators

Feature	Description	Selection Justifications

H. Market Structure indicators

Feature	Description	Selection Justifications

I. Other Indicators (Not Mentioned above)

Feature	Description	Selection Justifications

4. If you have any suggestions regarding the bank distress prediction problem , lets know about :

……
……
……
……
……
……
……
……
……
……
……
……
……

End of the Questionnaire

Thanks a lot of your cooperation

APPENDIX B

Description of financial ratios extracted from questionnaire results

INDEX	Description
SET	Shareholder's equity to total assets
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
SHN	Share-holder's equity + net profit to total assets + off balance sheet commitments
EAS	Ratio of equity capital to total asset.
LAS	Ratio of net loans to total assets.
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.
IBR	Interbank ratio measured as a ratio of deposits due From banks to deposits due to banks
LADF	Ratio of liquid assets to deposits and short term funds.
IDIVER	Finance related income to total income
PMA	Permanent assets to total assets
ITO	interest income to total operating income
PEA	Personnel expenses to average assets
DEA-EF	Efficiency score estimated based on the DEA-smoothed Bootstrap approach.
RFR	RainFall Ratios

Questionnaire Result

Table B.1: Initial Set of variables

Dept	Y of Exp	CA	AQ	MQ	Profitability	Liquidity	Non-Camel	Macroeconomic	Market	Other
Prudential Supervision	5	T1C,EAS,SET	LAS,LLP,NPL	NA	NIM	IBR	IDIVER	GDPG	NA	NA
Prudential Supervision	12	TCR	PMA	NA	ITO	LADF	IDIVER	INF	NA	NA
Prudential Supervision	2	EAS	LAS,LLP,NPL	NA	ROE,ROA	LADF	NA	GDPG	NA	NA
Prudential Supervision	6	T1C,EAS	NPL	NA	ROE,ROA,NIM	LADF	NA	GDPG	NA	NA
Prudential Supervision	2	EAS,SET	LLP	NA	ROA	LADF	NA	GDPG	NA	NA
Prudential Supervision	9	TCR,T1C	PMA,LLP	PEA	ROE	LADF	NA	GDPG	NA	NA
Prudential Supervision	12	T1C	LAS	DEA-EF	ROE	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	16	TCR,EAS,T1C	LLP,NPL	NA	ROA	LADF	NA	GDPG	NA	NA
Prudential Supervision	20	SET	LAS,LLP,NPL	NA	ROE,ROA,NIM	LADF	NA	GDPG	NA	NA
Prudential Supervision	4	TCR	NPL	NA	NIM,ROA	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	5	EAS	LAS,LLP	DEA-EF	ROE,ROA,NIM	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	9	EAS	LLP,NPL	NA	NIM,ROA	LADF	IDIVER	GDPG,INF	NA	NA

Prudential Supervision	8	T1C	NPL	NA	ROE,ROA,NIM	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	6	EAS,TCR	LAS	NA	ITO	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	9	SET	LAS	NA	ROE,ROA	LADF	NA	INF	NA	NA
Prudential Supervision	3	EAS	LAS	NA	ROA	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	5	TCR,T1C	LAS	NA	ROE,ROA	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	5	T1C,EAS,SET	LLP,NPL	NA	ROE,ROA,NIM	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	6	EAS,TCR	LLP	NA	ROE,ROA,NIM	LADF	NA	INF	NA	NA
Prudential Supervision	7	TCR	NPL	NA	NIM,ROA	LADF	NA	GDPG,INF	NA	NA
Prudential Supervision	9	TCR,T1C	PMA	NA	ROE	LADF	NA	GDPG	NA	NA
Prudential Supervision	11	T1C	NLP	NA	ROE	NA	NA	INF	NA	NA
Prudential Supervision	13	TCR,EAS,T1C	NPL	NA	ROA	IBR	NA	INF	NA	NA
Prudential Supervision	15	EAS	LLP	NA	ROE,ROA,NIM	NA	NA	GDPG	NA	NA
Prudential Supervision	2	EAS,TCR	PMA	NA	ROE,ROA	LADF	NA	GDPG	NA	NA
Prudential Supervision	4	EAS,TCR	LAS	NA	NIM,ROA	NA	NA	INF	NA	NA

Where

Table B.2: Questionnaire Results

Shortcut	Description
CA	Capital Adequacy
AQ	Asset Quality
MQ	Management Quality

Variables in table B.1 will form initial input factor for factor analysis task.

APPENDIX C

Banks operating in Sudan:

#	Bank's Name
1	Agricultural Bank
2	Savings and Social Development Bank
3	Industrial Development Bank.
4	El -Nilien Bank
5	Bank of Khartoum
6	Real Estates Commercial Bank
7	Faisal Islamic Bank
8	Sudanese French Bank
9	National Bank of Sudan
10	Blue Nile Mashreq Bank
11	Sudanese Islamic Bank
12	Tadamon Islamic Bank
13	Al Nile Bank For Commerce and Development
14	Baraka Bank (Sudan)
15	Export Development Bank
16	Saudi Sudanese Bank
17	Workers' National Bank
18	Animal Resources' Bank
19	Al -Shamal Islamic Bank
20	Farmer's Commercial Bank
21	Omdurman National Bank
22	African Bank for Trade and Development
23	Byblos Bank (Africa)
24	Alsalam Bank
25	Sudanese Egyptian Bank
26	United Capital Bank

