



Sudan University of Science & Technology  
Faculty of Computer Science and Information Technology

# **Modeling Crude Oil Price Chaotic Behavior Using Machine Learning**

نمذجة السلوك الفوضوي لأسعار النفط الخام باستخدام آلة التعلم

Submitted for the Degree of Doctor of Philosophy in  
Computer Science

**By**

Lubna Abdelkareim Gabralla

**Supervisor**

Professor Dr. Ajith Abraham

*May 2015*

# Dedication

*By the name of Allah, Most Gracious, Most Merciful*

I dedicate the current thesis:

***To my dear parents***

*They encouraged and prayed for me throughout the time of my research.*

***To my husband: Marwan***

*His presence, patience, understanding and encouragement have helped me struggle through the most difficult times. Without his help and encouragements, I could not have finished this work.*

***To my wonderful children: Ahmed, Aya, Mohammed and Qusai***

*They knew who they are my best buddies and my strengths.*

***To my brothers: Mohammed, Khalid, Tariq, Omar and Hatem And their wives***

*They stood by me through the good times and bad moments, encouraged me to have high expectations and to fight hard to achieve my goal.*

***To the soul of our colleague: Ismail Hamid***

*I deeply regret that he did not live long enough to see his dream come true, and to complete the PhD journey with us. I ask Allah by his mercy, which envelopes all things that forgive to Ismail and grant him Jannah Firdaus.*

# Acknowledgements

First and above all, I praise Allah, the almighty for providing me this opportunity and granting me the capability to proceed successfully. I would therefore like to offer my sincere thanks to several people who in one way or another contributed to the completion of this thesis by their assistance and guidance.

My sincere gratitude goes to my thesis advisor; Professor Ajith Abraham, who over the past three years provided excellent guidance, encouragements, caring, patience, motivation, enthusiasm, immense knowledge, and reassured from time to time. He always made himself available to answer any question I had, to give more advices and comments, which were very helpful in achieving my objectives in research and in my life. It has been an honor to be his Ph.D. student.

Besides my advisor, I would like to express my sincere thanks to my thesis committee members for their critical thoughts, insightful comments and expert advices on my dissertation.

I would like to express my deepest gratitude to the Ph.D. program members in SUST for their efforts to provide and allow us this wonderful opportunity. Therefore, I am particularly indebted to: Prof Izzeldin Mohammed Osman (Academic consultant for vice chancellor Sudan University), Dr. Mohammed Al Hafiz and Dr. Talaat M. Wahby (Deans of the Faculty of Computer Science), Dr. Awad M. Awad elkarim and Dr. Hisham Mansor (Program coordinators) and all the professors who taught me in the first and second semesters .

Special thanks go to my cousin Dr. Abdel Kadir A. Osman (Dean of students in SUST) for his encouragements and support throughout the PhD journey.

Furthermore, I am grateful to a number of people who offered helpful advice and comments: Dr. Rania Jammazi, first teacher for me in the oil field. I will never forget the time that we were together in her home in Tunisia and her friendly assistance with various problems all the time, especially for her help with the data and article. Dr. Eihab Basheer for his helpful comments on all parts of the thesis and for assisting me in many different ways, Dr. Haruna Chiroma and Dr. Hela Mahersia: I greatly appreciate their excellent assistance and their spiritual support for me. Varun Kumar Ojha from VSB-Technical University of Ostrava, Czech Republic for his help and assistance.

I would like to acknowledge all my friends for all their augmented encouragements from time to another special thanks to my close friends: Dr. Reem Ahmed and Dr. Hanan Al-Ghamdi.

Last but not the least, I want to thank Sudan petroleum ministry staff and all those who helped or participated in our questionnaire (For example, and not as a limitation Prof Mustafa Nawari, Mr. Mohammed Osman, Mr. Abdelrazig Abdelrahim, Dr.Elfatih M.Nour , My colleague Mohamed Awad, Dr.Abdul Rhman Al-Afandi, Dr. Hamza Awad, Dr. Fatai Anifowose (K.S.A), Ms. Fathia Kraiem (Tunisia), Dr. Salah Al-Shrhan (Kuwait), Dr. Azah Muda (Malaysia), Prof Adel Alimi (Tunisia), Mr. Ehsan Ul-Haq (U.K), Mr. Badr El-Din Saeed (U.A.E), Mr. Abdulla Al-Ghamdi (K.S.A), Dr.Ibrahim Magboul (Malaysia), and Dr. Tibebe Beshah (Ethiopia).

# Abstract

The oil prices and its future are one of the most prominent topics in our world nowadays. In 2014, oil prices are down sharply, losing over 50% of their value since June peak, when West Texas Intermediate was \$115 a barrel and it is now below \$51 [1]. This has raised a number of questions: What are the factors and reasons that led to this change in prices? What is the effect of this decline on the countries that are highly dependent on oil based economy or importing countries? Falling oil prices have both positive and negative impacts. On one hand for many people, cheaper oil means lower fuel prices and economies of importing countries may rebound a bit if the declining prices are exploited and on the other hand oil-exporting countries are very hardly hit. Similarly rise in prices have opposite effects in both directions. Prediction of oil prices is an important task and difficult challenge at the same time under a number of complex factors that influence the determination of oil prices such as political, economic, climate factors and so on. Experts and analysts indicated the importance of expected future oil prices to support the global economy, companies and institutions to hedge against surprise changes to make sound decisions and building a healthy and successful economy. This volatile behavior is predicted to prompt more new and interesting research challenges.

In the present research, machine learning and computational intelligence approaches are used to predict crude oil prices using direct prediction and combined prediction models.

In this research, before constructing a computational model, several aspects of initial preparation of data were selected, which consists of 14 input as attributes to predict the West Taxes Intermediate (WTI) as output. Normalization, feature selection and data partition are used for preparation of the inputs. Then several direct prediction models that have shown good (a priori) performance on datasets similar to the prediction of crude oil prices were examined. Radial basis function neural network outperformed other methods in obtaining the prediction error. In order to improve the accuracy of the direct models, different combined prediction models were used, which include Meta learning schemes, hybrid and ensemble models. For the generalized ensemble method,

Particle Swarm Optimization method was used to determine the optimal weights and obtained the best results. Volatility Implied Equity Index (VIX), West Texas Intermediate (WTI), New York Harbor conventional gasoline spot prices (GPNY), Exchange rate (ER), and Future contracts 1 (FC1) are the most important factors to determine the crude oil price. The generalized ensemble is a good model to explore and explain crude oil market's rules with 80% and 20% of the data for training and testing. Comparison with different results in the literatures presented further proved the effectiveness and superiority of the generalized ensemble model for the prediction of the WTI crude oil price.

Modern society relies on crude oil dramatically. Crude oil price is affected by many factors and coupled by an international network of thousands of producers, refiners, marketers, traders, and consumers for buying and selling physical volumes of crude oil. So there is also a great need to understand, organize, analyze and explain the behavior of the crude oil price market and different aspects of international crude oil pricing and prediction using novel information enterprise architecture for crude oil pricing and prediction based on Zachman framework were discussed. This novel information architecture leads to a deeper understanding of the comprehensive structure of the crude oil market.

**Keywords:** *Prediction crude oil price, Machine learning, Information enterprise architecture, Zachman framework.*

## المخلص

في الوقت الحاضر تعتبر أسعار النفط الحالية والمستقبلية واحدة من ابرز المواضيع . ففي عام 2014، انخفضت أسعار النفط بشكل حاد، وفقدت أكثر من 50٪ من قيمتها . وصلت الأسعار ذروتها في يونيو بوصول سعر برميل النفط الخام من متوسط غرب تكساس الى 115 دولارا في حين وصلت الآن اسعار النفط تحت 51 دولار . قد أثار هذا التراجع عددا من التساؤلات: من هو المسؤول عن هذا الانخفاض الحاد؟ ما هي العوامل والأسباب التي أدت إلى هذا التغيير في الأسعار؟ ما هو تأثير هذا الانخفاض على البلدان التي تعتمد اعتمادا كبيرا في اقتصادها على النفط و القائمة عليه أو البلدان المستوردة والمستهلكة للنفط ؟ ما هو مصير الاسعار المستقبلية للنفط ؟؟ ان لهبوط أسعار النفط آثار إيجابية وسلبية على حد سواء. فمن جهة انخفاض اسعار النفط لكثير من الناس، يعني انخفاض أسعار البنزين واطاحة الفرصة لاقتصاد البلدان المستوردة ان تنتعش قليلا إذا تم استغلال هذا الانخفاض بشكل جيد وعلى الجانب الآخر نجد ان الدول المصدرة للنفط الاكثر تضررا بهذا الانخفاض. بالمثل ارتفاع اسعار النفط له آثار ايجابية وسلبية لكلا الطرفين ولكن بشكل معاكس للانخفاض. التنبؤ بأسعار النفط مهمة هامة وفي نفس الوقت تحديا صعبا في إطار عدد من العوامل المعقدة التي تؤثر على تحديد أسعار النفط مثل العوامل السياسية والاقتصادية والعوامل المناخية وغيرها. أشار الخبراء والمحللين على أهمية التنبؤ بأسعار النفط المستقبلية لدعم الاقتصاد والشركات والمؤسسات العالمية و للتحوط ضد التغييرات المفاجئة باتخاذ القرارات السليمة وبناء اقتصاد سليم وناجح. ومن المتوقع مع هذا السلوك المضطرب المطالبة بإجراء المزيد من البحوث الجديدة والمثيرة للاهتمام.

في هذا البحث اخترنا عدة خطوات اولية قبل بناء النموذج الحوسبي وذلك لأعداد البيانات والتي تتألف من 14 متغير كمداخلات للتنبؤ بسعر نفط خام متوسط غرب تكساس كمخرج. الخطوات الأولية كانت استخدام التطبيق، واختيار الميزات وتقسيم البيانات لمجموعات تدريب واختبار ثم درسنا عدة نماذج من آلة التعلم على سبيل المثال نماذج التنبؤ المباشرة التي أظهرت نتائج جيدة على بيانات لمشاكل مشابهة لمشكلة التنبؤ بأسعار النفط الخام. الشبكات العصبية الشعاعية الأساسية تفوقت على الأساليب المباشرة الأخرى بحصولها على أدنى نسبة خطأ. من أجل تحسين دقة النماذج المباشرة استخدمنا نماذج التنبؤ المجتمعة والتي تشمل نظريات الميتا، الهجين ونماذج الفرقة. لأسلوب الفرقة المعمم استخدمنا طريقة سرب الجسيمات الأمثل لتحديد الوزن المثالي وللحصول على أفضل النتائج. بنهاية التجارب نستنتج مايلي: أن مؤشر تقلب الأسهم الضمني ، نفط وسط غرب تكساس، الاسعار الفورية لبنزين نيويورك، سعر الصرف والعقود الآجلة هي أهم العوامل لتحديد أسعار النفط الخام . نموذج الفرقة المعمم مع نسبة تدريب 80% واختبار 20% هو نموذج جيد لشرح قواعد سوق النفط الخام . عرضنا مقارنة بين نتائج النموذج المقترح و نتائج مختلفة من الأدب اثبتت المقارنة فعالية وتفوق النموذج .

نعتمد على النفط الخام بشكل كبير في المجتمع الحديث ويتأثر النفط بعوامل كثيرة كما يرتبط بشبكة دولية من الآف المنتجين، المسوقين، التجار، الوسطاء والمستهلكين لشراء وبيع كميات النفط الخام . لذلك كانت هناك حاجة كبيرة لفهم وتنظيم وتحليل وتفسير سلوك اسعار النفط الخام . ناقشنا في هذا البحث جوانب مختلفة للتسعير وللتنبؤ باسعار النفط الخام باستخدام هيكلية المؤسسات المعلوماتية الجديدة والتي تعتمد على استخدام اطار زامان هذه العمارة تؤدي لفهم عميق شامل لهيكل اسواق النفط الخام.

**الكلمات المفتاحية:** التنبؤ باسعار النفط الخام، آلة التعلم، هيكلية المؤسسات المعلوماتية، اطار زامان

# Table of Contents

Dedication.....	I
Acknowledgements.....	III
Abstract.....	IV
List of Figures.....	XI
List of Tables.....	XIII
List of Abbreviations.....	XV
List of Appendices.....	XVII
List of Publications.....	XVIII
CHAPTER ONE.....	1
1. INTRODUCTION.....	1
1.1 Introduction.....	1
1.2 Statement of the Problem and Motivation.....	2
1.2.1 Motivation and Research Significance.....	2
1.2.2 Problem Statement.....	5
1.3 Research Questions.....	5
1.4 Research Objectives.....	6
1.5 Methodology.....	6
1.6 Structure and Organization of the Thesis.....	7
CHAPTER TWO.....	9
2. RESEARCH BACKGROUND.....	9
2.1. Crude Oil Basics.....	9
2.2 Crude Oil Chaotic Behavior.....	11
2.3 Factors Affecting the Prices of Crude Oil.....	13
2.3.1 Fundamental Supply and Demand.....	14
2.3.2 Geopolitical Events.....	14

2.3.3	Seasonal Weather and Natural Disasters .....	15
2.3.4	Speculation and Market Conditions .....	15
2.4	What is Data Mining / Machine Learning? .....	16
2.4.1	Machine Learning Algorithms .....	17
2.5	Zachman Framework .....	22
2.6	Summary .....	24
CHAPTER THREE .....		26
3.	LITERATURE REVIEW .....	26
3.1	Introduction .....	26
3.2	Factors Affecting Crude Oil Price Fluctuation .....	27
3.3	Oil Price Volatility and Prediction Models .....	29
3.3.1	Statistical and Econometric Prediction Models .....	29
3.3.2	Machine Learning Techniques in Crude Oil Price Models.....	31
3.4	Enterprise Architecture Framework.....	37
3.5	An Overview of the limitations of Previous Research Works .....	39
3.6	Summary .....	40
CHAPTER FOUR .....		42
4.	RESEARCH METHODOLOGY .....	42
4.1.	Introduction .....	42
4.2.	Data Set Description .....	46
4.3.	Data Preprocessing.....	48
4.3.1	Feature Selection Methods.....	48
4.3.2	Data Partition .....	51
4.3.3	Normalization.....	52
4.4	Software Tools .....	52
4.5	Machine Learning Algorithms (ML) .....	53
4.5.1	Direct Prediction Models .....	53



4.5.2	Combined Prediction Models.....	55
4.6	Evaluation and Validation Approaches.....	61
4.6.1	Measuring Techniques: .....	62
4.6.2	Questioning Techniques:.....	62
4.7	Summary .....	63
CHAPTER FIVE .....		64
5. DIRECT PREDICTION MODELS.....		64
5.1	Experimental Results .....	64
5.1.1	First Phase Experiments and Results .....	65
5.1.2	Second Phase Experiments and Results.....	71
5.2	A Comparison Analysis of Direct Prediction Models .....	78
5.3	Conclusions .....	79
CHAPTER SIX.....		80
6. COMBINED PREDICTION MODELS .....		80
6.1	Meta Prediction Experiments.....	80
6.2	Hybrid Prediction Models .....	87
6.3	Ensemble Prediction Models .....	90
6.3.1.	The Basic Ensemble Method .....	90
6.3.2.	The Generalized Ensemble Method (GEM) .....	91
6.4	A Comparison Analysis of Combined Prediction Models.....	92
6.4.1	Comparison of the (Ensemble –PSO-ANFIS) Prediction Model Results with Other Machine Learning Approaches .....	93
6.5	Conclusions .....	93
CHAPTER SEVEN .....		95
7. INFORMATION ENTERPRISE ARCHITECTURE FOR CRUDE OIL PRICING AND PREDICTION .....		95
7.1	Information Enterprise Architecture for Crude Oil Pricing and Prediction..	95
7.1.1	Stages of Crude Oil Pricing and Prediction System .....	97

7.2	Zachman Framework for Crude Oil Pricing and Prediction .....	100
7.2.1	Contextual View- Scope/ Planners Perspective .....	101
7.2.2	Conceptual View- Owners Perspective.....	103
7.2.3	System Model: Logical View- Designer’s Perspective .....	107
7.2.4	Technology Model - Builder’s Perspective .....	112
7.2.5	Detailed Representations - Subcontractor’s Perspective .....	115
7.2.6	Functioning System-Real System Crude Oil Prediction and Pricing	117
7.3	Conclusions.....	117
CHAPTER EIGHT .....		118
8.	CONCLUSIONS AND RECOMMENDATIONS .....	118
8.1	Summary .....	118
8.2	Contributions of the Research.....	120
8.3	Future Research.....	120
REFERENCES .....		122
APPENDICES .....		135
Appendix A: Attribute selection methods and their features using WEKA.....		135
Appendix B: Samples of computations made for performance evaluation .....		140
Appendix C: Samples of of FFN and ANFIS code .....		147
Appendix D: Questionnaire item and document analysis.....		151

## List of Figures

<b>Figure 1.1:</b>	World energy consumption by sector, 2012.....	3
<b>Figure 1.2:</b>	World total primary energy supply from 1971 to 2010.....	3
<b>Figure 2.1:</b>	Oil formation.....	9
<b>Figure 2.2:</b>	The crude oil benchmark by density and sulfur content.....	11
<b>Figure 2.3:</b>	The volatility of the crude oil prices between 1970s and 2011.....	13
<b>Figure 2.4:</b>	Crude oil supply and demand (2002-2012).....	14
<b>Figure 2.5:</b>	Connection between two neurons i and j.....	20
<b>Figure 2.6:</b>	The architecture of the ANFIS.....	21
<b>Figure 4.1:</b>	Basic flowchart of the crude oil predicting model and its information enterprise architecture.....	43
<b>Figure 4.2:</b>	Data preparation for crude oil price model.....	44
<b>Figure 4.3:</b>	The proposed crude oil prediction model framework.....	46
<b>Figure 4.4:</b>	SBDS3 using WEKA.....	49
<b>Figure 4.5:</b>	Bagging ensemble methods.....	56
<b>Figure 4.6:</b>	Ensemble method framework.....	57
<b>Figure 4.7:</b>	Stacking structure for prediction crude oil prices.....	58
<b>Figure 5.1:</b>	MAE for six prediction models with 10 sub-data set.....	69
<b>Figure 5.2:</b>	Comparison between training algorithms.....	72
<b>Figure 5.3:</b>	Comparison among 3 type of NNs.....	78
<b>Figure 6.1:</b>	Bagging algorithm with 5 base prediction models.....	83
<b>Figure 6.2:</b>	Comparison between Bagging and Random Subspace using 7 sub datasets and 4 categories of training and testing.....	85
<b>Figure 6.3:</b>	Trapezoidal-shaped membership function for the first Input.....	88
<b>Figure 6.4:</b>	Developed TSK FIS using 3 inputs.....	88
<b>Figure 6.5:</b>	Developed ANFIS structure with 3 inputs.....	88
<b>Figure 7.1:</b>	Stages for crude oil pricing and prediction system.....	96
<b>Figure 7.2:</b>	Activity diagram representing the conceptual-function cell.....	105
<b>Figure 7.3:</b>	UML package representing conceptual-network cell.....	106
<b>Figure 7.4:</b>	Layer architecture representing logical –function cell.....	110
<b>Figure 7.5:</b>	Crude oil pricing and prediction model representing logical-content cell.....	110
<b>Figure 7.6:</b>	Use case model representing logical-people cell.....	111
<b>Figure 7.7:</b>	Deployment diagram representing logical –network.....	111

<b>Figure 7.8:</b>	State diagram representing logical– time cell.....	112
<b>Figure 7.9:</b>	Network architecture describing Technology –network.....	114
<b>Figure 7.10:</b>	Sequence diagram representing logical– function.....	114
<b>Figure A.1:</b>	Represent SBDS <sub>1</sub> using WEKA.....	135
<b>Figure A.2:</b>	Represent SBDS <sub>2</sub> using WEKA.....	135
<b>Figure A.3:</b>	Represent SBDS <sub>3</sub> using WEKA.....	136
<b>Figure A.4:</b>	Represent SBDS <sub>4</sub> using WEKA.....	136
<b>Figure A.5:</b>	Represent SBDS <sub>5</sub> using WEKA.....	137
<b>Figure A.6:</b>	Represent SBDS <sub>6</sub> using WEKA.....	137
<b>Figure A.7:</b>	Represent SBDS <sub>7</sub> using WEKA.....	138
<b>Figure A.8:</b>	Represent SBDS <sub>8</sub> using WEKA.....	138
<b>Figure A.9:</b>	Represent SBDS <sub>9</sub> using WEKA.....	139
<b>Figure A.10:</b>	Represent SBDS <sub>8</sub> using WEKA.....	139

## List of Tables

<b>Table 2.1:</b>	The basic structure for Zachman framework.....	24
<b>Table 3.1:</b>	Recent single CI applications in oil price prediction.....	34
<b>Table 3.2:</b>	Examples of recent hybrid and ensemble models.....	36
<b>Table 4.1:</b>	Attribute selection methods and their features.....	50
<b>Table 4.2:</b>	Training and testing percentages.....	52
<b>Table 4.3:</b>	Example of dataset after normalization process.....	52
<b>Table 4.4:</b>	ANFIS with different type of membership functions.....	59
<b>Table 5.1.a:</b>	MAE for first phase experiment using sub dataset from (SBDS <sub>1</sub> to SBDS <sub>5</sub> )	65
<b>Table 5.1.b:</b>	MAE for first phase experiment using sub dataset from (SBDS <sub>6</sub> to SBDS <sub>10</sub> )	66
<b>Table 5.2.a:</b>	RMSE for first phase experiment using sub dataset from(SBDS <sub>1</sub> to SBDS <sub>5</sub> )	67
<b>Table 5.2.b:</b>	RMSE for first phase experiment using sub dataset from(SBDS <sub>1</sub> to SBDS <sub>5</sub> )	68
<b>Table 5.3:</b>	Time schedule for direct prediction models.....	70
<b>Table 5.4:</b>	Summary of the results for direct prediction models.....	71
<b>Table 5.5:</b>	Performance of FFN.....	73
<b>Table 5.6:</b>	Performance of RCN.....	74
<b>Table 5.7:</b>	Performance of RBF.....	75
<b>Table 5.8:</b>	RMSE for FFN, RCN and RBF.....	76
<b>Table 5.9:</b>	Comparison between NNs based on time.....	77
<b>Table 5.10:</b>	Summary for the results which explain the comparison between NNs.....	79
<b>Table 6.1:</b>	Bagging algorithms with 5 base prediction models using MAE.....	81
<b>Table 6.2:</b>	Bagging algorithms with 5 base prediction models using RMSE.....	82
<b>Table 6.3:</b>	MAE for Random Subspace.....	84
<b>Table 6.4:</b>	RMSE for Random Subspace.....	84
<b>Table 6.5:</b>	MAE for meta learning (Stacking, Voting and Ensemble selection).....	86
<b>Table 6.6:</b>	RMSE for meta learning (Stacking, Voting and Ensemble selection).....	86
<b>Table 6.7:</b>	NNs results with Bagging and Random subspace.....	87
<b>Table 6.8:</b>	ANFIS results (MAE) for 7 sub-data sets.....	89
<b>Table 6.9.:</b>	ANFIS results (RMSE) for 7 sub-datasets.....	89

<b>Table 6.10:</b>	ANFIS training time.....	89
<b>Table 6.11:</b>	MAE for basic ensemble results for neural networks.....	90
<b>Table 6.12:</b>	Ensemble using Average method for Data (A).....	90
<b>Table 6.13:</b>	Ensemble using Average method for Data (B).....	91
<b>Table 6.14:</b>	Ensemble of PSO-ANFIS for Data (A).....	91
<b>Table 6.15:</b>	Ensemble of PSO-ANFIS for Data (A).....	91
<b>Table 6.16:</b>	Comparison among Meta learning models.....	92
<b>Table 6.17:</b>	Performance comparison between Ensemble prediction models.....	92
<b>Table 6.18:</b>	Comparison of models used in the literature to predict WTI crude oil price using the ANFIS-PSO Ensemble.....	93
<b>Table 7.1:</b>	Development of international oil pricing methods.....	98
<b>Table 7.2:</b>	Rows of the crude oil pricing and prediction information architectural.....	100
<b>Table 7.3:</b>	Dimension of the crude oil pricing and prediction information architectural.....	101
<b>Table 7.4:</b>	Contextual view-scope/planners perspective.....	103
<b>Table 7.5:</b>	Conceptual – people using process Vs organization matrix.....	106
<b>Table 7.6:</b>	Conceptual view-enterprise/owners perspective.....	107
<b>Table 7.7:</b>	System model: View-crude oil prediction /pricing designer’s perspective....	109
<b>Table 7.8:</b>	Technology model-Crude oil prediction /pricing builder’s perspective.....	115
<b>Table 7.9:</b>	Detailed representations of crude oil prediction and pricing.....	116
<b>Table 7.10:</b>	Real system crude oil pricing/prediction .....	117
<b>Table B.1:</b>	Feed forward performance for SBDS <sub>2</sub> .....	140
<b>Table B.2:</b>	Radial basis function performance for SBDS <sub>1</sub> .....	142
<b>Table B.3:</b>	Recurrent performance for SBDS <sub>4</sub> .....	143
<b>Table B.4:</b>	ANFIS and RBF performance for all sub datasets.....	145
<b>Table B.5:</b>	RCN and FFN performance for all sub datasets.....	146

## List of Abbreviations

<b>AI</b>	Artificial intelligence
<b>ANFIS</b>	Adaptive Neuro Fuzzy Inference System
<b>ANN</b>	Artificial neural networks
<b>API</b>	American Petroleum Institute
<b>ARFF</b>	Attributes Relation File Format
<b>ARIMA</b>	Auto Regressive Integrated Moving Average
<b>BFG-QN</b>	BFGS quasi-Newton
<b>BPNN</b>	Back-propagation neural network
<b>BR</b>	Bayesian regularization
<b>CFS</b>	Correlation based Feature Selection
<b>CFTC</b>	Commodity Futures Trading Commission
<b>CLI</b>	Command-Line Interface
<b>DODAF</b>	Department of Defense Architecture Framework
<b>EAF</b>	Enterprise Architecture Framework
<b>ECB</b>	European Central Bank
<b>ECM</b>	Error correction models
<b>EIA</b>	Energy Information Administration
<b>EMD</b>	Empirical mode decomposition
<b>ER</b>	Exchange rate
<b>ES</b>	Expert System
<b>FC1</b>	Future contracts 1
<b>FC2</b>	Future contracts 2
<b>FC3</b>	Future contracts 3
<b>FC4</b>	Future contracts 4
<b>FEAF</b>	Federal Enterprise Architecture Framework
<b>FFN</b>	Feed forward Neural Networks
<b>FFR</b>	Federal Fund rate
<b>FIS</b>	Fuzzy inference system
<b>GARCH</b>	Generalized Autoregressive Conditional Heteroskedastic
<b>GEM</b>	General Ensemble Model
<b>GP</b>	Gold prices
<b>GPMGA</b>	Generalized Pattern Matching Genetic Algorithm
<b>GPNY</b>	New York Harbor conventional gasoline spot prices
<b>GPUS</b>	US Gulf Coast conventional gasoline spot prices
<b>GRNN</b>	General Regression Neural network
<b>HKM</b>	Hierarchical multiple kernel machine
<b>HP</b>	New York Harbor No. 2 heating oil spot price
<b>IBL</b>	Instance-based learning
<b>IT</b>	Information Technology
<b>LinR</b>	Linear regression
<b>LM</b>	Levenberg – Marquardt
<b>LSM</b>	Least Square Methods

<b>MAE</b>	Mean Absolute Error
<b>MAPE</b>	Mean absolute percentage error
<b>ML</b>	Machine learning
<b>MOC</b>	Market on Close
<b>NNs</b>	Neural Networks
<b>NYMEX</b>	New York Mercantile Exchange's
<b>OCED</b>	Organization for Economic Cooperation and Development
<b>OPEC</b>	Organization of Petroleum Exporting Countries
<b>OSVM</b>	Orthogonal Wavelet Support Vector Machine
<b>PCA</b>	Principal Component Analysis
<b>PRAs</b>	Pricing Reporting Agencies
<b>PSO</b>	Particle Swarm Optimization
<b>RBF</b>	Radial Basis Function
<b>RCN</b>	Recurrent Neural Network
<b>REPtree</b>	Reduced Error Pruning Tree
<b>RMSE</b>	Root Mean Squared Error
<b>RSM</b>	The Random Subspace Method
<b>RSTM</b>	Rough-Set-Refined Text Mining
<b>RWM</b>	Random walk model
<b>SPX</b>	The regional Standard and Poor's equity index
<b>SVM</b>	Support Vector Machines
<b>SVMR</b>	Support Vector Machine Regression
<b>SVR</b>	Support Vector Regression
<b>TEAF</b>	Treasury Enterprise Architecture Framework
<b>TOGAF</b>	The Open Group Architectural Framework
<b>TSK</b>	Takagi, Sugeno and Kang
<b>VAR</b>	Vector auto-regression
<b>VECM</b>	Vector error correction
<b>VIX</b>	Volatility Implied Equity Index
<b>WTI</b>	West Texas Intermediate
<b>ZF</b>	Zachman Framework



## List of Appendices

<b>Appendix A:</b>	Attribute selection methods and their features using WEKA.....	132
<b>Appendix B:</b>	Samples of computations made for performance evaluation .....	137
<b>Appendix C:</b>	Samples of FFN and ANFIS code.....	144
<b>Appendix D:</b>	Questionnaire item and document analysis.....	148

## List of Publications

### Journal Publications

- 1- GABRALLA, L. A. & ABRAHAM, A. 2013. Computational Modeling of Crude Oil Price Forecasting: A Review of Two Decades of Research. *International Journal of Computer Information Systems and Industrial Management Applications.* , 5 729-740.

---

- 2- Gabralla, L. A. & Abraham,A. 2014 Comparison of Soft Computing Approaches for Prediction of Crude Oil Price. *Journal of Network and Innovative Computing* 2160-2174 , 2 pp. 318-330.
- 3- Gabralla, L. A. & Abraham, A. 2015.Comparison of Hybrid Intelligent Approaches for Prediction of Crude Oil Price. *International Journal of Computer Information Systems and Industrial Management Applications.* 2150-7988, pp. 053-065

### Conference Publications

- 1- GABRALLA, L. A., JAMMAZI, R. & ABRAHAM, A. Oil price prediction using ensemble machine learning.2013 International Conference on Computing,Computing, Electrical and Electronics Engineering (ICCEEE), Khartoum, Sudan. IEEE, 674-679,2013
- 2- GABRALLA, L. A. & ABRAHAM, A. Prediction of Oil Prices Using Bagging and Random Subspace. Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA), Ostrava, Czech Republic, Springer, 343-354,2014.
- 3- GABRALLA, L. A. & ABRAHAM, A. Hybrid soft computing methods for prediction of oil prices. 6th International Conference on Soft Computing and Pattern Recognition (SoCPaR), Tunis,Tunisia,. IEEE, 140-144,2014 .
- 4- GABRALLA, L. A., MAHERSIA, H. & ABRAHAM, A. Ensemble Neurocomputing Based Oil Price Prediction. Afro-European Conference for Industrial Advancement, Addis Ababa, Ethiopia Springer, 293-302, 2015.
- 5- GABRALLA, L. A., WAHBY, T. M., OJHA, V. K. & ABRAHAM, A. 2015 b. Ensemble of Adaptive Neuro-Fuzzy Inference System Using Particle Swarm Optimization for Prediction of Crude Oil Prices. International Conference on Hybrid Intelligent Systems (HIS). Kuwait, Kuwait: IEEE

# CHAPTER ONE

## 1. INTRODUCTION

### 1.1 Introduction

The industrial revolution began by using traditional sources with coal combustion and hydropower as main sources of electricity, then later began to use oil and nuclear energy as a major participant of energy [2]. Oil is a wealth depletable and is distributed randomly all over the earth [3]. It is an important source of energy and represents the indispensable raw material as a major component in many manufacturing processes and transportation fuel. It is used as a fuel for cars, airplanes, factories and agricultural equipment, trucks, commercial and military ships and electric power generation for homes, workplaces and other places.

Oil prices have undergone many changes and instabilities over the years. It was known as oil shocks, and the first shock was in the October war in 1973, where the price rose from 2.29\$ to 10.73\$ for a barrel until 1974, and these prices continued to rise, and even achieved strong jump to 32.51\$ per barrel in 1981, this what is known as the second oil shock [4].

Despite all these changes and challenges the Organization of Petroleum Exporting (OPEC)<sup>1</sup> was no longer the only dominant player in oil prices in the global markets, where the intertwined political factors [5], with economic reasons, such as the exchange rates of currencies [6], demand for international oil [7], and other factors like climatic factors [8] played an important role.

Crude oil price prediction is a challenging task due to its complex nonlinear and chaotic behavior. During the last couple of decades, both academicians and practitioners have devoted proactive knowledge to address this issue. A strand of them has focused on some key factors that may influence the crude oil price prediction accuracy [9-11],

---

<sup>1</sup> The Organization of the Petroleum Exporting Countries (OPEC) was founded in Baghdad, Iraq, with the signing of an agreement in September 1960 by five countries namely Islamic Republic of Iran, Iraq, Kuwait, Saudi Arabia and Venezuela. They were to become the founder Members of the Organization. These countries were later joined by Qatar (1961), Libya (1962), the United Arab Emirates (1967), Algeria (1969), Nigeria (1971), Ecuador (1973) and Angola (2007). Currently, the Organization has a total of 12 Member Countries

while others concentrated on designing models that will assist to predict crude oil prices with near accurate results [12-14]. To support the global economy, companies and institutions to hedge against surprise changes to make sound decisions and building a healthy and successful economy.

In the past decades, traditional statistical and econometric techniques [15] [16] have been widely applied to crude oil price prediction [17]. However, several experiments proved that the prediction performance might be very poor if one continued using these traditional statistical and econometric models [18].

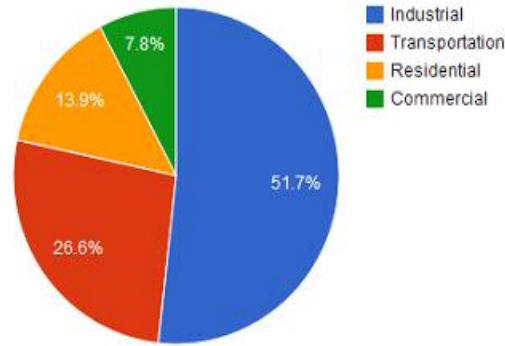
New single techniques such as artificial neural networks [19], genetic algorithm [20] and support vector machine [21] have emerged to remedy this inefficiency. However, experiments and evidence [22] illustrate that the hybrid techniques are superior to the single techniques.