27	Aljazeera Sudanese Jordanian Bank
28	Family Bank
29	Financial Investment Bank
30	Abu Dhabi National Bank
31	Qatar National Bank
32	Arab Sudanese Bank
33	National Bank of EGYPT (Khartoum)
34	Alkhaleej Bank
35	Qatar Islamic Bank
36	Abu Dhabi Islamic Bank

APPENDIX D

Description of Selected financial Ratios

ID	INDEX	Description
1	T1C	Tier 1 capital Ratio measured as a ratio of Tier 1 capital to risk weighted assets.
2	TCR	Total Capital ratio measured as a ratio of (Tier 1 + Tier 2 capital) to risk weighted assets
3	EAS	Ratio of equity capital to total asset.
4	LAS	Ratio of net loans to total assets.
5	LLP	Ratio of loan loss provisions to total loans.
6	NPL	Ratio of non-performing loans to total loans
7	ROE	Return on equity measured as a ratio of net Income to Capital equity.
8	ROA	Return on assets measured as a ratio of net Income to total assets.
9	LADF	Ratio of liquid assets to deposits and short term funds.
10	IDIVER	Finance related income to total income
11	RFR	Rain Fall Ratio

D.1: Selected financial ratios

Field	Source	Calculation	Table
Paid-up capital	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=4110
Legal Reserve	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=4130
General Reserve	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=4140
Retained earnings	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=4150
MonthLoss	Balance Sheets DB - Direct		BBS_A(RL_CUR(-)) ; CD_EN=4440
risk weighted assets	Calculated (Balance Sheets DB)	MonthLoss- revaluation Reserve - 5* (General provision) - .5(EquityParticipation+ total asset)	A_CAP_Sum-Calculated Item
revaluation Reserve	Calculated (Balance Sheets DB)	Capital reserve*.45	BBS_A(RL_CUR) ; CD_EN=4440
General provision	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=4120*.45
EquityParticipation	Balance		BBS_A(RL_CUR) ;

	Sheets DB - Direct		CD_EN=1960
total asset	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=100
ReserveOnCentral Bank	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=1110
Loan	Calculated (Balance Sheets DB)	Loan account + CICs Securities + Of which: Consortium financing + Securities + Advances + Securities + Claims on Federal Government + Securities purchased + Claims on state and local government + Claims on public nonfinancial enterprises + Advances for financing foreign trade + Of which: Securities + Advances for financing foreign trade + Claims on households + Advances for financing foreign trade	BBS_A - Calculated Item
Provsion for bad loans	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=4010
NonPerforming Loan	Financial Data - Direct		FinData(TOTAL_NON_PERF) ;Status='Non-Performing'
Net Income	Calculated (Income Statement	TOTAL Income – (Paid TO Bank OF Sudan + Paid TO Other Deposit Money Bank	BBS_B - Calculated Item

	DB)	+ Paid TO Other Financial Institutions+ Paid ON Deposits+ Paid On Debt Securities+ Other)	
CASH	Balance Sheets DB - Direct		BBS_A(RL_CUR) ; CD_EN=1000
Current Deposit	Calculated (Balance Sheets DB)	Demand Deposits (current account) + Demand deposits (Banks) + Demand deposits (Nonbanks)	BBS_A - Calculated Item
Short-Term Securities	Calculated (Balance Sheets DB)	CICs Securities+ Securities (Banks) + Securities (Nonbanks) + Of which: GMCs + GICs + Securities purchased(South Sudan) + Securities purchased (LocalGov) + Of which: Securities (Nonfinancial)	BBS_A - Calculated Item
Deposits	Calculated (Balance Sheets DB)	Demand deposits (current account) + Savings deposits + Margins on L/C and L/G + Restricted deposits	BBS_A - Calculated Item
Income from finance	Income Statement DB		BBS_B(AMT_EN) ; CD_EN=5040
Total Income	Income Statement DB		BBS_B(AMT_EN) ; CD_EN=590