This thesis is focused on the concepts of hybrid data mining models for the prediction crude oil prices and the development of information enterprise architecture to serve the petroleum industry. This Chapter discusses the problem statement and motivation, research questions, objectives, and methodology of the research.

## **1.2 Statement of the Problem and Motivation**

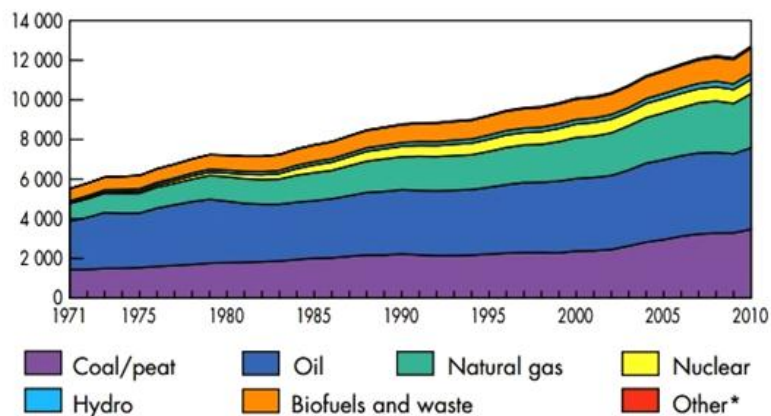
### **1.2.1 Motivation and Research Significance**

Without any doubt, the physical production of energy is the basis of the global economy. Development of the economy depends on the different resources of energy that even most of economic sectors such as commercial, industrial and transportation are impossible to operate without energy. Figure 1.1 shows the proportion of energy use in a number of sectors [23].



**Figure 1.1** World energy consumption by sector, 2012 [23].

Among the different energy sources, oil plays a significant role to become the most important source of energy. It represents the largest proportion of the world's energy consumption compared to other source. According to the Energy Information Administration (EIA), in 2012 the world currently consumes 85.64 million barrels of crude oil daily [12]. Oil plays a vital role in the national and international economies as the backbone and the source of indispensable raw inputs for numerous industries and represents a major component in many manufacturing processes such as plastics, and chemicals. More than half of the world's oil are used as fuel for aircraft, engines and cars. Figure 1.2 shows the oil supply compared to the other sources of energy [24].



**Figure 1.2** World total primary energy supply from 1971 to 2010 [24]

Oil price is suffering from high volatility and fluctuations. In the global market, it is the most active and heavily traded commodity. This volatility is up to approximately 25% per annum [25]. This rate cannot be ignored for its influence in the world economy,

particularly in developing countries. Sharp oil price movements, dramatic uncertainty for the global economy and trends in changing oil prices have an impact on world politics, economy, military and all sectors of society. Therefore, understanding oil price evolution, follow-up and monitoring the prices and forecasting of its future price movement are very important.

For owners of the oil industry and investors there is a need for determining and reducing risk and help producers to make strategic planning, to managing their oil supply and their forward contracts for oil trades. In addition to the need for development and transporting them, it represents an integrated part of the decision-making process for the production and export. In this context, predictability of oil prices is a crucial input into the policymaking process. For example, the European Central Bank (ECB) uses oil futures prices to construct a proxy of inflation and output-gap forecast that guides monetary policy [26].

From the Government's side, the issue of predicting oil prices in the short term and long term has a significant impact on the public policy of the state and national decision-making and to build a local budget and grow their economy. Narayan et al. [27] provided evidence that the nominal oil price predicts economic growth for 37 developing and developed countries. According to the volatility of oil price, Government policy-makers can use policy tools to adjust the stock market, reduce risk of financial market and reduce the probability of extreme risk, reduction of investment [28] and recession [29] among others.

For researchers, academics and economics strategists, perspective of oil prediction represents a prominent role and deeper understanding of the issues in the economy, the financial theories, market hypotheses and pricing of consumer goods to the citizens. [30] provided evidence of co-integration between oil and commodity prices such as wheat, corn and soybean.

## 1.2.2 Problem Statement

In the last several years crude oil prices have presented large variations and appear highly nonlinear and even chaotic, which makes it rather difficult to forecast the future oil prices, which not only directly affect global economic activities, also bring risk to oil related enterprises [31], [32] noted in terms of statistical regularities, that change in the real price of oil have historically tended to be (1) permanent, (2) difficult to predict, and (3) governed by very different regimes at different points in time. Therefore, understanding evolution of oil prices and its forecasting are very important for the oil industry and investors in determining and reducing the risk and helping producers to make strategic planning in managing their oil supply and their forward contracts for oil trades. According to the above reasons, there is a great need for oil price volatility measuring and modeling of oil price chaotic behavior.

At the moment, there is no single information architecture, which explains how the information flows from oil price movements and prediction results could be useful to consumers, policy makers, and entire nations, etc.

## 1.3 Research Questions

The main question behind this research is how should crude oil pricing and prediction functions and tasks be organized under a common architecture? There are also other important questions behind this research such as:

- Q (1):** How to design a model-using machine learning that can predict crude oil prices accurately?
- Q (2):** Which set of features can better describe the performance?
- Q (3):** Can we achieve comparable or relatively higher prediction performances by introducing time lags?
- Q (4):** What is the best percentage for training data and testing data?

## 1.4 Research Objectives

The main aim of this study is to design suitable information architecture so that the results could be useful for the users, policy makers and other nations, etc.

To investigate the prediction of oil prices through the application of hybrid model using machine learning tools and perform comparative analysis to find the best model.

As additional objectives, on the one hand the behavior of oil prices will be examined and on the other hand the fundamental factors contributing to their variation.

## 1.5 Methodology

Most of the studies in the literature focused on constructing a new model using a fixed percentage of training and testing. For example [33] used the size ratio of training to testing sets of 4:1, and [34] partitioned data into 70% training, 15% validation and 15% testing. On the other hand, most researchers were interested in using one input for testing [35] [36] despite the large number of factors influencing oil prices. So in this work a variety of the training and testing percentages with a set of different inputs using several kinds of attributes selection algorithms were provided to get high accuracy for the model.

The hybridization of the artificial intelligence techniques can provide solutions to the complex, nonlinear, and volatile crude oil price prediction. Neuro-fuzzy approach refers to combinations of artificial neural network learning and fuzzy logic. Neuro-fuzzy incorporates the human-like reasoning style of fuzzy systems through the use of fuzzy sets and a linguistic model consisting of a set of *if-then* fuzzy rules. The main strength of neuro-fuzzy systems is that they are universal approximates with the ability to solicit interpretable *if-then* rules and accuracy [37] .

Moreover ensemble methods are one of the latest techniques that promises results more effective. The ensemble method depends on the behavior that a collection of predictor such as machine learning algorithms (neural network, support vector machine, decision trees and so on) can do better than the individual approaches.



Predictors are combined using a weighted average method. However, finding the optimal values of weight is not an easy task. A particle swarm optimization (**PSO**) method was used to determine the optimal weights. Performance evaluation of different models are compared using several evaluation criteria.

Finally, information enterprise architecture for crude oil pricing and prediction based on Zachman framework is presented. The framework helps to explain how to use a system in a useful way, understand the concept of the crude oil price under the umbrella of complicated factors, organizing of varied processes and follow up the overlap between them and to explain the comprehensive structure of the crude oil market and constructing infrastructure for information technology in this field.

## **1.6 Structure and Organization of the Thesis**

The rest of this thesis is organized as follows:

### **Chapter Two: Research Background**

The Chapter provides a brief research background includes: crude oil basics, crude oil chaotic behavior, factors affecting the prices of crude oil and the basic concepts of data mining /machine learning and Zachman framework. This will helps in the understanding of the circumstances and conditions surrounding of the crude oil prediction.

### **Chapter Three: Literature Review**

It describes an exhaustive overview of the existing literature of the applications for predicting crude oil prices with a focus on computational intelligence techniques.

### **Chapter Four: Research Methodology**

This chapter answers the questions: What are the methodologies used in our research? This chapter is considered as a backbone of the thesis.

## **Chapter Five: Direct Prediction Models**

To guarantee building a successful machine-learning model in predicting crude oil prices, practical steps for selecting best learning algorithm by running it over our data were applied. In this Chapter, the prediction process and analyzing the single prediction models were modeled, Furthermore, the comparison of these algorithms is presented based on a root mean squared error (RMSE) and mean absolute error (MAE) to find out the best suitable approaches.

## **Chapter Six: Combined Prediction Models**

In order to improve the results of direct models that was explored in the Chapter five, several types of combined models are illustrated in this Chapter. The Chapter starts with Meta prediction followed by hybrid prediction and finally Ensemble model. Various comparisons are done between different combined models and finally with the results of the final model.

## **Chapter Seven: Development of an Information Enterprise Architecture for Crude Oil Pricing and Prediction**

The Chapter provides new information enterprise architecture for crude oil pricing and prediction.

## **Chapter Eight: Conclusion and Recommendation**

It presents the summary, contributions, and conclusion of the research and finally, identified future research directions.

# CHAPTER TWO

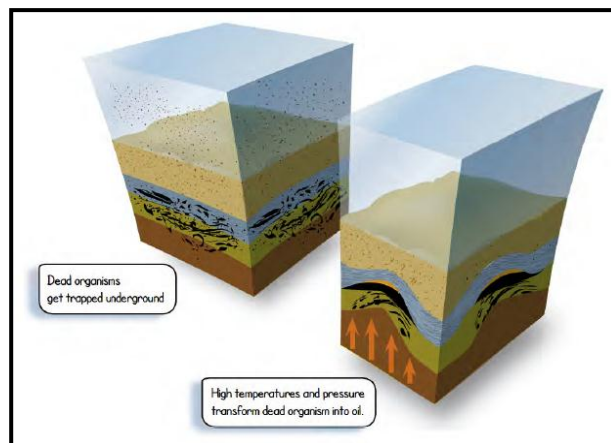
## 2. RESEARCH BACKGROUND

### Overview

This Chapter provides a brief research background, which includes: crude oil basics, crude oil chaotic behavior, factors affecting the prices of crude oil and the basic concepts of data mining /machine learning technology and their applications. Finally, the chapter ends with highlighted the Zachman framework approach. This chapter will help in the understanding of the circumstances and conditions surrounding the crude oil prediction.

### 2.1. Crude Oil Basics

Oil is created by the remains of prehistoric animals and plants, which died between 245 and 544 million years ago and is extremely compressed on the sea bed by billions of tons of silt, then decomposed and mixed with sand and silt and exposed to temperatures in the earth's crust [38] as illustrated in Figure 2.1.



**Figure 2.1** Oil formation [39]

Crude oil formation occurs by a mix of hydrocarbons that exists in liquid phase in natural underground reservoirs with certain minerals such as sulfur under extreme

pressure and remains liquid at atmospheric pressure after passing through surface separating facilities [38, 40]. Sometimes the term ‘petroleum’ refers to crude oil, but it may also refer to other related hydrocarbons [39] .

There are various types of crude oil and more than 300 different types of crude oil are produced around the world with different qualities and characteristics [41]. Properties of the crude oil determine the mix of final petroleum products and determine an appropriate price for it [42]. The most important of these properties<sup>2</sup> are *density* (API gravity)<sup>3</sup> and *sulfur content*. Density property is essential to determine whether a specific type of crude oil. Crude oils with lower density, referred to as *light crude* and *heavy crude* have a high density and low share of light hydrocarbons and require a much more complex refining process. Sulfur is a naturally occurring element in crude oil, is an undesirable property. The higher sulfur level, the bigger effect it will have on the environment and refiners make heavy investments in order to remove it and more corrosive effect will have on the equipment. Crude oil with high sulfur are referred to as *sour* crudes while those with low sulfur content are referred to as *sweet* crudes [41, 42]. High-quality crude oils are characterized by low density (light) and low sulfur content (sweet) and are typically more expensive than their heavy and sour counterparts [41].

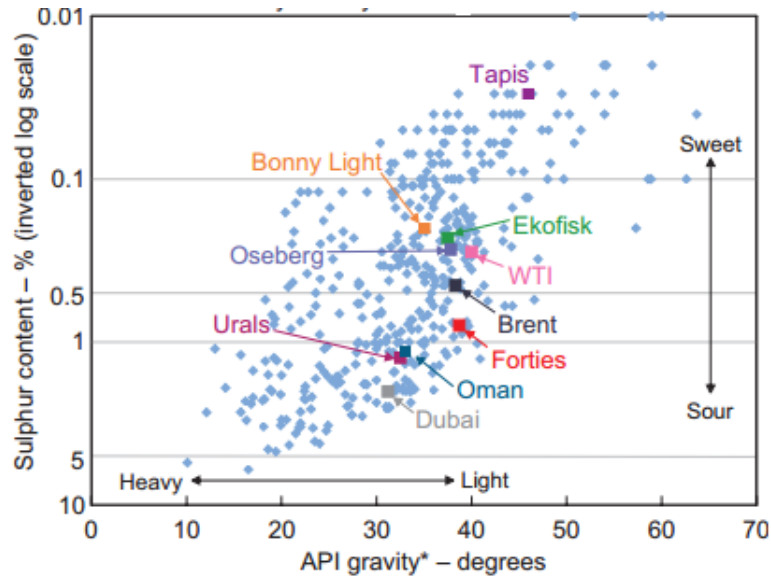
The price of a particular crude oil is usually set at a discount or at a premium to a marker or reference price (benchmarks) for buyers and sellers of crude oil [42]. There are several international benchmarks of pricing system and the most popular are West Texas Intermediate (WTI). In the US it is light crude oil and has low sulfur content and these properties make it excellent for making gasoline [43] and it is mostly used for crude oil price in the United States. In the North Sea (Brent), Brent 2/3 of the world (Europe or Africa) use it as a benchmark for pricing the crude oil. On the other hand, Dubai is mainly used in the Middle East region for exporting crude oil price to Asia and pricing index [44]. The other well-known benchmark includes (Tapis) from Malaysia, (Minas) of Indonesia, OPEC Reference Basket used by OPEC, Bonny Light used in

---

<sup>2</sup> Other important characteristics include the amount of salt water, sediment and metal impurities.

<sup>3</sup> API gravity express the gravity or density of liquid petroleum products devised jointly by the American Petroleum Institute and the NIST - National Institute of Standards and Technology.

Nigeria, Urals oil used in Russia and Mexico's Isthmus [45]. Figure 2.2 illustrates crude oil benchmarks and their characteristics.



**Figure 2.2** The crude oil benchmark by density and sulfur content [41]

Recently world had suffered from political instability, wars and conflicts, especially in the Middle East oil-rich areas, such as the Arab Spring movements in Tunisia, Libya, Egypt, Syria and Yemen. With the acceleration of technological development, these factors and others had influences on the oil market and volatile behavior of trading. Therefore crude oil prices are characterized by high volatility and some drastic shocks [46] , and the dominant feature of the behavior of the oil prices is becoming is very chaotic.

## 2.2 Crude Oil Chaotic Behavior

Crude oil price movement has a ‘chaotic behavior’ because of the large fluctuations in the crude oil prices. Panas and Ninni [47] illustrated strong evidence for the presence of chaos and non-linear dynamics in daily oil prices for the Rotterdam and Mediterranean petroleum markets. The WTI price, which traded in the New York Mercantile Exchange's (NYMEX) during 1970’s to 2011 and the crude oil price behavior is reviewed as follows:

- In 1972, the price of crude oil was under \$3.50 per barrel<sup>4</sup>. Major geopolitical events occurred, such as the Arab Embargo by several Arab-exporting nations joined by Iran as a reaction to the support of the United States and many countries in the western world for Israel in its attack by Syria and Egypt on October 5, 1973. The net loss of four million barrels per day extended through March of 1974. By the end of 1974, the nominal price of oil had quadrupled to more than \$12.00 [4, 8].

- From 1974 to 1978, the world crude oil price was relatively flat ranging from \$12.52 per barrel to \$14.57 per barrel. In 1979 the Iranian revolution and the Iran-Iraq war that started in 1980, led to another round of crude oil price increase to more than double. The nominal price went from \$35 in 1978 to \$71 per barrel in 1981. In the early 1980s a recession, created and develop non -OPEC production and production from Canada's oil sands, lead OPEC to reducing their share then significant downward impact on crude oil market prices declined to below \$40 per barrel [8].

- The price of crude oil spiked in 1990-1991 because of the Iraqi invasion of Kuwait and the ensuing Gulf War with the lower production. In 1994, the inflation adjusted oil price reached the lowest level since 1973 [4, 8].

- When crude oil prices plummeted \$10 per barrel, following crisis in Asia (1998) OPEC increased its quota by 2.5 million barrels per day (10 percent) to 27.5 million barrels per day .In 1998, Asian Pacific oil consumption declined for the first time since 1982. In response, OPEC cut quotas by 1.25 million barrels per day in April and another 1.335 million in July. The price continued down through December 1998 this allowed OPEC to regain control over the oil market. The cuts were sufficient to move prices above \$25 per barrel. With growing U.S. and world economies, the price continued to rise throughout 2000 to a post 1981 high [4, 8].

- In 2001, a weakened US economy and increase in non-OPEC production by Russian production increases put downward pressure on prices. Terrorist attacks in September 11, 2001 reversed the price situation. By the end of 2002, problems in Venezuela led to a strike at PDVSA causing Venezuelan production to plummet. It

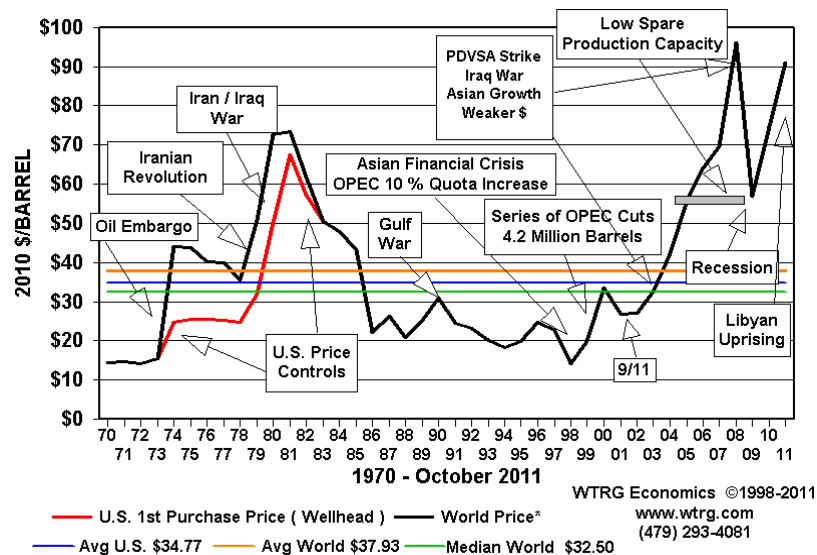
---

<sup>4</sup> Barrel: unit of value for crude oil or petroleum products

increased drastically, since the middle of 2004, with the average price of \$31.14 per barrel in 2003 to the average price of \$56.47 per barrel in 2005. The price was on an upward curve for the whole year of 2007 and reached its peak on July 2008 [4].

- The crude oil price reached \$147.27 on July 11, 2008 [48], and it later dropped to \$ 61.95 in 2009 [49] . In 2010-2011 Crude oil prices recovered again and continued to rise to \$79.48 and \$94.88 respectively [50]. The volatility of the crude oil prices between 1970s and 2011 with related events was presented by [4] in Figure 2.3.

- The main reason is attributed to the irregularity and the sudden changes in the oil price behavior that marked the last three decades its complexity and irregularity. The complexity is mainly due to its dependence on many global and national economic factors. Section 2.3 presents the most important of these factors effecting the crude oil prices according to the economic expert opinion.



**Figure 2.3** The volatility of the crude oil prices between 1970’s and 2011 [4].

## 2.3 Factors Affecting the Prices of Crude Oil

Crude oil price movements are even chaotic, and very complex nonlinear time series, which is frequently, influenced not only by control the economic rules but also by

numerous complicated factors. Academics, analysts and politicians seem to disagree on what is the main driver for the oil price development. Usual explanations are as follows:

### 2.3.1 Fundamental Supply and Demand

OPEC and non-OPEC members are suppliers of oil at the global level, these countries, especially the members of OPEC oil reserves are keen to maintain it to remain at an appropriate level of safety against risks, therefore increasing demand for this stock leads to an increase in oil prices and the continuous increase in oil demand is the fundamentals behind the volatility of oil prices. This increase of oil demand produces economic growth and we all know that oil wealth are depleted and their widening gap over the time between the increase in demand and production. This indicates that a rise in oil prices in the future will be a realistic issue [7]. Figure 2.4 indicates the crude oil supply and demand during the period between 2002-2012.

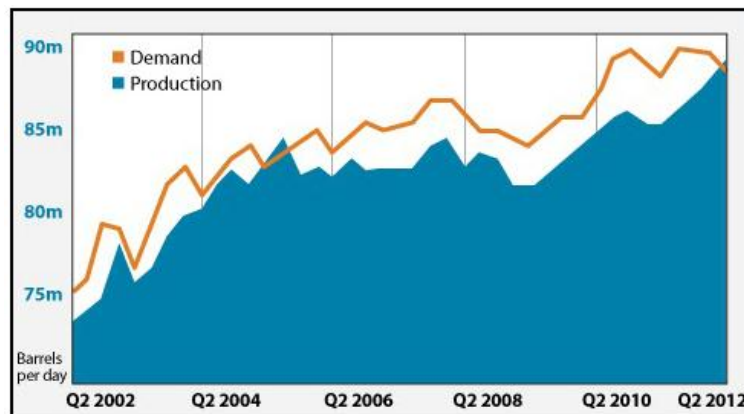


Figure 2.4 Crude oil supply and demand (2002-2012) [51]

### 2.3.2 Geopolitical Events

Much of the world's crude oil is located in regions that have been prone historically to political upheaval such as the Middle East. Several major oil price shocks have occurred at the same time as supply disruptions triggered by political events, most notably the first and second Gulf war, the search for weapons of mass destruction in Iraq and their consequence, Iran nuclear threats and political tensions arising from that, also in recent years the Arab revolutions in Tunisia, Libya, Egypt, Syria and Yemen. Or



curbs on potential development of resources such political events have been seen in Nigeria, Venezuela, Iraq, Iran, and Libya. All these events have an impact on the political and security situation in the region and the rising oil prices and their instability [5, 7].

### **2.3.3 Seasonal Weather and Natural Disasters**

Cold weather, snow and storms lead to volatility in oil prices in several European countries and the U.S. dependence on oil for heating and increasing demand for supplies of oil as well as natural disasters may cause damage of HR and oil production facilities, leading to a reduction of the amount produced and then rising oil prices. In summer during the travel season, increase in gasoline demand also leads to increase of oil prices [7, 8] Hurricane, floods and earthquakes can have a significant influence on oil prices. In 2005 Hurricane Katrina<sup>5</sup> and Rita<sup>6</sup> [52] caused extensive production breakdown with offshore oil and gas platforms destroyed and pipeline damages. Traders' concerns rose in May 2011 when Mississippi River floods threatened oil refineries this lead to pushing the oil prices [49] . On other hand, oil prices fell on after a massive earthquake shook Japan, shutting refineries and other industrial facilities in the world's third-largest oil consumer [53] .

### **2.3.4 Speculation and Market Conditions**

Commodity traders registered with the Commodity Futures Trading Commission (CFTC) are playing significant role in oil prices through the establishment of agreements to buy or sell oil at a specified date in the future for an agreed price by traders, these agreements known as bidding on oil futures contracts [54]. Commodities traders are divided into two categories:

1. Representatives of companies who actually use the oil.
2. Speculators who want to make money from changes in the price of oil.

---

<sup>5</sup> Katrina was one of the most devastating natural disasters in United States history. it was an extraordinarily powerful and deadly hurricane that carved a wide swath of catastrophic damage and inflicted large loss of life

<sup>6</sup> Rita was an intense hurricane that reached Category 5 strength over the central Gulf of Mexico, Rita produced significant storm surge that devastated coastal communities in southwestern Louisiana, and its winds, rain, and tornadoes caused fatalities and a wide swath of damage from eastern Texas to Alabama. Additionally, Rita caused floods due to storm surge in portions of the Florida Keys.

Speculators invest in oil futures, essentially bets on how much oil will cost at a later date ,and these futures are traded in the NYMEX, as well as the International Petroleum Exchange. In addition to exchange rate shock has a significant negative impact on crude oil prices while the impulse response of the exchange rate variable to a crude oil price shock was statistically insignificant [6].

Next Section presents the basic concepts of machine learning and data mining techniques available and their applications.

## 2.4 What is Data Mining / Machine Learning?

Numerous definitions and descriptions for data mining and machine learning exist in the literature. In [55] data mining is defined as “the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories”. It is also defined as “the extraction of implicit, previously unknown, and potentially useful information from data” [56].

Although in some literature, the concepts, data mining and machine learning are often used mutually, according to [57] the building blocks of data mining is the evolution of a field with the confluences of various disciplines, like database management systems, statistics, artificial intelligence (AI), and machine learning (ML). According to Gorunescu [58], machine learning (ML) represents an extremely important scientific discipline in the development of data mining, using techniques that allow the computer to learn with training.

Literature chooses to classify data mining technology into either *predictive* or *descriptive* modeling. The goal of a predictive modeling is to predict the value of one column based on the value of other columns. It refers to the process of building a model that will permit the value of one variable to be predicted from the known values of other variables [58, 59]. *Classification* and *regression* are the two most common tasks in predictive modeling. If the label is discrete (containing a fixed set of values), the task is called classification. If the label is a continuous value, the task is called regression. Descriptive modeling called unsupervised technique. Being unsupervised task, it helps

us to see patterns and segments that behave similarly. The two most common descriptive modeling tasks are *association* and *clustering*, achieved by the identification of patterns that describe data and that can be easily understood by the user.

There are various successful applications of data mining techniques in real-life situations. In the banking and financial services domain, identifying customers who are most likely interested in a new credit product is one instance of a data mining application. Another example is telecommunications fraud prediction in mobile telephony and services [58]. The application of data mining in biomedical and DNA data analysis is also discussed in [55]. Recently data mining is used in discovering large-scale sequencing pattern, in identifying and study human gene for the development of new pharmaceuticals in cancer therapies [60].

The crude oil prediction problem is one case of data mining using regression approach. [61, 62] showed that in the last years this particular area of research development. A review of the existing literature hitherto published on predict crude oil prices and their application of ML techniques present with more detail in section 3.2.2

## **2.4.1 Machine Learning Algorithms**

### **A. Support Vector Regression**

Support Vector Machines (SVM) are supervised learning models used for classification and regression analysis. It offers one of the most robust and accurate methods among all well-known algorithms [63]. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Support Vector Regression (SVR) is an SVM algorithm to handle nonlinear prediction [64]. SMOreg is an iterative optimization algorithm proposed by Smola and Schölkopf [65] for using SVR regression. SMOreg uses constraints structural risk minimization as the model and has the good ability to model regression, prediction with non-linear data.

## **B. Instance Based Learning**

Instance-based learning (IBL) algorithms are derived from the nearest neighbor machine learning philosophy. IBK is the number of nearest neighbors ( $k$ ) can be set manually or determined automatically. Each unseen instance is always compared with existing ones using a distance metric. Instance-based algorithms have numerous advantages one benefit of this approach is its simplicity[66].

## **C. K Star**

K Star ( $K^*$ ) is an instanced based classifier [67]. A new data instance is classified by comparing it to the stored examples in order to find the most similar ones. This approach is also called nearest neighbor classification and the main advantage of this approach is that arbitrary complex structures in the data can be captured and training and retraining this model is fast.

## **D. Isotonic Regression**

The isotonic regression [68] finds a non-decreasing approximation of a function while minimizing the mean squared error on the training data. The algorithm sweeps through the data and adjusts the estimate to the best possible fit with constraints. Sometimes it also needs to modify previous points to make sure the new estimate does not violate the constraints. The benefit of such a model is that it does not assume any form for the target function such as linearity [69].

## **E. Extra-Tree**

The Extra-Trees algorithm constructs an ensemble of unpruned decision or regression trees. At each node number of attributes were selected randomly and splitting a node with minimum sample size. It is generated numerous times with the original learning sample to produce an ensemble model. The predictions of the trees are combined to get the final prediction, by majority vote in classification problems and average in regression problems. The basic differences with other tree based ensemble methods are it uses the whole learning sample and splits nodes by choosing cut points fully randomly [70].

## F. REP-Tree

Reduced Error Pruning Tree (REPTree) is a fast decision tree learner. It builds a decision tree based on information gain or reducing the variance and prunes it using reduced-error pruning (REP) with back over fitting [71].

## G. Neural Networks (NNs)

NNs are computer models constructed to mimic the functions of the human brain through parallel computation of several input vectors. NNs are composed of neurons distributed in the input, hidden, and output layers [72]. The most common types of neural network are feed forward neural networks, recurrent neural network, and radial basis function networks.

### 1. Feed Forward Neural Networks (FFN)

Backpropagation [73] method is a supervised learning scheme and the most popular technique in multilayer networks when a set of input produces its own actual output and then compare it with the target value by calculating the error, after that error is fed back through the network. The weights of each connection are adjusted to reduce the error in several ways, such as gradient descent etc. until sufficient performance is achieved. To improve the generalization, there are several learning methods such as Levenberg – Marquardt (LM), Bayesian regularization (BR) and BFGS quasi-Newton (BFG-QN) back propagation algorithm [74].

In addition, each neuron in a particular layer is connected with all neurons in the next layer. The connection between the  $i^{\text{th}}$  and  $j^{\text{th}}$  neuron (in a different layer) is characterized by the weight coefficient  $\omega_{ij}$  and the  $i^{\text{th}}$  neuron itself is characterized by the threshold coefficient  $v_i$  (Fig. 2.5). The weight coefficient reflects the degree of importance of the given connection in the neural network. The output value of the  $i^{\text{th}}$  neuron  $x_i$  is determined by Eqs. (2.1) and (2.2)):

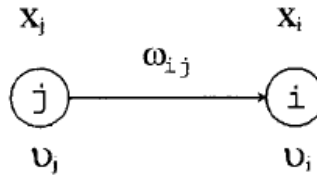
$$x_i = f(\xi_i) \quad (2.1)$$

$$\xi_i = v_i + \sum_{j=1}^N \omega_{ij} x_j \quad (2.2)$$

Where  $N$  is the neurons' number,  $\xi_i$  is the potential of the  $i^{\text{th}}$  neuron and function  $f(\xi_i)$  is the so-called transfer function. The supervised adaptation process varies the threshold coefficients  $v_i$  and weight coefficients  $\omega_{ij}$  to minimize the sum of the squared differences between the computed and required output values. This is accomplished by minimization of the objective function  $E$ , given in equation (2.3):

$$E = \sum_O \frac{1}{2} (x_o - x_d)^2 \quad (2.3)$$

Where  $x_o$ , and  $x_d$ , are vectors composed of the computed and desired output neurons and summation runs over all output neurons  $O$ .



**Figure 2.5** Connection between two neurons  $i$  and  $j$

## 2. Recurrent Neural Network (RCN)

RCN is the state of the art in nonlinear time series prediction, system identification, and temporal pattern classification. As the output of the network at time  $t$  is used along with a new input to compute the output of the network at time  $t + 1$ , the response of the network is dynamic [75].

$$x_i(t) = v_i(t) + \sum_{j=1}^N \omega_{ij} x_j(t-1) \quad (2.4)$$

### 3. Radial Basis Function (RBF)

Radial Basis Function (RBF) [75] is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. RBF is successful in numerous fields especially for system control, time series and prediction.

### H. Adaptive Neuro Fuzzy Inference System (ANFIS)

Mathematical and statistical methods are not well suitable for expression of human experiences such as perception, logic and uncertain concepts. A fuzzy inference system [77] employing fuzzy *if-then* rules can provide a framework to model human knowledge. Takagi, Sugeno and Kang (TSK) [78] proposed a fuzzy inference method in which the conclusion of a fuzzy rule is constituted by a weighted linear combination of the crisp inputs rather than a fuzzy set.

There is no systematic way to transform experiences of knowledge of human experts to the knowledge base of a fuzzy inference system (FIS). On the other hand, in an Artificial Neural Network (ANN) it is hard to extract structured knowledge from either the weights or the configuration of the ANN. To overcome these drawbacks and to take advantages of these two approaches integrated system was built by combining the concepts of (FIS) and (ANN) modeling this system called Adaptive

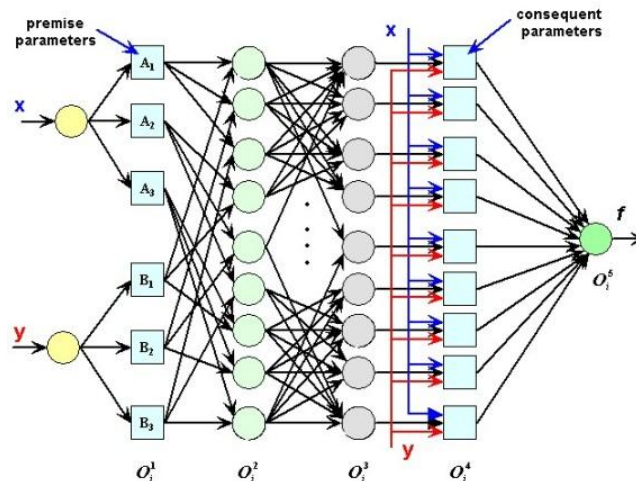


Figure 2.6 The architecture of the ANFIS [76]

Neuro Fuzzy Inference System (ANFIS) [76]. ANFIS implements a Takagi Sugeno Kang (TSK) fuzzy inference system. For a first order TSK model, a common rule set with two fuzzy *if-then* rules is represented as follows:

*Rule 1: If x is A1 and y is B1, then f1 = p1x + q1y + r1*

*Rule 2: If x is A2 and y is B2, then f2 = p2x + q2y + r2*

Where x and y are linguistic variables and A1, A2, B1, B2 are corresponding fuzzy sets and p1, q1, r1 and p2, q2, r2 are linear parameters. ANFIS makes use of a mixture of back propagation to learn the premise parameters and least mean square estimation to determine the consequent parameters. Figure 2.6 illustrates the ANFIS structure.

## **I. Particle Swarm Optimization**

Particle swarm optimization [79] is a technique for simulating the social and cooperative behavior of different types such as birds, fish, bees and human beings. The PSO composed of a population (swarm) of possible solutions called particles. These particles move through the search domain with a specified velocity in search of optimal solutions. Each particle maintains a memory, which helps it in keeping the track of its previous best position.