D.2: Fields Sources and Calculation Formula

APPENDIX E

E.1 Features Selection using hybrid discriminant analysis and genetic algorithm (DAGA-FS) code:

```
<?xml version="1.0" encoding="UTF-8"?><process
version="7.3.001">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process"
compatibility="6.0.002" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve"
compatibility="7.3.001" expanded="true" height="68"
name="Retrieve Pure_Training" width="90" x="45" y="34">
        <parameter key="repository_entry"
value="../data/Pure_Training"/>
      </operator>
      <operator activated="true"
class="nominal_to_binominal" compatibility="7.1.001"
expanded="true" height="103" name="Nominal to Binominal"
width="90" x="179" y="30">
        <parameter key="return_preprocessing_model"
value="false"/>
        <parameter key="create_view" value="false"/>
        <parameter key="attribute_filter_type"
value="single"/>
        <parameter key="attribute" value="Label"/>
        <parameter key="attributes" value=""/>
        <parameter key="use_except_expression"
value="false"/>
        <parameter key="value_type" value="nominal"/>
```

```
        <parameter key="use_value_type_exception"
value="false"/>
        <parameter key="except_value_type"
value="file_path"/>
        <parameter key="block_type" value="single_value"/>
        <parameter key="use_block_type_exception"
value="false"/>
        <parameter key="except_block_type"
value="single_value"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes"
value="true"/>
        <parameter key="transform_binominal" value="false"/>
        <parameter key="use_underscore_in_name"
value="false"/>
      </operator>
      <operator activated="true"
class="linear_discriminant_analysis" compatibility="7.3.001"
expanded="true" height="82" name="LDA" width="90" x="313"
y="34"/>
      <operator activated="true"
class="optimize_selection_evolutionary"
compatibility="7.3.001" expanded="true" height="103"
name="Optimize Selection (Evolutionary)" width="90" x="447"
y="34">
        <parameter key="use_exact_number_of_attributes"
value="true"/>
        <parameter key="restrict_maximum" value="false"/>
        <parameter key="min_number_of_attributes"
value="1"/>
        <parameter key="max_number_of_attributes"
value="1"/>
        <parameter key="exact_number_of_attributes"
value="9"/>
        <parameter key="initialize_with_input_weights"
value="false"/>
        <parameter key="population_size" value="5"/>
        <parameter key="maximum_number_of_generations"
value="10"/>
```

```
        <parameter key="use_early_stopping" value="true"/>
        <parameter key="generations_without_improval"
value="2"/>
        <parameter key="normalize_weights" value="false"/>
        <parameter key="use_local_random_seed"
value="false"/>
        <parameter key="local_random_seed" value="1992"/>
        <parameter key="show_stop_dialog" value="false"/>
        <parameter key="user_result_individual_selection"
value="true"/>
        <parameter key="show_population_plotter"
value="true"/>
        <parameter key="plot_generations" value="10"/>
        <parameter key="constraint_draw_range"
value="false"/>
        <parameter key="draw_dominated_points"
value="true"/>
        <parameter key="maximal_fitness" value="Infinity"/>
        <parameter key="selection_scheme"
value="tournament"/>
        <parameter key="tournament_size" value="0.25"/>
        <parameter key="start_temperature" value="1.0"/>
        <parameter key="dynamic_selection_pressure"
value="true"/>
        <parameter key="keep_best_individual"
value="false"/>
        <parameter key="save_intermediate_weights"
value="false"/>
        <parameter key="intermediate_weights_generations"
value="10"/>
        <parameter key="p_initialize" value="0.5"/>
        <parameter key="p_mutation" value="-1.0"/>
        <parameter key="p_crossover" value="0.5"/>
        <parameter key="crossover_type" value="uniform"/>
        <process expanded="true">
          <operator activated="true"
class="split_validation" compatibility="7.3.001"
expanded="true" height="145" name="Validation" width="90"
x="380" y="34">
```

```
            <parameter key="create_complete_model"
value="false"/>
            <parameter key="split" value="relative"/>
            <parameter key="split_ratio" value="0.7"/>
            <parameter key="training_set_size" value="100"/>
            <parameter key="test_set_size" value="-1"/>
            <parameter key="sampling_type" value="shuffled
sampling"/>
            <parameter key="use_local_random_seed"
value="false"/>
            <parameter key="local_random_seed"
value="1992"/>
            <process expanded="true">
              <operator activated="true"
class="support_vector_machine" compatibility="7.3.001"
expanded="true" height="112" name="SVM (2)" width="90"
x="112" y="30">
                <parameter key="kernel_type" value="dot"/>
                <parameter key="kernel_gamma" value="1.0"/>
                <parameter key="kernel_sigma1" value="1.0"/>
                <parameter key="kernel_sigma2" value="0.0"/>
                <parameter key="kernel_sigma3" value="2.0"/>
                <parameter key="kernel_shift" value="1.0"/>
                <parameter key="kernel_degree" value="2.0"/>
                <parameter key="kernel_a" value="1.0"/>
                <parameter key="kernel_b" value="0.0"/>
                <parameter key="kernel_cache" value="200"/>
                <parameter key="C" value="0.0"/>
                <parameter key="convergence_epsilon"
value="0.001"/>
                <parameter key="max_iterations"
value="100000"/>
                <parameter key="scale" value="true"/>
                <parameter key="calculate_weights"
value="true"/>
                <parameter
key="return_optimization_performance" value="true"/>
                <parameter key="L_pos" value="1.0"/>
                <parameter key="L_neg" value="1.0"/>
```

```
                <parameter key="epsilon" value="0.0"/>
                <parameter key="epsilon_plus" value="0.0"/>
                <parameter key="epsilon_minus" value="0.0"/>
                <parameter key="balance_cost"
value="false"/>
                <parameter key="quadratic_loss_pos"
value="false"/>
                <parameter key="quadratic_loss_neg"
value="false"/>
                <parameter key="estimate_performance"
value="false"/>
              </operator>
              <connect from_port="training" to_op="SVM (2)"
to_port="training set"/>
              <connect from_op="SVM (2)" from_port="model"
to_port="model"/>