## **2.5 Zachman Framework**

The energy market, which consumes and stores large quantities of crude oil and trading in millions of barrels between the process of buying and selling every day at prices, needs to be agreed upon by several parties. These prices reflect supply and demand situation in the current market [42]. So the flow of data and understanding is more complex, which justifies the absence of any framework to facilitate the process of following up crude oil prices and understanding what is referred to by these prices.

Despite the vitality and importance of the crude oil pricing system in the oil market, it is worth mentioning that the process of predicting future crude oil prices depends on a clear understanding of the crude oil pricing system, which is an integral part of it. So the main objective of this section is to provide an architectural model for



enterprise knowledge based on the Zachman Framework (ZF) for the international crude oil pricing and prediction system to organize the data, process, and technology.

Many authors [80-82] attempted to compare between Enterprise Architecture Frameworks (EAFs) to provide guidelines in determining and selecting the best EAF.

However, most EAF differ in their approach and level of detail. Some are proposed guidelines, whereas others have specific methodologies and aspects to follow, some of these frameworks were developed for very specific areas, whereas others have broader functionality [81]. Zachman framework was selected for following reasons:

- Its robust nature lends itself to an application of any size because the framework is flexible [83].
- It is the most comprehensive using a number of viewpoints (dimensions and perspectives) and related to different aspects [81].
- ZF is widely accepted and frequently used as the main framework in EA and serves as the basis for numerous other models e.g., FEAF, TOGAF and TEAF [84].
- Its wide usage in practice and application of information architecture as an approach in several fields [85-87].
- ZF does not execute a method and it does not restrict any user to a set of pre-defined artifacts [88] [89].

In 1987, John A. Zachman, proposed a simple classification schema for classifying and organizing the descriptive representations of an Enterprise that are significant to the management and development of the Enterprise's systems, which is now called Zachman Framework [90, 91]. ZF contain a two dimensional matrix (Table 2.1), which includes six rows denoted to different perspectives of the enterprise from the points of view of the planner (row 1), the owner (row 2), the designer or architect (row 3), the builder (row4), sub-contractors (row 5) and user (row 6). Meanwhile the six columns explain different aspects of the process (Data, Function, Network, People, Time, and Motivation) in formula of primitive linguistic interrogatives: What, How, Where, Who, When and Why. So, by answering all of these questions, while taking into the account the six different perspectives provided a very comprehensive documentation system. .

**TABLE 2.1** The basic structure for Zachman framework.

	What	How	Where	Who	When	Why
Planner						
Owner						
Designer						
Builder						
Subcontractor						
User						

The framework arises with the following set of rules to maintain the reliability of architectural descriptions:

- **Rule1:** No order in the columns.
- **Rule2:** Each column has a simple generic model.
- **Rule3:** The basic model of each column must be unique.
- **Rule4:** Each row describes a distinct, unique perspective.
- **Rule5:** Each cell is unique.
- **Rule6:** The composite or integration of all cell models in one row constitutes a complete model from the perspective of that row.
- **Rule7:** The logic is recursive.

## 2.6 Summary

This chapter presented a summary of research background which includes: In first section crude oil basics. It provides a method of formation of oil, types of crude oil and basic concepts related to the oil industry, such as benchmark crude oil, second section described crude oil chaotic behavior through explained evidence for the presence of chaos and non-linear dynamics in oil prices according to global events. Then section three showed factors affecting the prices of crude oil, usually academics, analysts and politicians seem to disagree on what is the main driver for the oil price development, in this section four fundamental factors that had played an important role in oil prices were presented. Followed by defining the basic concepts of machine learning/data mining technology, their applications and presented examples of popular machine learning algorithms. Finally the end of the chapter by presented Zachman framework. The main

objective in this chapter is explain basic concepts in the oil industry to help in the understanding of the circumstances and conditions surrounding of the crude oil prediction.

# CHAPTER THREE

## 3. LITERATURE REVIEW

### Overview

This Chapter describes an exhaustive overview of the existing literature for the application of predicting crude oil prices. The Chapter begins with the overview of crude oil prediction in the literature, and presents three major factors affecting crude oil price fluctuation, oil price volatility and prediction models and enterprise architecture frameworks. Finally, highlighted some of the problems that may limit the success of prediction models to avoid it.

### 3.1 Introduction

Oil embodies a vital role in the world economy because it is the backbone and origin of many industries. It is an important source of energy representing an indispensable raw material and a major component in many manufacturing process and transportation. Oil price suffers from high volatility and fluctuations. In global markets, it is the most active and heavily traded commodity. Therefore, it has attracted the attention of researchers and authors to study it from different aspects and under different categories. In the academic literature, prediction of oil prices has been discussed in different formats. For example, Xie, et al. [35] divided the literature as: a) Oil prices are confined between demand and supply framework, b) Oil price volatility analysis and c) Oil price forecasting. Pan, et al. [5] analyzed it in three parts: a) The future as a predictor to spot oil prices, b) Economic models for explanation and prediction and c) Intelligent computing models to predict oil price. Additionally Tuo and Yanbing [3] classified oil price forecasting literature into three parts: a) In the formal model, those factors affecting the analysis of oil prices. b) Theoretical model was more disbursed to notice the behavior of market structure and c) Simulation models as represented by information technology.

According to our objectives, the literature review was divided into three main parts:

- Factors affecting crude oil price fluctuation
- Oil price volatility and prediction models
- Enterprise architecture frameworks.

## **3.2 Factors Affecting Crude Oil Price Fluctuation**

In Section 2.3, fundamental supply and demand, geopolitical events, seasonal weather and natural disasters, and speculation and market conditions, were identified as the most important factors that affect the prices of crude oil. For the purpose of understanding the factors that affect the volatility of oil prices, many scholars have divided and studied these factors in an attempt to explain their role and analysis of their impact in oil prices.

Pan, et al. [5] attributed the volatility of oil prices to three main factors: Increase in demand and supply shortages, possibly caused by economic growth or the behaviors of oil producing countries, exogenous events such as wars, natural disasters, etc. and endogenous factors such as speculations in the markets. Ji [92] categorized the major factors responsible for influencing crude oil price fluctuation into six categories: Macroeconomics, speculation, stock market, supply and demand, exchange rate, and commodity market. In a related study Hamilton [32] analyzed the factors that play roles in the fluctuation of oil prices, includes: commodity price speculation, OPEC monopoly pricing, strong world demand, geological limitations on increasing production, and an increasingly important contribution of the scarcity rent, in terms of statistical regularities and perspective of economic theory. The author concluded that although scarcity rent made an insignificant contribution to the price of oil in 1997, it might be an important feature of the most recent data. Wang, et al. [93] classified the main patterns that affect crude oil price volatility into six sets, which includes military and political, OPEC policy, Non-OPEC policy, natural disaster, world economic condition and other factors such as speculation and exchange rates, this category led to 24 patterns such as war, revolution, terrorist attack and so on.

Morana [11] investigated that financial shocks have played an important role in determining the oil price, the author found out that during the period of 2004-2010 oil prices increased by 65%, while 33 % represented financial shocks. Moreover, Fattouh, et al. [94] pay special attention to investigating whether financial market information can help to forecast the price of oil in physical markets. In a recent study by Kilian and Murphy [9], it presented an evidence that speculative demand and inventories played an important role during earlier oil price shock in 1979, 1986, and 1990 in addition to an unexpected increase in world oil consumption.

On the other hand numerous researchers tried to understand the relationship between the price of oil and fundamental factors. Lizardo and Mollick [10] concluded that an inverse relationship between oil prices and the value of the U.S. Dollar in the long-term led more to an increase in oil prices to a decrease in US Dollar rate for the oil exporting countries, while the importers of oil currencies goes down compared to the USD value. Bénassy-Quéré, et al. [95] provided evidence that a 10% rise in the oil price led to a 4.3% appreciation of the dollar in the long term, but they couldn't explain the increase in the oil with the decline in the U.S. Dollar exchange rate in the period from 2002 to 2004. Climatic changes could also cause significant impacts on oil activities. Huang, et al. [96] developed an expert system (ES) for assessing climate-change impact within the petroleum sector. The results indicated that the impacts of increased temperature and natural hazards would be very significant on most of the petroleum-related processes.

To assess and measure the degree of OPEC market power in the oil market, Golombek, et al. [97] utilized estimate a parsimonious dominant model used quarterly data on oil prices for the 1986-2009 period. The experiments indicated that OPEC exercised its market power during the sample period, world GDP is the main driver of long-run oil prices, and supply factors have become more important in recent years. Similarly Yang et al. [98] examined the market structure of OPEC, the stable and unstable demand structure, and related elasticity of demand to study the price volatility of the crude oil market by using the error correction model. Their results indicated that given the 4% cut in OPEC production, the oil price is expected to increase unless the

recession is severe and oil price would be diminished if non-OPEC or domestic production were greatly expanded. However, the success of pricing models that focus on OPEC behavior lasted for only a short time.

### **3.3 Oil Price Volatility and Prediction Models**

There is a vast and still growing literature that aims to explain and address the stochastic behavior of oil prices. Chen, et al. [99] concluded that the fluctuation of crude oil prices in the global market at present has caused a growing interest and efforts in examining current models and proposing new ones and identifying improved approaches in order to avoid the effects of crude oil price unpredictability.

Emerged different classifications of prediction models, For example, Azadeh, et al. [22] categorized forecasting methods according to the time schedule long-term, short-term and medium term, also authors mentioned other classifications based on qualitative methods, regression methods, multiple equation methods and time series methods. Furthermore Weiqi, et al. [100], Wang, et al. [31], Xie, et al. [35] classified forecasting methods as qualitative and quantitative. Otherwise Zhang, et al. [101] grouped the studies for analysis and predicting of crude oil price structure into two categories: structure models and data-driven methods. Moreover Fan, et al. [20] included oil price forecast approaches in two classes: Single-factor time series models (considered time as an independent variable) and multi-factor models (takes into account the main influential factors of oil prices). Khashman and Nwulu [102] classified research in predicting the price of crude oil into two categories: research using econometric models or tools and research using soft computing methods. In this section the crude oil prediction models will be clustered into:

- Statistical and econometric prediction models.
- Machine learning techniques in crude oil price models.

#### **3.3.1 Statistical and Econometric Prediction Models**

In the past decades, traditional statistical and econometric techniques, such as linear regression (LinR), co-integration analysis, Generalized Autoregressive Condition-

al Heteroskedastic (GARCH) models, naive random walk, vector auto-regression (VAR) and error correction models (ECM) have been widely applied to crude oil price prediction [17].

Early attempts to model and forecast volatility Huntington [103] implemented a sophisticated econometric model to predict crude oil prices in the 1980s. Abramson and Finizza [15] suggested a probabilistic method for predictions of average annual oil prices. Gulen (1998) followed them and used co-integration analysis to predict the WTI price, using monthly data to cover the period of March 1983 to Oct 1995, and Barone-Adesi [104], [105] suggested a semi-parametric approach for oil price forecasting. Similarly Morana [106] proposed a semi-parametric approach based on the bootstrap approach, using daily oil prices for the period from 4 January 1982 to 21 January 1999 to predict the oil prices.

Furthermore Lanza, et al. [107] proved the relationships between heavy crude oil and product price using Co-integration and error correction models (EC) and evaluated the predictive power of the specification in forecasting crude oil prices. Weiqi, et al. [100] presented a structural practical econometric model to forecast oil price three months ahead of the Brent crude spot price, using the explanatory variable of defined relative inventory and OPEC production and compared their results with the ARIMA model according to the statistical properties of the data evaluation, the results indicated that the proposed structural model has a better performance than the ARIMA model. Coppola [16] proved that vector error correction (VECM) surpassed the random walk model (RWM) in forecasting oil price movements of 1-month futures contracts.

In 2014, Zhao and Wang [108] proposed an autoregressive integrated moving average (ARIMA(4, 3, 0)) to predict world crude oil. Data from 1970 to 2006 was used as a model, experiment shows the model had mean absolute percentage error (MAPE) was 4.059%. In the same year, Xie, et al. [109] used the decomposition-based VAR model, scrutinizes the additional information of high-low extreme values to predict crude oil prices, the experiments over year 1986–2013 for WTI spot crude oil price.



Results reported the dominance of decomposition-based VAR over both efficient market model and ARIMA model.

The above models can provide good prediction results when the price series under study is linear or near linear. However, several experiments have proved that the prediction performance might be very poor if one continued using these traditional statistical and econometric models [18]. The major reason causing this phenomenon was that the traditional statistical and econometric models were built on linear assumptions and they cannot capture the nonlinear patterns hidden in the crude oil price series [17].

In addition, the econometric model depends on making a strong assumption about the problem. This means if the assumptions are not correct, the model could generate misleading results [19]. Due to the limitations of the traditional statistical and econometric models, some nonlinear and emerging Machine Learning (ML) models, such as artificial neural networks (ANN) are viewed as non-parametric, nonlinear, and assumption-free models [110]. More methods and techniques using ML in literature are presented in the next Section.

### **3.3.2 Machine Learning Techniques in Crude Oil Price Models**

Single and hybridized ML techniques have been applied to predict crude oil prices using voluminous historical data to build prediction models [61]. The Sections below discuss the different kinds of application of ML in the predicting oil prices as follows:

- A. Single prediction of crude oil price models.
- B. Hybrid prediction crude of oil price models.

#### **A. Single Crude Oil Price Prediction Models**

Artificial neural networks (ANN) are designed to represent data by simulating the work of the human brain. ANN's emerged in different areas such as industrial, medical and business, and has achieved successful results. Therefore, many researchers also used ANN in the prediction of oil prices. Haidar, et al. [111] suggested a network

to predict the oil prices using two groups of inputs, crude oil futures data, and Dollar index, S&P500, gold price and heating oil price. The authors measured performance by heat rate, root mean square error, correlation coefficient, mean squared error and mean absolute error. The authors concluded that heating oil spot price support forecast crude oil spot price in numerous steps prediction. Alizadeh and Mafinezhad [112] proposed General Regression Neural network (GRNN) using six factors monthly data to predict Brent crude oil price. Experiment results show that the model achieved high accuracy in normal and crisis situations. Mingming and Jinliang [113] collected data covering Brent and West Texas Intermediate (WTI) from 1946 to 2010 and adopted multiple wavelet recurrent neural networks (MWRNNs) to forecast crude oil prices. The study showed that the model has high prediction accuracy. In recent study, Godarzi, et al. [114] developed a dynamic ANN to predict oil price using train data from 1974 to 2004 and from 2004 to 2009 to test the model. The results indicated that the proposed model was more accurate than time series and static ANN. This was based on the fact that ANN models often suffer from the local minima, over fitting problems and the difficulty of determining appropriate network architectures [31].

Support Vector Machines (SVM) provide a class of competitive learning algorithms to improve generalization performance of neural networks and accomplish global optimum solutions simultaneously [31]. Khashman and Nwulu [115] designed an intelligent system based on SVM to predict the price of crude oil involving eight input factors (global demand; a random world event; among others). Empirical results show high prediction accuracy.

The success of the application of support vector machine in solving several problems depend on the appropriate selection, and use of a kernel function[116], therefore Chiroma, et al. [116] compared the performances of five different kernel functions of the support vector machine to provide better understanding of the behavior of the kernel functions for support vector machine and improve the accuracy of crude oil prediction. Monthly data from 1987 to 2012 were utilized. The empirical results exhibited that the wave kernel function is significantly better than that of radial basis function, polynomial, exponential, and sigmoid kernel functions on crude oil prediction. However, the

prediction accuracy of the single forecasting model needs to be improved and long time consuming when there is a large size of data, so SVM with approximate algorithms need to be integrated to reduce the time [117].

Furthermore, Fan, et al. [20] predicted crude oil prices using generalized pattern matching based on genetic algorithm (GPMGA). One-month forecasting of the Brent and WTI crude oil prices results show that the effectiveness and superiority of GPMGA when compared with PMRS and Elman network. Kaboudan [118] designed comp metric prediction of the crude oil price model based on genetic programming and neural networks using monthly closing prices of crude oil from January 1993 to December 1998. The prediction accuracies of both models were compared using naive random walk fit forecasting accuracy and the empirical results indicated that the genetic programming model was superior to the multi-layer feed forward neural network model also GP has an advantage over random walk.

Yu, et al. [14] proposed a new model based on rough-set-refined text mining (RSTM) for crude oil price predicting. The authors evaluated the model by comparing it with statistical models, time series models and neural network models, the empirical results display that RSTM outperforms other predicting models. Although, in the text mining model unlike other well-defined problem domains, expert opinions on the crude oil markets can vary wildly [5].

Fuzzy sets deal with problems which involve uncertainty and vagueness for this reason Zhang, et al. [12] applied fuzzy time series to predicting oil price using WTI oil and evaluated the performance model by root mean square error. Experimental results show that fuzzy time series can produce good forecast results. But, experiments and evidence such as [22] shows that the hybrid CI technique is superior to the single CI technique.