              <portSpacing port="source_training"
spacing="0"/>
              <portSpacing port="sink_model" spacing="0"/>
              <portSpacing port="sink_through 1"
spacing="0"/>
            </process>
            <process expanded="true">
              <operator activated="true" class="apply_model"
compatibility="7.1.001" expanded="true" height="76"
name="Apply Model (2)" width="90" x="45" y="30">
                <list key="application_parameters"/>
                <parameter key="create_view" value="false"/>
              </operator>
              <operator activated="true" class="performance"
compatibility="7.3.001" expanded="true" height="76"
name="Performance (2)" width="90" x="179" y="30">
                <parameter key="use_example_weights"
value="true"/>
              </operator>
              <connect from_port="model" to_op="Apply Model
(2)" to_port="model"/>
              <connect from_port="test set" to_op="Apply
Model (2)" to_port="unlabelled data"/>
```

```
            <connect from_op="Apply Model (2)"
from_port="labelled data" to_op="Performance (2)"
to_port="labelled data"/>
            <connect from_op="Performance (2)"
from_port="performance" to_port="averagable 1"/>
            <portSpacing port="source_model" spacing="0"/>
            <portSpacing port="source_test set"
spacing="0"/>
            <portSpacing port="source_through 1"
spacing="0"/>
            <portSpacing port="sink_averagable 1"
spacing="0"/>
            <portSpacing port="sink_averagable 2"
spacing="0"/>
          </process>
        </operator>
        <connect from_port="example set"
to_op="Validation" to_port="training"/>
        <connect from_op="Validation"
from_port="averagable 1" to_port="performance"/>
        <portSpacing port="source_example set"
spacing="0"/>
        <portSpacing port="source_through 1" spacing="0"/>
        <portSpacing port="sink_performance" spacing="0"/>
      </process>
    </operator>
    <operator activated="true"
class="support_vector_machine" compatibility="7.3.001"
expanded="true" height="124" name="SVM" width="90" x="581"
y="34">
      <parameter key="kernel_type" value="polynomial"/>
      <parameter key="kernel_gamma" value="1.0"/>
      <parameter key="kernel_sigma1" value="1.0"/>
      <parameter key="kernel_sigma2" value="0.0"/>
      <parameter key="kernel_sigma3" value="2.0"/>
      <parameter key="kernel_shift" value="1.0"/>
      <parameter key="kernel_degree" value="3.0"/>
      <parameter key="kernel_a" value="1.0"/>
      <parameter key="kernel_b" value="0.0"/>
```

```
        <parameter key="kernel_cache" value="500"/>
        <parameter key="C" value="0.0"/>
        <parameter key="convergence_epsilon" value="0.001"/>
        <parameter key="max_iterations" value="100000"/>
        <parameter key="scale" value="true"/>
        <parameter key="calculate_weights" value="true"/>
        <parameter key="return_optimization_performance"
value="true"/>
        <parameter key="L_pos" value="1.0"/>
        <parameter key="L_neg" value="1.0"/>
        <parameter key="epsilon" value="0.0"/>
        <parameter key="epsilon_plus" value="0.0"/>
        <parameter key="epsilon_minus" value="0.0"/>
        <parameter key="balance_cost" value="false"/>
        <parameter key="quadratic_loss_pos" value="false"/>
        <parameter key="quadratic_loss_neg" value="false"/>
        <parameter key="estimate_performance"
value="false"/>
      </operator>
      <operator activated="true" class="apply_model"
compatibility="7.1.001" expanded="true" height="82"
name="Apply Model" width="90" x="715" y="34">
        <list key="application_parameters"/>
        <parameter key="create_view" value="false"/>
      </operator>
      <operator activated="true"
class="performance_binominal_classification"
compatibility="7.3.001" expanded="true" height="82"
name="Performance" width="90" x="849" y="34">
        <parameter key="main_criterion" value="first"/>
        <parameter key="accuracy" value="true"/>
        <parameter key="classification_error" value="true"/>
        <parameter key="kappa" value="false"/>
        <parameter key="AUC (optimistic)" value="false"/>
        <parameter key="AUC" value="false"/>
        <parameter key="AUC (pessimistic)" value="false"/>
        <parameter key="precision" value="true"/>
        <parameter key="recall" value="true"/>
        <parameter key="lift" value="false"/>
```

```
        <parameter key="fallout" value="true"/>
        <parameter key="f_measure" value="true"/>
        <parameter key="false_positive" value="true"/>
        <parameter key="false_negative" value="true"/>
        <parameter key="true_positive" value="true"/>
        <parameter key="true_negative" value="true"/>
        <parameter key="sensitivity" value="true"/>
        <parameter key="specificity" value="true"/>
        <parameter key="youden" value="true"/>
        <parameter key="positive_predictive_value"
value="true"/>
        <parameter key="negative_predictive_value"
value="true"/>
        <parameter key="psep" value="true"/>
        <parameter key="skip_undefined_labels"
value="true"/>
        <parameter key="use_example_weights" value="true"/>
      </operator>
      <connect from_op="Retrieve Pure_Training"
from_port="output" to_op="Nominal to Binominal"
to_port="example set input"/>
      <connect from_op="Nominal to Binominal"
from_port="example set output" to_op="LDA" to_port="training
set"/>
      <connect from_op="LDA" from_port="exampleSet"
to_op="Optimize Selection (Evolutionary)" to_port="example
set in"/>
      <connect from_op="Optimize Selection (Evolutionary)"
from_port="example set out" to_op="SVM" to_port="training
set"/>
      <connect from_op="Optimize Selection (Evolutionary)"
from_port="weights" to_port="result 3"/>
      <connect from_op="SVM" from_port="model" to_op="Apply
Model" to_port="model"/>
      <connect from_op="SVM" from_port="exampleSet"
to_op="Apply Model" to_port="unlabelled data"/>
      <connect from_op="Apply Model" from_port="labelled
data" to_op="Performance" to_port="labelled data"/>
```

```
      <connect from_op="Performance" from_port="performance"
to_port="result 1"/>
      <connect from_op="Performance" from_port="example set"
to_port="result 2"/>
      <portSpacing port="source_input 1" spacing="0"/>
      <portSpacing port="sink_result 1" spacing="0"/>
      <portSpacing port="sink_result 2" spacing="0"/>
      <portSpacing port="sink_result 3" spacing="0"/>
      <portSpacing port="sink_result 4" spacing="0"/>
    </process>
  </operator>
</process>
```