Section *B* explores hybrid models for predicting oil prices in the literature. Some other recent studies in the literature which were applied for predicting oil price using single model are listed in Table 3.1

**Table 3.1** Recent single CI applications in oil price prediction

Ref.	Study description	Results
[13]	Tested the relation of crude oil prices and the prediction model based on ANN to forecast and compared with LSM.	The results shown that in the short-term, the best prediction model for ANN of four, three, two and one hidden layers, respectively. The ANN of one to four hidden layers is found to be able to forecast better than the LSM.
[34]	An orthogonal wavelet support vector machine (OSVM) model for predicting the monthly prices of West Texas Intermediate crude oil prices.	Experimental results suggest that the OSVM is effective, robust, and can efficiently be used for crude oil price prediction.
[119]	A new kernel methods-hierarchical multiple kernel machine (HMKM) to predicting oil price	Empirical results demonstrate that our new system robustly outperforms traditional neural networks and regression models.
[120]	The application of two types of the Artificial Neural Networks, including Feed Forward, Back Propagation Network and Radial Basis Function Network in forecasting return volatilities of crude oil future prices using some intra markets variables especially with focus on the speculation activity.	The study show that both Radial Basis Function Network and the Feed Forward Back Propagation Network are working better than the GARCH model.

**B. Hybrid Prediction Models for Crude Oil Price**

Hybrid methods have emerged in the recent years, the basic idea of which was to complement the disadvantages of the individual models and generate synergy effect on the results to predict oil prices. Wang, et al. [93] developed a hybrid AI system of neural networks rule based expert system and web-based text mining using historical data of monthly spot prices of crude oil collected from January 1970 to December 2002. The results illustrated that the performance of hybrid forecast were more accurate than single neural network forecasting.

Mathematical and statistical methods are not well suitable for expression of human experiences such as perception, logic and uncertain concepts. A fuzzy inference system employing fuzzy rules can then provide a framework to model human. On the

other hand, Artificial Neural Network (ANN) learning mechanism is hard to extract structured knowledge from either the weights or the configuration of the ANN. To overcome these drawbacks and to take advantages of these two approaches an integrated neuro-fuzzy system was built called ANFIS. Panella, et al. [121] collected data from Europe (Brent crude oil) and the US (West Texas Intermediate crude oil) from 2001 to 2010 to forecast crude oil, natural gas, electricity, and coal prices using three different models radial basis function neural networks, adaptive neuro-fuzzy inference system networks and least-square approximation. The experimental results showed the superiority of adaptive neuro-fuzzy inference system. Ghaffari and Zare [122] applied an adaptive network-based fuzzy inference system for forecasting WTI crude oil spot price. Using daily data from 5 January 2004, to 30 April 2007, 68.18% prediction accuracy was achieved.

In order to enhance the effectiveness of the artificial intelligence techniques, an ensemble machine learning approach was built for the prediction of crude oil price Yu, et al. [17] constructed an empirical mode decomposition (EMD) based on neural network ensemble learning. They used daily West Texas Intermediate (WTI) data from 1/1/1986 to 30/9/2006 as training and Brent from 20/5/1987 to 30/9/2006 as test data. Results proved that EMD based neural network ensemble can be used for oil price prediction. Yu, et al. [123] obtained different prediction results based on different training sets and variety of models including (ARIMA) model, support vector machine regression (SVMR) model, back-propagation neural network (BPNN) model and, radial basis function network (RBFN) model, and then the results were combined using a fuzzy ensemble model. Their results indicated that the proposed model was superior to the single models for oil price prediction. Even though, the success of the hybrid model was highly dependent on selecting optimal parameters which significantly determined the performance ability and choosing the correct and compatible hybridization models [61]. Recent hybrid an ensemble models reported in the literature are mentioned in Table 3.2

**Table 3.2** Examples of recent hybrid and ensemble models.

Ref.	Study Description	Results
[124]	Ensemble neural network (feed-forward, recurrent and radial basis function networks) model for prediction of oil price.	Ensemble methods were found to be superior when compared to the individual neural networks and learning methods
[125]	ARIMA and BP neural network combinatorial algorithm is presented for world crude oil price forecasting.	By comparing the crude oil price prediction results of the ARIMA model, the BP neural network model and the combining algorithm, it is shown that the combination algorithm has the highest prediction precision.
[126]	A Neural Network-based Ensemble Prediction used PMRS and ECMPMRS. First used to model the trend of crude oil price, and then ECM is offered to establish to forecasting errors. Finally, NN is employed to integrate the results from the ones of PMRS and ECM to make the final forecasting values.	The empirical results show that the proposed integrated model can significantly improve the forecasting performance, compared with other four forecasting models, and it can be an alternative tool to predict crude oil prices.
[33]	Hybrid learning paradigm is proposed, through incorporating compressed sensing based de-noising (CSD) and certain artificial intelligence (AI) forecasting tool.	In the case of different data samples with different time ranges, the proposed model performs the best, indicating that the proposed CSD-AI learning paradigm is an effective and robust approach in crude oil price prediction.
[127]	A hybrid model integrating wavelet and multiple linear regressions (MLR) is. Selected to decompose an original time series into several subseries with different scale. Then, the principal component analysis (PCA) is used in processing subseries data in MLR for crude oil price forecasting. The particle swarm optimization (PSO) is used to adopt the optimal parameters of the MLR model.	The performance of the WMLR model is compared with the MLR, ARIMA, and GARCH models using various statistics measures. The experimental results show that the proposed model outperforms the individual models in forecasting of the crude oil price series.
[128]	The wavelet de-noising ARMA models ensemble by least square support vector regression with the reduced forecasting matrix dimensions by independent component analysis.	Empirical studies show the significant performance improvement when the proposed model is tested against the benchmark models.

### 3.4 Enterprise Architecture Framework

Review of viewpoints and relevant research in information architecture is required for understanding the concept of the domain area, so this subsection covers researches presenting the application of information architecture and enterprise architecture concepts. **Enterprise** is considered set of physical and complex logical business processes, which combine to constitute the infrastructure of the work. **Architecture** is defined as the art or science of building. **Framework** is defined as a basic conceptual structure [129]. **An Enterprise Architecture Framework (EAF)** is a comprehensive structural scheme of all the elements and relationships that represent the enterprise to manage and provide software and information systems to achieve the flexibility and complementarities between business perspective and its information technology (IT) within the organization [81].

It provides many advantages and features such as ability to understand and analyze weaknesses and contradictory aspects to be identified and overcome them, understand how the business rules, organizational strategies and resources are turned into a physical system, reduction of information system complexities, enabling information sharing and interaction between systems, help predict return on investment, to ensure interoperability of systems and helps control the cost of developing systems and understand where the money comes from [130]. Several popular frameworks have been proposed to architect enterprises:

- **Department of Defense Architecture Framework (DoDAF):** provides visualization infrastructure for specific stakeholder concerns through three various views. View that describes/interrelated operational elements, tasks and activities to accomplish mission operations, this called operational, view describes systems and interconnections to support the operation, namely the systems view and finally Technical Standards describe the rules governing the arrangement, interaction and interdependence of system components to augment the systems view [84].

- **Federal Enterprise Architecture Framework (FEAF):** corresponding U.S. "Federal Enterprise Architecture Framework" (FEAF), it provides a common approach for the integration of strategic, business and technology management with entities (what), activities (how) and locations (where) in respect to the views, it provides five level perspectives in order of details [89] .
- **Treasury Enterprise Architecture Framework (TEAF):** TEAF includes descriptions of a number of work products for documenting and modeling enterprise architectures. Consistent practices and common terminology within the various bureaus and offices at the Department of Treasury. The TEAF describes four views (comparable to the columns in the Zachman framework) and four perspectives (comparable to the rows in the Zachman framework), creating a matrix (comparable to the one suggested by Zachman) [89]
- **The Open Group Architectural Framework (TOGAF):** TOGAF is a framework for enterprise architecture, which covers four architectural fields; Business, Application, Data, and Technology. It also provides architectural development methods with respect to architectural domains mentioned [84].
- **Zachman Framework (ZF):** In 1987, John A. Zachman, proposed simple classification schema for classifying and organizing the descriptive representations of an Enterprise that are significant to the management and development of the Enterprise's systems, which is now called Zachman Framework [90, 91]. More details will be presented in Chapter 7.

Many authors such as [80-82] attempted to compare between EAFs and to provide guidelines in determining and selecting a best EAF. However, most EAF differ in their approach and level of detail. Some were proposed guidelines, whereas others have specific methodologies and aspects to follow, some of these frameworks were developed for very specific areas, but others have broader functionality [81]. The architecture was constructed using Zachman framework.



Zachman Framework is used as a reference model in various fields for several purposes. Ramadan and Hefnawi [86] applied top three levels of the Zachman to develop descriptive network security architecture of academic center, the results of their work show that a conceptual model can be applied to real organization. Bahill, et al. [130] designed model for Baseball based on Zachman framework, the authors concluded that Zachman framework is useful and the model is easy to understand and represent the way to learn Baseball. Abdullah [131] used Zachman framework in to determine the contents in the design of a collaborative digital library, the propose framework contributes to another dimension of a framework for digital library research and added to the digital library framework as user testing and evaluation of the system. In a similar study Fazil, et al. [132] selected Malaysia for data collection to determine the content of a semantic theses digital library and design a formal framework using Zachman framework in managing theses and specifically focused on the content organization. The study showed that a framework is more robust, offers user-friendly interface besides adaptable search and browsing interface. Piho, et al. [133] studied Zachman Framework with archetypes and archetype patterns in University of Leeds, UK through the business domain, which consists of a clinical laboratory. The study revealed that ZF helps developers to better understand business domains to design more robust and cost effective enterprise applications.

### **3.5 An Overview of the limitations of Previous Research Works**

There exist various attempts to address the predicting oil prices problem, and through literature reviewed a number of problems which were figured as shortcomings affecting the performance and reliability for the models.

[1] Most of the studies in the literature focused on constructing a model using one percentage of training and testing for example Yu, et al. [33] used the size ratio of training to testing sets is set to 4:1, Chiroma, et al. [34]divided data into 70% training, 15% validation and 15% testing and Mingming and Jinliang [113]used 60% data to train, 10% to validate, and 30% to test. However, use of a diverse range of the percentages allows a wider opportunity to achieve better results.

[2] The data gathered at different intervals (daily, weekly, monthly and quarterly) depending on the respective research objective. However, compared with weekly data and monthly data, the daily data might be more complex in terms of higher level of noise [33] and it was required to represent the real oil markets.

[3] Some authors such as [123] limited their data period to that of only five months, which constitutes the shortest period of data used [61]. For example, in [123] only 20 days of data were used as a prediction sample(test), also in [14] only 49 data points were reserved for out-of-sample tests. All this data period limits the training and testing data significantly.

[4] Despite the large number of factors that influencing oil price and it was known that the volatility of the crude oil price market was a result of the dependency on the market on numerous factors. Most researchers were interested in using one input for testing usually the benchmark WTI and Brent such as [36, 123] they do not take into consideration other inputs associated with the market. Ignore these factors in predicting the market can detract from the credibility of the prediction model. A model with good prediction results shows good interconnections between inputs and the output, which suggests the state of dependence [49].

[5] Pre-processing such as attributes (Feature) selection in case of use several factors focused on removing those features, which do not contribute to the enhancement of the prediction. Also data representation helps to clean and reduce noise in data sets and normalize them in condensing the process of prediction, then increase the effectiveness of the model. All these important processes were absent in most research.

[6] Information enterprise architecture for crude oil pricing and prediction was missing despite the vitality and importance of the crude oil pricing system in the international oil market.

### **3.6 Summary**

The purpose of this chapter is two-fold: To provide an overview of the existing literature and of recent developments for the applications of predicting crude oil prices.

On the other hand, it attempts to detect the gaps and problems faced by the success of prediction models to avoid them in future research works. This chapter divided the literature review according to the objectives of this research into three essential parts:

[1] Factors affecting crude oil price fluctuation, this section concerned to display the literature that studied the factors that affecting in crude oil prices to explain their role and analysis of their impact in oil prices.

[2] Oil price volatility and prediction models show a wide variety of models address the stochastic behavior of oil prices by cluster the models of the two fundamental groups: statistical and econometric prediction models and machine learning techniques.

[3] Enterprise architecture frameworks cover research works presenting the application of information architecture and enterprise architecture concepts to review of viewpoints and relevant research in information architecture to employment this concepts in correct way when development information architecture and enterprise architecture for crude oil pricing and prediction.

Based on the literature reviewed, a number of limitations were identified in order to face and address them.

# CHAPTER FOUR

## 4. RESEARCH METHODOLOGY

### Overview

This chapter is considered as a backbone of the thesis because it presents the approaches and techniques that provide solutions to the problem of crude oil price prediction and also attempts to answer the following research question:

*How to design a model using machine learning that can predict crude oil prices accurately?*

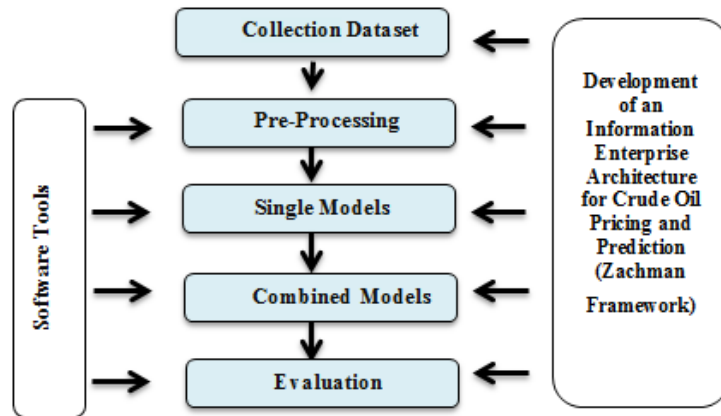
The chapter begins the discussion of the steps involved in all experiments such as dataset description, methods selected for the data preprocessing, and displays for software tools that used in this research, followed by illustrating of machine learning algorithms which include two main categories direct prediction models and combined prediction models. Several types of combined methods are presented such as a Meta prediction model, hybrid models and ensemble prediction models. Finally, the chapter ends with evaluation techniques, which assess and measure the performance and accuracy of a variety of methodologies.

### 4.1. Introduction

Monitoring the crude oil prices and the prediction of its movement is critically and represent an integrated part of the decision-making process for the production, export, development and transport for the owners of industries and investors. Therefore, the main purpose of this research is to explain, define and develop Information enterprise architecture for crude oil pricing and prediction through conceptual modeling by utilizing crude oil data collection and analysis factors that affect crude oil prices.

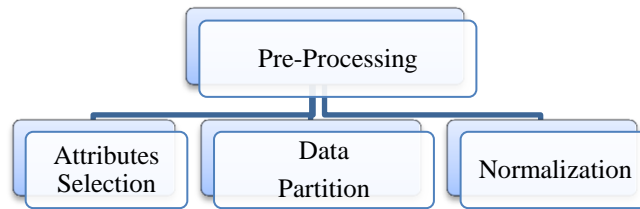
Figure 4.1 displays the overall research framework and major phases of this work and the parts to be discussed in this Chapter are highlighted. The research work

started with data collection and analysis, which was done in parallel with the analysis of crude oil pricing and prediction situations through interviews, documentation, and observations. The daily data was used because it is more complex in terms of higher level of noise compared with weekly data and monthly data. Description of the each input factor exhibited is Section 4.2.



**Figure 4.1** Basic flowchart of the crude oil predicting model and its information enterprise architecture

The next task was to prepare data via sequence steps of pre-processing. The feature is synonymous with the input variable or attribute. The irrelevant input features will lead to greater computational cost, increase utilization, and training times and may lead to over-fitting. So finding a good input data representation is very important to reduce the dimensionality of the data and reducing training, reducing the measurement and storage requirements, and utilization times. Before constructing a model several aspects of initial preparation of data were selected. Normalization, feature selection and data partition are used for preparation the inputs as shown in Figure 4.2. Section 4.3 defines these steps in more detail. It is worth mentioning that these steps are often used when designing any model in this research.



**Figure 4.2** Data preparation for crude oil price model

In the next stage, the research follows and uses the direct model in order to designing the combined model of crude oil prediction information architecture. The role of the learning task is to search more efficiently for a solution of problems. Learning systems include different components such as a set of examples, a set of possible learning results and a learning algorithm [134]. Machine learning algorithms (ML) take a different scenario, according to the scope of the level of adaptation. ML algorithms in this research were grouped into direct learning and combined learning (e.g., Meta learning), while learning at the single (base) level is focused on collecting experience on a specific learning task and learning at the Meta level is concerned with accumulating experience on the performance of multiple applications of a learning system [135]. Selecting a suitable subset of ML algorithms is considered as a search problem. The aim is to identify the set of learning algorithms with the best performance. A set of ML algorithms that have shown good (a priori) performance on datasets similar to the crude oil prediction were chosen. Another phase of selecting learning algorithm by reducing ML algorithms is by using various performance measures (e.g. Accuracy, precision, etc.) to evaluate the learning algorithm [135].