APPENDIX F

Optimized Parameters Ensemble SVM Code:

```
<?xml version="1.0" encoding="UTF-8"?><process
version="7.3.001">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process"
compatibility="6.0.002" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve"
compatibility="7.3.001" expanded="true" height="68"
name="Retrieve Pure_Training" width="90" x="112" y="34">
        <parameter key="repository_entry"
value="../data/Pure_Training"/>
      </operator>
      <operator activated="true"
class="replace_missing_values" compatibility="7.3.001"
expanded="true" height="103" name="Replace Missing Values"
width="90" x="313" y="34">
        <parameter key="return_preprocessing_model"
value="false"/>
        <parameter key="create_view" value="false"/>
        <parameter key="attribute_filter_type" value="all"/>
        <parameter key="attribute" value=""/>
        <parameter key="attributes" value=""/>
        <parameter key="use_except_expression"
value="false"/>
        <parameter key="value_type"
value="attribute_value"/>
```

```
        <parameter key="use_value_type_exception"
value="false"/>
        <parameter key="except_value_type" value="time"/>
        <parameter key="block_type"
value="attribute_block"/>
        <parameter key="use_block_type_exception"
value="false"/>
        <parameter key="except_block_type"
value="value_matrix_row_start"/>
        <parameter key="invert_selection" value="false"/>
        <parameter key="include_special_attributes"
value="false"/>
        <parameter key="default" value="average"/>
        <list key="columns"/>
      </operator>
      <operator activated="true" class="split_validation"
compatibility="7.3.001" expanded="true" height="124"
name="Validation" width="90" x="447" y="34">
        <parameter key="create_complete_model"
value="false"/>
        <parameter key="split" value="relative"/>
        <parameter key="split_ratio" value="0.7"/>
        <parameter key="training_set_size" value="100"/>
        <parameter key="test_set_size" value="-1"/>
        <parameter key="sampling_type" value="shuffled
sampling"/>
        <parameter key="use_local_random_seed"
value="false"/>
        <parameter key="local_random_seed" value="1992"/>
        <process expanded="true">
          <operator activated="true" class="bagging"
compatibility="7.3.001" expanded="true" height="82"
name="Bagging" width="90" x="191" y="34">
            <parameter key="sample_ratio" value="0.9"/>
            <parameter key="iterations" value="10"/>
            <parameter key="average_confidences"
value="true"/>
            <parameter key="use_local_random_seed"
value="false"/>
```

```
            <parameter key="local_random_seed"
value="1992"/>
            <process expanded="true">
              <operator activated="true"
class="support_vector_machine_evolutionary"
compatibility="7.3.001" expanded="true" height="82"
name="SVM" width="90" x="431" y="34">
                <parameter key="kernel_type"
value="radial"/>
                <parameter key="kernel_gamma" value="1.0"/>
                <parameter key="kernel_sigma1" value="1.0"/>
                <parameter key="kernel_sigma2" value="0.0"/>
                <parameter key="kernel_sigma3" value="2.0"/>
                <parameter key="kernel_degree" value="3.0"/>
                <parameter key="kernel_shift" value="1.0"/>
                <parameter key="kernel_a" value="1.0"/>
                <parameter key="kernel_b" value="0.0"/>
                <parameter key="C" value="0.0"/>
                <parameter key="epsilon" value="0.1"/>
                <parameter key="start_population_type"
value="random"/>
                <parameter key="max_generations"
value="10000"/>
                <parameter
key="generations_without_improval" value="30"/>
                <parameter key="population_size" value="1"/>
                <parameter key="tournament_fraction"
value="0.75"/>
                <parameter key="keep_best" value="true"/>
                <parameter key="mutation_type"
value="gaussian_mutation"/>
                <parameter key="selection_type"
value="tournament"/>
                <parameter key="crossover_prob"
value="1.0"/>
                <parameter key="use_local_random_seed"
value="false"/>
                <parameter key="local_random_seed"
value="1992"/>
```

```
                <parameter key="hold_out_set_ratio"
value="0.0"/>
                <parameter key="show_convergence_plot"
value="false"/>
                <parameter key="show_population_plot"
value="false"/>
                <parameter
key="return_optimization_performance" value="false"/>
              </operator>
              <connect from_port="training set" to_op="SVM"
to_port="training set"/>
              <connect from_op="SVM" from_port="model"
to_port="model"/>
              <portSpacing port="source_training set"
spacing="0"/>
              <portSpacing port="sink_model" spacing="0"/>
            </process>
          </operator>
          <connect from_port="training" to_op="Bagging"
to_port="training set"/>
          <connect from_op="Bagging" from_port="model"
to_port="model"/>
          <portSpacing port="source_training" spacing="0"/>
          <portSpacing port="sink_model" spacing="0"/>
          <portSpacing port="sink_through 1" spacing="0"/>
        </process>
        <process expanded="true">
          <operator activated="true" class="apply_model"
compatibility="7.1.001" expanded="true" height="82"
name="Apply Model" width="90" x="45" y="30">
            <list key="application_parameters"/>
            <parameter key="create_view" value="false"/>
          </operator>
          <operator activated="true"
class="performance_classification" compatibility="7.3.001"
expanded="true" height="82" name="Performance" width="90"
x="179" y="34">
            <parameter key="main_criterion" value="first"/>
            <parameter key="accuracy" value="true"/>
```

```
            <parameter key="classification_error"
value="false"/>
            <parameter key="kappa" value="false"/>
            <parameter key="weighted_mean_recall"
value="false"/>
            <parameter key="weighted_mean_precision"
value="false"/>
            <parameter key="spearman_rho" value="false"/>
            <parameter key="kendall_tau" value="false"/>
            <parameter key="absolute_error" value="false"/>
            <parameter key="relative_error" value="false"/>
            <parameter key="relative_error_lenient"
value="false"/>
            <parameter key="relative_error_strict"
value="false"/>
            <parameter key="normalized_absolute_error"
value="false"/>
            <parameter key="root_mean_squared_error"
value="false"/>
            <parameter key="root_relative_squared_error"
value="false"/>
            <parameter key="squared_error" value="false"/>
            <parameter key="correlation" value="false"/>
            <parameter key="squared_correlation"
value="false"/>
            <parameter key="cross-entropy" value="false"/>
            <parameter key="margin" value="false"/>
            <parameter key="soft_margin_loss"
value="false"/>
            <parameter key="logistic_loss" value="false"/>
            <parameter key="skip_undefined_labels"
value="true"/>
            <parameter key="use_example_weights"
value="true"/>
            <list key="class_weights"/>
          </operator>
          <connect from_port="model" to_op="Apply Model"
to_port="model"/>
```

```
          <connect from_port="test set" to_op="Apply Model"
to_port="unlabelled data"/>
          <connect from_op="Apply Model" from_port="labelled
data" to_op="Performance" to_port="labelled data"/>
          <connect from_op="Performance"
from_port="performance" to_port="averagable 1"/>
          <portSpacing port="source_model" spacing="0"/>
          <portSpacing port="source_test set" spacing="0"/>
          <portSpacing port="source_through 1" spacing="0"/>
          <portSpacing port="sink_averagable 1"
spacing="0"/>
          <portSpacing port="sink_averagable 2"
spacing="0"/>
        </process>
      </operator>
      <connect from_op="Retrieve Pure_Training"
from_port="output" to_op="Replace Missing Values"
to_port="example set input"/>
      <connect from_op="Replace Missing Values"
from_port="example set output" to_op="Validation"
to_port="training"/>
      <connect from_op="Validation" from_port="model"
to_port="result 2"/>
      <connect from_op="Validation" from_port="averagable 1"
to_port="result 1"/>
      <portSpacing port="source_input 1" spacing="0"/>
      <portSpacing port="sink_result 1" spacing="18"/>
      <portSpacing port="sink_result 2" spacing="0"/>
      <portSpacing port="sink_result 3" spacing="0"/>
    </process>
  </operator>
</process>
```

APPENDIX G

G1. Islamic credit risk analysis using: Information gain and logistic regression code:

```
<?xml version="1.0" encoding="UTF-8"?><process
version="7.3.001">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process"
compatibility="6.0.002" expanded="true" name="Process">
    <parameter key="logverbosity" value="init"/>
    <parameter key="random_seed" value="2001"/>
    <parameter key="send_mail" value="never"/>
    <parameter key="notification_email" value=""/>
    <parameter key="process_duration_for_mail" value="30"/>
    <parameter key="encoding" value="SYSTEM"/>
    <process expanded="true">
      <operator activated="true" class="retrieve"
compatibility="7.3.001" expanded="true" height="68"
name="Retrieve CreditRiskDS" width="90" x="45" y="34">
        <parameter key="repository_entry"
value="../data/CreditRiskDS"/>
      </operator>
      <operator activated="true"
class="weight_by_information_gain" compatibility="7.3.001"
expanded="true" height="82" name="Weight by Information
Gain" width="90" x="179" y="34">
        <parameter key="normalize_weights" value="true"/>
        <parameter key="sort_weights" value="true"/>
        <parameter key="sort_direction" value="descending"/>
      </operator>
      <operator activated="true"
class="h2o:logistic_regression" compatibility="7.3.000"
expanded="true" height="103" name="Logistic Regression"
width="90" x="380" y="34">
        <parameter key="solver" value="IRLSM"/>
        <parameter key="reproducible" value="true"/>
```

```
        <parameter key="maximum_number_of_threads"
value="4"/>
        <parameter key="use_regularization" value="false"/>
        <parameter key="lambda" value="0.0"/>
        <parameter key="lambda_search" value="false"/>
        <parameter key="number_of_lambdas" value="0"/>
        <parameter key="lambda_min_ratio" value="0.0"/>
        <parameter key="early_stopping" value="true"/>
        <parameter key="stopping_rounds" value="3"/>
        <parameter key="stopping_tolerance" value="0.001"/>
        <parameter key="standardize" value="true"/>
        <parameter key="non-negative_coefficients"
value="false"/>
        <parameter key="compute_p-values" value="true"/>
        <parameter key="remove_collinear_columns"
value="true"/>
        <parameter key="add_intercept" value="true"/>
        <parameter key="missing_values_handling"
value="Skip"/>
        <parameter key="max_iterations" value="0"/>
        <parameter key="max_runtime_seconds" value="0"/>
      </operator>
      <operator activated="true" class="apply_model"
compatibility="7.1.001" expanded="true" height="82"
name="Apply Model" width="90" x="581" y="34">
        <list key="application_parameters"/>
        <parameter key="create_view" value="false"/>
      </operator>
      <operator activated="true"
class="performance_binominal_classification"
compatibility="7.3.001" expanded="true" height="82"
name="Performance" width="90" x="715" y="34">
        <parameter key="main_criterion" value="first"/>
        <parameter key="accuracy" value="true"/>
        <parameter key="classification_error" value="true"/>
        <parameter key="kappa" value="false"/>
        <parameter key="AUC (optimistic)" value="false"/>
        <parameter key="AUC" value="false"/>
        <parameter key="AUC (pessimistic)" value="false"/>
```

```
        <parameter key="precision" value="true"/>
        <parameter key="recall" value="true"/>
        <parameter key="lift" value="false"/>
        <parameter key="fallout" value="true"/>
        <parameter key="f_measure" value="true"/>
        <parameter key="false_positive" value="true"/>
        <parameter key="false_negative" value="true"/>
        <parameter key="true_positive" value="true"/>
        <parameter key="true_negative" value="true"/>
        <parameter key="sensitivity" value="true"/>
        <parameter key="specificity" value="true"/>
        <parameter key="youden" value="true"/>
        <parameter key="positive_predictive_value"
value="true"/>
        <parameter key="negative_predictive_value"
value="true"/>
        <parameter key="psep" value="true"/>
        <parameter key="skip_undefined_labels"
value="true"/>
        <parameter key="use_example_weights" value="true"/>
      </operator>
      <connect from_op="Retrieve CreditRiskDS"
from_port="output" to_op="Weight by Information Gain"
to_port="example set"/>
      <connect from_op="Weight by Information Gain"
from_port="weights" to_port="result 3"/>
      <connect from_op="Weight by Information Gain"
from_port="example set" to_op="Logistic Regression"
to_port="training set"/>
      <connect from_op="Logistic Regression"
from_port="model" to_op="Apply Model" to_port="model"/>
      <connect from_op="Logistic Regression"
from_port="exampleSet" to_op="Apply Model"
to_port="unlabelled data"/>
      <connect from_op="Apply Model" from_port="labelled
data" to_op="Performance" to_port="labelled data"/>
      <connect from_op="Performance" from_port="performance"
to_port="result 1"/>
```

```
      <connect from_op="Performance" from_port="example set"
to_port="result 2"/>
      <portSpacing port="source_input 1" spacing="0"/>
      <portSpacing port="sink_result 1" spacing="0"/>
      <portSpacing port="sink_result 2" spacing="0"/>
      <portSpacing port="sink_result 3" spacing="0"/>
      <portSpacing port="sink_result 4" spacing="0"/>
    </process>
  </operator>
</process>
```