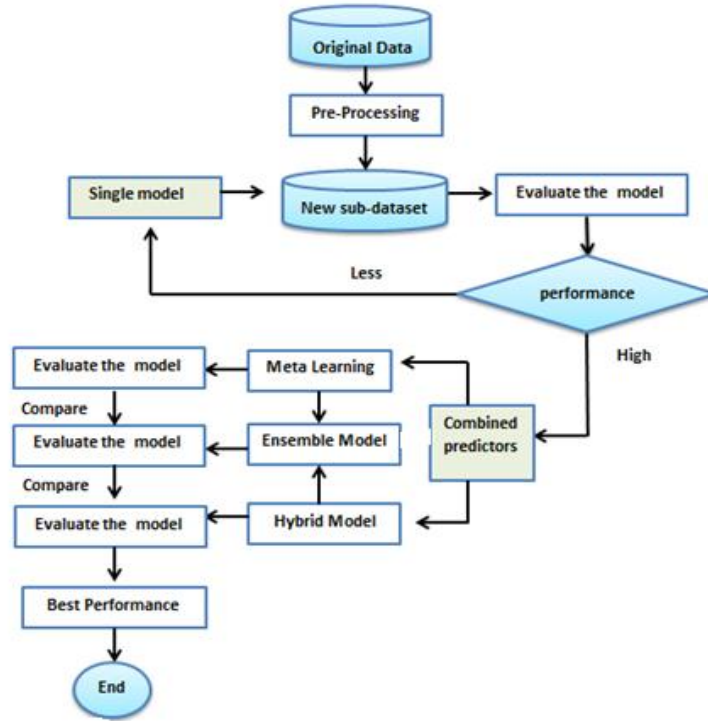
Combined predictors can be found in different styles in the literature such as Meta learning [136], Ensemble based prediction [137], Hybrid methods [138] and more. In Section 4.5.2, three different aspects for combined predictors, which includes, Meta prediction models, Hybrid prediction models, and Ensemble prediction model were illustrated.

Meta-learning succeeded on the appropriate selection of a suitable predictive model (or combination of models) depending on the domain of application by providing

automatic mapping for a suitable model to a particular task. Verikas, et al. [137] explained a distinction between a hybrid system and an ensemble of predictors. A system is considered as hybrid if several soft computing approaches are exploited for data analysis, but only one direct predictor is applied to make a final decision and to obtain a final decision. In an ensemble, outputs of multiple predictors are combined in various ways. The purpose of combination models to build an integrated system able to overcome the drawbacks in each approach. The benefits of integrated models include robustness, improved performance and increased problem-solving capabilities [139].

Ensemble methods have two phases: the first phase is the production of the different models [140] such as Bagging and Boosting. Sometimes this phase is also recognized as Meta learning systems. For instance Vilalta, et al. [141] considered Stacked generalization and Boosting is a form of meta-learning, while Maclin and Opitz [142] indicated that Bagging and Boosting are two popular methods for creating accurate ensembles in addition to Džeroski and Ženko [143] used Stacking to build an ensemble of classifiers. Similarly Blachnik [144] presented Voting, Staking, Bagging and Boosting as examples of ensemble learning. Menahem, et al. [145] defined Meta-learning as the process of learning from basic classifiers (ensemble members) where the inputs of the meta-learner are the outputs of the ensemble-member classifiers". The second phase of an Ensemble method is the combination of the models [140]. The basic ensemble method and generalized ensemble method are the most popular techniques used in this phase. Figure 4.3 illustrates the proposed crude oil prediction model framework.

The research findings and architectural artifacts were evaluated based on appropriate validation techniques (See Section 4.6). Finally, choosing of the right methodology is very important because the success of the experimental results depends largely on a good selection of the appropriate tool. Section 4.4 explains the software tools used in the experimental stages.



**Figure 4.3** The proposed crude oil prediction model framework.

## 4.2. Data Set Description

The dataset for experiments are obtained cooperative by Faculty of Management and Economic Sciences, Sousse University, Tunisia. It consists of 3337 records as instances and 14 attributes to predict the West Texas Intermediate (WTI) as output. The data set was taken from different sources such as [146, 147]. Attributes are listed as below:

- [1] **Date (DT):** The daily data from 4 January 1999 to 10 October 2012. Dates<sup>7</sup> are converted to numeric form when the input file is read.
- [2] **West Texas Intermediate (WTI):** It is the most famous benchmark [17], and plays an important role as a reference point to determine the price, and it constitutes a crucial factor in the configuration of prices of all other commodities [148].

<sup>7</sup> Weka contains a mechanism for defining a date attribute to have a different format.



- [3] **Federal Fund Rate (FFR):** One of the most influential interest rates in the U.S. economy, because it effects on monetary and financial conditions, which in turn have an impact on fundamental aspects of the broad economy including employment, growth and inflation [149].
- [4] **Volatility Implied Equity Index (VIX):** Measures the contribution of the instability of the market.
- [5] **The Regional Standard and Poor's Equity Index (SPX):** Represent the market performance.
- [6] **New York Harbor Conventional Gasoline Spot Prices (GPNY):** As example to assesses oil products.
- [7] **US Gulf Coast Conventional Gasoline Spot Prices :( GPUS):** As example to assesses oil products.
- [8] **New York Harbor No. 2 Heating Oil Spot Price (HP):** As indication of seasonality in the energy market.
- [9] **Future Contracts 1 (FC1):** For WTI to maturity traded on NYMEX
- [10] **Future Contracts 2 (FC2):** For WTI to maturity traded on NYMEX
- [11] **Future Contracts 3 (FC3):** For WTI to maturity traded on NYMEX
- [12] **Future Contracts 4 (FC4):** For WTI to maturity traded on NYMEX.
- [13] **Exchange Rate (ER):** The price of oil and exchange rates of other currencies against the U.S. Dollar price.
- [14] **Gold Prices (GP):** Gold is that less volatile than crude oil and could reflect the real trend in the commodity market rather than the noise and gold used as the results of investors hedge against inflation caused by the oil price shock [150].

## 4.3. Data Preprocessing

Two methods as data preparation steps were implemented: first method is feature selection methods which is defined as a process of selecting a subset of features,  $d$ , out of the larger set of  $D$  features, which maximize the classification or prediction performance of a given procedure over all possible subset data. The second method is normalization, which shifts the instance values in specific and obviously means to represent information contained within the data and the data set [151].

### 4.3.1 Feature Selection Methods

10 different sub datasets, which were derived from the original dataset after implementing the several attribute selection algorithms, were formulated. For instance  $SBDS_1$  and  $SBDS_2$  are as a result of Correlation based Feature Selection (CFS) algorithm by evaluating the value of a group of attributes by concerning the individual predictive ability of each feature as well with the possibility of redundancy among the features with several search methods such as best-first, which keeps a list of all attribute subsets evaluated so far, sorted in order of the performance measure. Forward selection was used, where it starts with no attributes and add them one at a time and Backward, where it start with all the attributes and delete each one at a time, stops when the addition/deletion of any residual attributes results in a decrease in evaluation. In a case of one, begin with all the attributes or with none of them and this called bidirectional search method [56].

In  $SBDS_3$  and  $SBDS_4$ , Genetic algorithm, which is based on search processes on the principle of natural selection was utilized [56].  $SBDS_5$  is formulated after a Random search in the space of attribute subsets. Random search starts from a random point and reports the best subset found. If a start set is supplied, Random searches randomly for subsets, which is useful or better than the start point with the same or fewer attributes. [152].

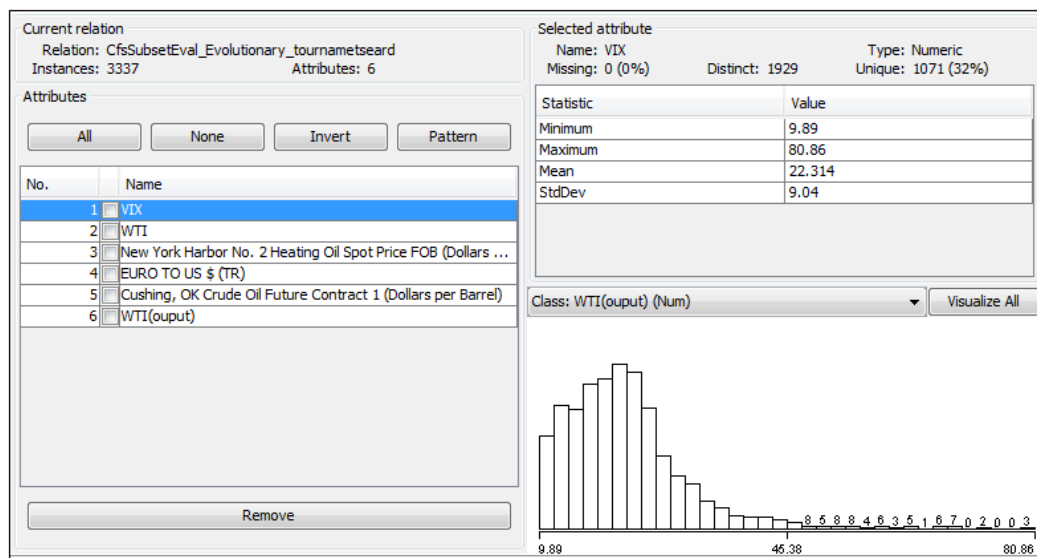
Forward selections with a limited number of  $k$  attributes were performed, based on the ranking using training data to decide, which attribute is added in each iteration of forward selection, and the test data is only used to evaluate the “best”,  $P$  best subsets of

a particular size. To determine the “optimal” subset size, the P scores on the test data for each subset size was averaged, and the size with the highest average is chose. Then, a final forward selection is performed on the complete data set to find a subset of that optimal size and SBDS<sub>6</sub> is created.

When Classifier subset evaluator algorithm was used, SBDS<sub>7</sub> by evaluating attribute subsets on training data or a separate hold out testing set using Support vector regression to estimate the 'merit' of a set of attributes with genetic search method.

SBDS<sub>8</sub> was obtained by using Relief attribute evaluation algorithm, which evaluates the quality of attributes according to the value of the given attribute for the near instance to each other and different predicted (class) value [153]. Ranker search method was used, which Ranked the list of attributes based on individual evaluation of each attribute [154].

SBDS<sub>9</sub> and SBDS<sub>10</sub> used wrapper algorithm, which evaluate attribute sets by using SMOreg algorithm. It is called wrapper because the learning algorithm is wrapped into a selection task [56]. The best-first search method in two directions: forward and backward respectively were implemented. Example of SBDS<sub>3</sub> in WEKA is shown in Figure 4.4, Table 4.1 illustrates the categories and attributes for each algorithm, and other sub-data are attached in Appendix A.



**Figure 4.4** SBDS3 using WEKA

**Table 4.1** Attribute selection methods and their features

Sub Dataset	Attributes evaluator	Search method	Attributes													
			1	2	3	4	5	6	7	8	9	10	11	12	13	14
SBDS <sub>1</sub>	Correlation based Feature Selection subset evaluator	Best-first Forward	WTI	SPX	FG1											
SBDS <sub>2</sub>	Correlation based Feature Selection subset evaluator	Best-first Backward	DT	VIX	WTI	SPX	GPNY	GPUS	HP	ER	FC1	FC2	FC3	FC4		
SBDS <sub>3</sub>	Correlation based Feature Selection subset evaluator	Genetic	VIX	WTI	GPNY	ER	FC1									
SBDS <sub>4</sub>	Correlation based Feature Selection subset evaluator	Genetic	WTI	GPNY	FC1											
SBDS <sub>5</sub>	Correlation based Feature Selection subset evaluator	Random	WTI	SPX	ER	FC1										
SBDS <sub>6</sub>	Correlation based Feature Selection subset evaluator	Subset Size Forward Selection	VIX	WTI	GPNY	FC1										
SBDS <sub>7</sub>	Classifier subset evaluator	Genetic SMOreg	VIX	WTI	SPX	GPNY	ER	FC1	FC2							
SBDS <sub>8</sub>	Relief attribute evaluation	Ranker	WTI	FC1	FC2	FC3	FC4	VIX	GPNY	GPUS	HP	GP	FFR	SPX	DT	ER
SBDS <sub>9</sub>	Wrapper subset evaluator (SMOreg)	Best-first Forward	WTI	GPUS												
SBDS <sub>10</sub>	Wrapper subset evaluator (SMOreg)	Best-first Backward	WTI	FC1												

### 4.3.2 Data Partition

To build the predictor it is important to use as much as possible of the data as training, and also uses hidden data as testing as much as possible to test its performance more comprehensively [155]. However, over-train occur if we use all data for training and the same data for testing because the predictor will learn well the available data and fails to hidden data [155]. So, for this reason it is important to have a separate data set on which to examine the final product.

There are various alternatives to recognize the training and testing split process such as cross-validation, bootstrap and holdout [56]. Hold- out was used because the others might be too time-consuming and the computational overhead, which a known as defects of cross-validation [156]. According to holdout method, dataset was divided randomly into two parts, one half of training and the other half for testing. Predictions from the model for the test set is compared with the actual values from the test set to evaluate the predictive accuracy. In the literature, it is becoming a public practice to use three data sets: part for training, the other part of validation, and the last one for testing. However, not all data sets are large enough to allow for a validation part to be cut out such as our data. Therefore, our experiments are restricted to training and testing.

It is common to hold out one-third of the data for testing and use the remaining two-thirds for training [56]. However, several researchers achieved good results with other divisions, for example Lai, et al. [157] created their model using 60% for training and 40% for testing while Yu, et al. [33] utilized 80% for training and 20% for testing. The effect of training and testing data was investigated by randomly splitting them as shown in Table 4.3. Several percentages were used to increase the opportunities for achieving better results. In the literature, there are also some studies conducted by using such divisions for training and test data [49] .

**Table 4.2** Training and testing percentages

Training	Testing	Label
90%	10%	(A)
80%	20%	(B)
70%	30%	(C)
60%	40%	(D)

### 4.3.3 Normalization

Most models work well with normalized data sets. The data were normalized using Eq. (4.1) by scaling the instance to the range between -1 and 1 to improve prediction accuracy and CPU processing time [158].

$$n_i = \frac{k_i - x_{\min}}{p_{\max} - x_{\min}} \quad (4.1)$$

Where  $n_i$  = normalized dataset  $k_i$  = raw dataset,  $x_{\min}$  = minimum value of the dataset and  $p_{\max}$  = maximum value of the dataset. Table 3.4 shows an example of a dataset, which is presented in normalized form.

**Table 4.3** Example of dataset after normalization process

VIX	WTI	SPX	GPNY	GPUS	HP	ER	FC1
-0.54121	-0.98447	0.241408	-0.96326	-0.97564	-0.96736	-0.24138	-0.98551
-0.59194	-0.97596	0.335104	-0.95437	-0.96955	-0.9642	-0.2069	-0.97431
-0.62266	-0.97491	0.347168	-0.95674	-0.97086	-0.95999	-0.24138	-0.97461
-0.56122	-0.96939	0.321937	-0.94963	-0.96564	-0.9542	-0.2069	-0.96909
-0.48683	-0.97715	0.267088	-0.9597	-0.97173	-0.9642	-0.17241	-0.9773
-0.43018	-0.98581	0.255587	-0.96563	-0.97912	-0.9721	-0.2069	-0.98596
-0.3493	-0.98731	0.2056	-0.96978	-0.98086	-0.97736	-0.2069	-0.98835
-0.4547	-0.98761	0.275528	-0.97393	-0.98347	-0.97999	-0.24138	-0.98895
-0.4547	-0.9888	0.295199	-0.97748	-0.98217	-0.97947	-0.17241	-0.9894
-0.47273	-0.99343	0.305597	-0.98578	-0.98565	-0.98421	-0.2069	-0.99343
-0.40736	-0.98402	0.257298	-0.97985	-0.98304	-0.97947	-0.2069	-0.98372
-0.37833	-0.98148	0.234859	-0.97748	-0.98173	-0.97684	-0.2069	-0.98029
-0.40144	-0.98462	0.254642	-0.97926	-0.98521	-0.97999	-0.2069	-0.98402

## 4.4 Software Tools

WEKA [159] is an open source tool produced by the University of Waikato (New Zealand). The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results. It also has a general API, so it is

easy to embed WEKA, like any other library, in any applications [160]. WEKA is very stable and powerful as it implements machine learning algorithms and data mining tasks such as data pre-processing, classification, regression, clustering, association rules and also includes visualization [161]. After loading the dataset by using Attribute-Relation File Format (ARFF) or CSV, the model was created and interpreted using four options: command-line interface (CLI), Explorer, Experimenter, and Knowledge flow. Finally it is maintainable, and modifiable, without depending on any particular institution or company [162].

MATLAB is one of most useful facilities for efficient exploratory data analysis and data mining. It has several advantages over other methods or languages such as plotting the data very easily by using the graphical interactive tools. MATLAB toolboxes are sets of specific functions that provides more specialized functionality. E.g. Excel link allows data to be written in a format recognized by Excel and Statistics Toolbox allows more specialized statistical manipulation of data [163].

The experiments are conducted on a computer with Intel(R) core i7 4700MQ, CPU 2.40- GHZ, RAM 16 GB, MATLAB 2013 and WEKA 3.7.9

## **4.5 Machine Learning Algorithms (ML)**

Numerous algorithms for learning tasks have been addressed in a large variety of applications such as classification, control of dynamic systems and prediction. In principle, most of the learning tasks can be reduced to each other. For example, a prediction problem can be reduced to a classification problem by defining classes for each of the possible predictions. Equally, a classification problem can be reduced to a prediction problem, etc. [134]. As mentioned in the Introduction Section, ML was grouped to direct models and combined models as follows:

### **4.5.1 Direct Prediction Models**

**A. Support Vector Regression (SMOreg):** In this research, the SVM regression was addressed, as mentioned earlier, which is an iterative optimization algorithm proposed by Smola and Schölkopf [65] for using SVR regression. SVM generalization

performance depends on a good setting of their parameters. RBF was used as a kernel function and  $C = 1$ , which indicates the complexity.

**B. Instance-Based Learning (IBL):** For Instance-Based K-nearest neighbors Learning (IBK), a default value of  $K=1$  based on cross-validation was selected.