G.2: Sectors

Table G.2: Islamic Finance Sectors

Sector Name (Arabic)	Sector Name (English)
الزراعى	Agricultural
زراعى مطري	Rain Agriculture
زراعى مروي	Irrigated Agriculture
زراعى صيفى	Summer Agriculture
زراعى شتوي	Winter Agriculture
حيوانى	Animal
القطاع العام غبر المالى_الزراعه_تمويل داخلى قصير الاجل	Public Non-Financial Sector : Short term Agriculture
القطاع العام غبر المالى_الزراعه_تمويل داخلى متوسط و طويل الاجل	Public Non-Financial Sector : Mid & long term Agriculture
القطاع العام غبر المالى_الزراعه_تمويل استيراد قصير الاجل	Public Non-Financial Sector : Short term Import
القطاع الخاص و التعاونى غير المالى_الزراعه_تمويل داخلى قصير الاجل	Private Non-Financial Co-Operative Sector : Short term Agriculture
القطاع الخاص و التعاونى غير المالى_الزراعه_تمويل داخلى متوسط و طويل الاجل	Private Non-Financial Co-Operative Sector : Mid-term & Long Term Agriculture

القطاع الخاص و التعاونى غير المالى_الزراعه_تمويل استيراد متوسط و طويل الاجل	Private Non-Financial Co-Operative Sector : Mid-term & Long Term Import
الزراعي	Agricultural finance
القطاع العام غبر المالى_الزراعه_تمويل استيراد متوسط و طويل الاجل	Public Non-Financial Sector : Mid and long term Agriculture
القطاع الخاص و التعاونى غير المالى_الزراعه_تمويل استيراد قصير الاجل	Public Non-Financial Sector : Short term Agriculture
القطاع الخاص و التعاونى غير المالى_الزراعه_تمويل استيراد متوسط و طويل الاجل	Private Non-Financial Co-Operative Sector :Mid-term & Long Term Agriculture
محاصيل	Crops
دواجن	Poultry
تجاري (وارد)	Commercial (Import)
التجاره المحليه	Local Trade
تجاري	Commercial
الصناعى	Industrial
القطاع العام غبر المالى_الصناعه_تمويل داخلى قصير الاجل	Public Non-Financial Sector : Short term Industry
القطاع العام غبر المالى_الصناعه_تمويل داخلى متوسط و طويل الاجل	Public Non-Financial Sector : Mid and long term Industry
القطاع العام غبر المالى_الصناعه_تمويل استيراد قصير الاجل	Public Non-Financial Sector : Short term Import Industry
القطاع العام غبر المالى_الصناعه_تمويل استيراد متوسط و طويل الاجل	Public Non-Financial Sector : Mid and long term Import Industry
القطاع الخاص و التعاونى غير المالى_الصناعه_تمويل داخلى قصير الاجل	Private Non-Financial Co-Operative Sector : Short term Industry
القطاع الخاص و التعاونى غير المالى_الصناعه_تمويل داخلى متوسط و طويل الاجل	Private Non-Financial Co-Operative Sector : Mid-term & Long Term Industry
القطاع الخاص و التعاونى غير المالى_الصناعه_تمويل استيراد قصير الاجل	Private Non-Financial Co-Operative Sector : Short term Import
القطاع الخاص و التعاونى غير المالى_الصناعه_تمويل	Private Non-Financial Co-Operative Sector :

استيراد متوسط و طويل الاجل	Mid-term & Long Term Import Industry
استيراد و تصدير	Export & Import
صادر	Export
القطاع العام غبر المالى_الصادر_تمويل داخلى قصير الاجل	Public Non-Financial Sector : Short term Export
القطاع العام غبر المالى_الصادر_تمويل داخلى متوسط و طويل الاجل	Public Non-Financial Sector : Mid & long term Export
القطاع العام غبر المالى_الصادر_تمويل استيراد قصير الاجل	Public Non-Financial Sector : Short term Export
القطاع العام غبر المالى_الصادر_تمويل استيراد متوسط و طويل الاجل	Public Non-Financial Sector : Mid & long term Export Import
القطاع الخاص و التعاونى غير المالى_الصادر_تمويل داخلى قصير الاجل	Private Non-Financial Co-Operative Sector : Short term Export
القطاع الخاص و التعاونى غير المالى_الصادر_تمويل داخلى متوسط و طويل الاجل	Private Non-Financial Co-Operative Sector : Mid & Long Internal Finance Export
القطاع الخاص و التعاونى غير المالى_الصادر_تمويل استيراد قصير الاجل	Private Non-Financial Co-Operative Sector : Short term Import
الاستيراد لاغراض اخري	Import for Other Purposes
الاستيراد لاغراض اخري_تمويل داخلى قصير الاجل	Import for Other Purposes : Short Term
الاستيراد لاغراض اخري_تمويل داخلى متوسط و طويل الاجل	Import for Other Purposes : Short & Mid Term
القطاع الخاص و التعاونى غير المالى_الصادر_تمويل استيراد متوسط و طويل الاجل	Private Non-Financial Co-Operative Sector Export : Mid & Long Import Finance
نقل	Transportation
نقل و تخزين	Transportation and Storage
القطاع العام غبر المالى_التخزين_تمويل داخلى قصير الاجل	Public Non-Financial Sector Transportation and Storage: Short Term
القطاع العام غبر المالى_التخزين_تمويل داخلى متوسط و طويل الاجل	Public Non-Financial Sector Transportation and Storage:Mid & Long Term Internal Finance

القطاع العام غبر المالى_التخزين_تمويل استيراد قصير الاجل	Public Non-Financial Sector Transportation and Storage: Short Term Import
القطاع العام غبر المالى_التخزين_تمويل استيراد متوسط و طويل الاجل	Public Non-Financial Sector Transportation and Storage: Mid & Short Term Import
القطاع العام غبر المالى_النقل_تمويل داخلى قصير الاجل	Public Non-Financial Sector Transportation: Short Term Local Finance
القطاع العام غبر المالى_النقل_تمويل داخلى متوسط و طويل الاجل	Public Non-Financial Sector Transportation: Mid & Long Term Internal Finance
القطاع العام غبر المالى_النقل_تمويل استيراد قصير الاجل	Public Non-Financial Sector Transportation: Short Term Import
القطاع العام غبر المالى_النقل_تمويل استيراد متوسط و طويل الاجل	Public Non-Financial Sector Transportation: Mid & Long Term Import
القطاع الخاص و التعاونى غير المالى_التخزين_تمويل داخلى قصير الاجل	Private Non-Financial Co-Operative Sector Storage : Short term Local Finance
القطاع الخاص و التعاونى غير المالى_التخزين_تمويل داخلى متوسط و طويل الاجل	Private Non-Financial Co-Operative Sector Storage : Mid & Long Term Local Finance
اسر منتجه	Productive Families
تشغيل	Operative
تنمية	Development
بترول	Petrol
بترول و تعدين	Petrol and Mining
تجاري (تعدين)	Commercial (Mining)
طاقة	Energy
مهنى و حرفى	Professional
خدمى	Services
مصارف و صرافات	Banks and Exchanges Offices
اوراق ماليه	Stocks
اخري	Others
القطاعات غير المتفرعه	Non-Branched Sectors
المؤسسات المصرفيه	Financial Institutions

المؤسسات الماليه غير المصرفيه	Non-Banks financial Institutions
الصحة	Health
القطاع الخاص و التعاوني غير المالي	Private Non-Financial Co-Operative Sector
الحكومات المركزيه	Central Governments
الحكومات الولائيه	States Governments
القطاع الدبلوماسي	Diplomatic Sector
القطاع العام غير المالي	Public Non-Financial Sector
البنية التحتية والمنشأت	Infrastructure and Building
عقارات	Real states
مبانى	Buildings
طوق و جسور	Roads and Bridges

G.3: Finance Modes

Table G.3: Islamic Finance Modes

Finance Mode (Arabic)	Finance Mode (English)
مرابحة	Murabaha
مشاركة	Musharaka
مضاربة	Mudaraba
سلم	Salam
إستصناع	Istisnaa
مقاولة	Mugawala
مزارعة	Muzaraa
أخرى	Other
إعتماد مؤجل	Deferred Credit
إطلاع	Sight
خطابات ضمان	Letters of Guarantee
إجارة	Igara
بيع بالتقسيط	Installments
قرض حسن	Qard hasan
محافظ	Portfolios
خطاب ضمان-حسن التنفيذ	LG-Good Result
خطاب ضمان-مراسلين	LG-Correspondents
خطاب ضمان-مبدئى	LG-Preliminary