**C. K-Star:**  $K^*$  represent another type of instance-based learners and use an entropy-based distance function [67].

**D. Isotonic Regression:** The isotonic regression, which picks the attribute that results in the lowest squared error was applied [68].

**E. Extra-Tree:** Random number of attributes at each node was selected, and a node with minimum sample size was splitted. To produce regression trees numerous times with the original learning sample was generated. The predictions of the trees are combined to get the final prediction by average.

**F. Reduced Error Pruning Tree (REPTree):** It is a fast decision tree learner. It builds a decision tree based on information gain or reducing the variance and prunes it by using reduced-error pruning (REP) with back over-fitting [71].

**G. Artificial Neural Networks (ANN):** Several types of supervised networks such as FFN, RBF, and dynamic (RCN) were used. FFN networks are most frequently used for prediction and pattern recognition. RBF provides an alternative, fast method for designing nonlinear feed-forward networks. Dynamic networks use memory and recurrent feedback connections to recognize spatial and temporal patterns in data. They are commonly used for time-series prediction, nonlinear dynamic system modeling, and control systems applications (Demuth et al., 2008, Demuth, Beale, 2000).

Neural networks with one and sometimes two hidden layers are widely used, for the large majority of problems and have performed very well (Panchal et al., 2011). Increasing the number of hidden layers are extremely hard to train, increases computation time and may lead to over-fitting which leads to poor out-of-sample predicting performance for this reason one hidden layer in this work was used. One of the most



important characteristics of a network is the number of neurons in the hidden layer (s). If an insufficient number of neurons are used, the network will be unable to model complex data, and the resulting fit will be poor. Despite its importance, there is no formula for selecting the optimum number of hidden neurons. Therefore, scholars depend on experimentation.

40-45-50-55-60 neurons in the hidden layer based on trial and error approach were used. In most cases, the literature suggests the use of a trial-and-error approach to configuring network parameters, where the user sets the performance goal. For instance, (Rene et al., 2013) used trial-and-error approach to determine network parameters. Selection of the training algorithm, which is suitable for our problem, depends on many factors such as the complexity of the problem and the number of inputs and others (Demuth et al., 2008). The Levenberg-Marquardt (LM), Bayesian regularization (BR) and BFGS Quasi-Newton (BFG-QN) algorithms because they are commonly used for regression problems and is easy to compare with other algorithms. The transfer function is tan-sigmoidal for the hidden layer and pure linear function in the output layer were applied, the maximum number of epochs is set to 1000 and the training goal is set to 0. Experiments and results for NNs prediction models are discussed in Section 5.1.2 of Chapter 5.

## **4.5.2 Combined Prediction Models**

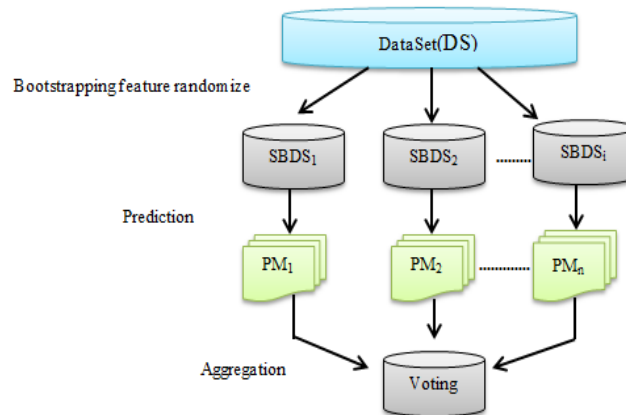
Since the last few years the researches in data mining and ML are transformed to another direction called combination techniques, which seeks to build combined learning systems for classification and regression has stronger generalization and performance than their single components [135].

### **A. Meta Prediction Models**

Meta-learning involving several algorithms. A popular set of this technique namely Bagging, Random subspace, Ensemble selection, Voting and Stacking. Meta depend on their mechanism were grouped into two parts: Bagging, Random subspace which separate data into subparts each part train by same predictor. Another part in-

cludes Ensemble selection, Voting, and Stacking, which providing same input to a number of predictors and combine their output using a given decision logic. As mentioned previously, Meta learning helps to create optimal predictive models and reuse previous experience from analysis of other problems, such that the modified learner is better than the original learner at learning from additional experience.

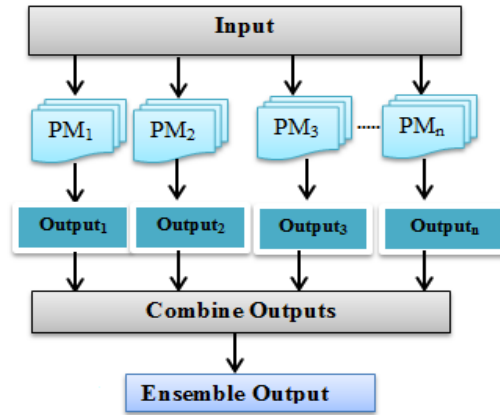
**1- Bagging:** Breiman [164] introduced bagging methods, which is basically a multiple predictor combined to get the final result through bootstrap replicates. As illustrated in Figure 4.5, the subset  $SBDS_i$  was generated by replacing the original data set DS (bootstrapping) many times, then compute a sequence predictor  $P_i$  by classifier  $C_i$  and then the same machine learning scheme was used for each sub-dataset  $SBDS_i$ . Finally, aggregate the results by voting (averaging) to access the last from  $P_{final}$ . This method is characterized that it could enhance performance [165] and reduce variance to improve generalization [166] [167].



**Figure 4.5** Bagging ensemble methods

**2- The Random Subspace Method (RSM):** RSM is the combining technique proposed by [168]. An  $r$ -dimensional random subspace  $X^b = (X_1^b, X_2^b, \dots, X_n^b)$  from the original  $p$ -dimensional feature space  $X$ , where  $r < p$  was selected. Then one constructs predictor in the random subspaces  $X^b$  and combines them by simple majority voting in the final decision rule. This process was repeated for  $b = 1, 2, 3, 4, 5$  according to number of direct predictors.

**3- Ensemble Selection Algorithm** [169]: The first step a “model library” was created. This library should be a large and diverse set of direct prediction models and the presence of a number of Prediction Models (PM) are denoted by  $PM_1, PM_2, PM_3, PM_n$  with different parameters. The second step is to combine the outputs of these models from our library with the Ensemble Selection algorithm. To prevent over-fitting, an ensemble have been with only one model in it. Then, models was added one at a time to our ensemble, to figure out which model to add, each time: individually average the predictions of each model from the library currently being considered with the current ensemble. Then pick the model that provided the most performance improvement Figure 4.6 explains these steps.



**Figure 4.6** Ensemble method framework

**4- Voting:** The simplest kind of ensemble is the way of aggregating a collection of prediction values with each base level giving different voting power for its prediction. The final prediction obtains the highest number of votes. Voting includes the weighted average (of each base classifier holds) when using regression problem and majority voting when doing classification, the weighted-majority output is:

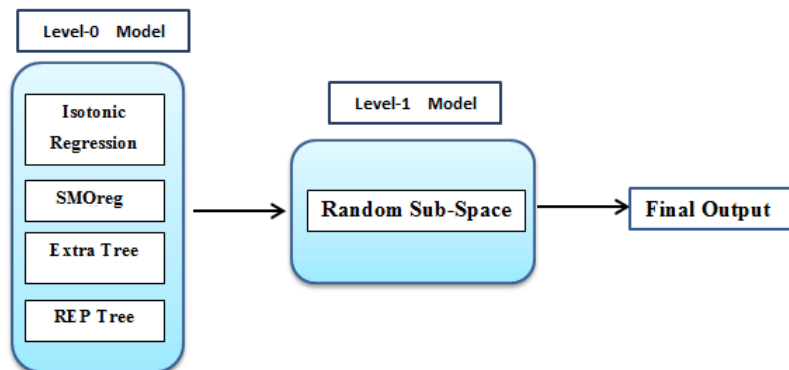
$$\text{Argmax} \left( \sum_{j=1}^k P_j(x), w_j \right) \quad (4.2)$$

$P_i(x)$  is the results of the prediction of  $i^{\text{th}}$  prediction model and  $P_i(x, w)$  is indicator function defined as:

$$p_i(x, w) = \begin{cases} 1 & x = w \\ 0 & \text{otherwise} \end{cases} \quad (4.3)$$

Majority voting has some benefits that it does not require any additional complex computation and any previous knowledge. However, this approach leads to the result that it is difficult to analyze and interpret. The second strategy is un-weighted, which gives some predictor higher weight if they achieve more accuracy than others (the winner is the one with the most number of votes) [170, 171].

**5- Stacking:** Another popular approach to combine predictors is called stacking or stacked generalization. It is a function that depends on the Meta learner (level-1 model), which concerns with the combination of the predictions of the numerous predictors generated by using different learning processes on a direct dataset, and base classifiers (level-0 models) to obtain the final prediction [143], according to which features and algorithms are used in the Meta. According to Figure 4.7, four direct predictors were used in level-0 model as inputs to Random subspace in level-1 and training data for level-1 was generated by using cross-validation model. Experiments for Meta prediction model is illustrated in Chapter Six.



**Figure 4.7** Stacking structure for prediction crude oil prices

## B. Hybrid Prediction Model

Hybrid intelligence techniques are a combination of multiple methods to build an efficient solution to deal with a particular problem and in recent years it is considered as a powerful tool to improve the accuracy [136]. On the other hand, the purpose of this research is not just helping to predict the market, but also to help make decision as investors use their expectation skill based on ‘rumors’ to hedge the price in the market.

**Adaptive Neuro Fuzzy Inference System (ANFIS):** ANFIS is one distinct example of the hybrid system. It is a good model to explore and propose a decision making system by extracts information (input) and compute it in the system automatically, thus producing a decision (output) based on information from the extracted crude oil market’s rules. To design ANFIS model these steps were followed:  
Specify number and type of membership functions. Different types of membership functions (MF) include Trapezoidal, Guassian, Gbell and Triangular shapes were tested for the inputs and output. A two (Trapezoidal) MF for each input variable type ANFIS resulted in high accurate modeling and minimal training time. Table 4.4 shows results of examples of several MF with SBDS<sub>1</sub>.

**Table 4.4** ANFIS with different type of membership functions

Data Sub Data set	A		B	
	2 MF	3MF	2MF	3MF
SBDS <sub>1</sub>	<b>Trapezoidal</b>			
	1.23522E-05	6.04313E-05	1.56636E-05	1.31261E-04
	<b>Gaussian</b>			
	7.86182E-04	3.00735E-05	1.70752E-04	1.27876E-04
	<b>Gbell</b>			
	9.10802E-05	1.31191E-03	2.29898E-04	8.87822E-04
	<b>Triangular</b>			
	7.4537E-01	5.4370E-01	3.232E-02	6.275E-02

- Generate initial TSK based FIS model (grid partitioning).
- Configure ANFIS system and adjust the membership functions and consequent parameters using the hybrid learning method which combining the least squares method and the gradient descent method for 100 epochs.
- The training process is terminated after 100 epochs. Simulation of ANFIS model is illustrated in Chapter Six (Tables 6.8, 6.9 and 6.10). Appendix C lists the sample of ANFIS program code:

### **C. Ensemble Prediction Model**

In an ensemble, outputs of multiple predictors are combined in various ways. Ensemble methods are one of the latest techniques that promises results more effective in different applications such as pattern recognition [172], machine learning, data mining [173] and medical applications [174].

Ensemble methods have two phases: the first phase is the production of the different models [140] using Bagging and Boosting. Sometime this phase is also recognized as Meta learning, for instance, Vilalta, et al. [141] considered that Stacked generalization and Boosting a form of meta-learning while Maclin and Opitz [142] indicated that Bagging and Boosting are two popular methods for creating accurate ensembles in addition to Džeroski and Ženko [143] used Stacking to ensemble of classifiers. Similarly Blachnik [144] presented Voting, Staking, Bagging and Boosting as examples of ensemble learning. Menahem, et al. [145] defined Meta based on Ensemble “Meta-learning is the process of learning from basic classifiers (ensemble members); the inputs of the meta-learner are the outputs of the ensemble-member classifiers”.

The second phase of an Ensemble method is the combination of the models [140]. The basic ensemble method and generalized ensemble method are the most popular techniques used in this phase. In this research, the basic ensemble method (BEM) as defined by:

$$F_{BEM} = \frac{1}{n} \sum_{i=1}^n F_i(X) \quad (4.4)$$

Where  $F_i(x)$  is the output produced by the different models. This approach by itself can lead to improved performance, but does not take into account the fact that some networks may be more accurate than others. It has the advantage of being easy to understand and implement and is often found not to increase the expected error [175].

The ensemble method depends on the behavior that a collection of predictor such as machine learning algorithms (neural network, support vector machine, decision trees and so on) can do better than the individual approaches. Predictors are combined through some weighted average or weighted combination. The generalized ensemble method find weights for each output that minimizes the Root Mean Square Error (RMSE) or Mean Absolute Error (MAE) of the ensemble. The general ensemble model (GEM) is defined by:

$$F_{GEM} = \sum_{i=1}^n \alpha F_i(X) \quad (4.5)$$

$$\sum_{i=1}^m \alpha = 1 \quad (4.6)$$

Where  $\alpha F_i(x)$  are chosen to minimize the MAE between the outputs and the desired values. Finding the optimal values of  $\alpha$  is not an easy task. A Particle swarm optimization (PSO) method was used to determine the optimal weights. Section 6.3 of Chapter 6 illustrates the Ensemble prediction experiments

## 4.6 Evaluation and Validation Approaches

There are a number of criteria to evaluate the result of learning, related to out-of-sample performance prediction and to assess information enterprise architecture. These are involved in various aspects of measuring techniques and questioning techniques to

guarantee final performance of the model that are “significant”, to predict crude oil prices.

#### 4.6.1 Measuring Techniques:

To judge the prediction performances and evaluate the accuracy of prediction, there are two basic criteria: the Mean Absolute Error (MAE) and Root Mean Square error (RMSE). The smaller the value of the evaluation indexes, the higher the performance of the algorithm. Willmott and Matsuura [176] indicated that MAE is a more natural measure of average error, it is unambiguous and comparisons of average model performance error should be based on MAE. e.g. [127], [34] and [177].

$$\mathbf{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \tilde{y}_i| \quad (4.7)$$

Before introducing the RMSE it is useful to introduce the Mean Squared Error:

$$\mathbf{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2 \quad (4.8)$$

RMSE is a widely used measure to calculate differences between the values predicted by a model or a predictor and the values actually observed e.g. [127],[33], and [126] . RMSE is defined by:

$$\mathbf{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2} \quad (4.9)$$

$$\mathbf{RMSE} = \sqrt{\mathbf{MSE}} \quad (4.10)$$

Where  $y_i$  and  $\tilde{y}_i$  are the predicted and actual values of crude oil price at time  $i$ , respectively and  $n$  is the numbers of predictions.

#### 4.6.2 Questioning Techniques:

The study adopted multiple data collection techniques, which included a questionnaire, interviewing of experts from the Oil Industry/Academia and document



/literature review with experts in the oil industry as a means of building and validating the information architecture for the Oil Industry. For the design of information framework for crude oil prices, Government and private oil sector in a group of different countries were chosen such as the Saudi Arabia, as one of the petroleum-exporting countries with a large reserve capacity, Tunisia, the first Arab country which erupted the revolutions of the Arab Spring and Sudan, which has suffered from the separation and division of the country into a republic of Sudan and South Sudan. It is worth mentioning that the country produces about 500,000 barrels of oil a day, about 75% is located in South Sudan [178]. Questionnaire item and document analysis are illustrated in Appendix D.

## **4.7 Summary**

The purpose of this Chapter is to present and discuss the approach and methods of the research. Hence, it covers the methodological aspects that have guided the present work. It starts with an introduction, which gives an overview of the methodology of work then data description and preprocessing approach. Several learning algorithms used this research are described, which are mainly used to determine the performance of proposed model. Finally the methodology used for evaluation of the research process and results are also presented.

# CHAPTER FIVE

## 5. DIRECT PREDICTION MODELS

### Overview

In the previous Chapter, a vast array of machine learning approaches were studied. All are sound, robust techniques that are extremely applicable to practical prediction problems. To guarantee build successful machine learning model in predicting crude oil prices, practical steps for selecting best learning algorithm must be applied by running it over our data. In this Chapter, the prediction process was modeled and the direct prediction models were analyzed, which includes isotonic regression, SMOreg, Kstar, IBK, ExtraTree, REPTree and several types of NNs includes FFN, RCN and RBF. Furthermore, the comparison of these algorithms is presented based on a root mean squared error (RMSE) and mean absolute error (MAE) to find out the best suitable approaches. The detailed finding of machine learning experiments and trend analysis on prediction crude oil price is presented in (Gabralla, Abraham, 2014, Gabralla et al., 2013, Gabralla et al., 2015).

### 5.1 Experimental Results

The purpose of this Section is to measure the performance of direct prediction models. On the one hand, ten sub datasets ( $SBDS_1$ ,  $SBDS_2$ ,  $SBDS_3$ ,  $SBDS_4$ ,  $SBDS_5$ ,  $SBDS_6$ ,  $SBDS_7$ ,  $SBDS_8$ ,  $SBDS_9$  and  $SBDS_{10}$ ) were used, which is derived from the original dataset by using several attribute selection algorithms mentioned in Table 4.1 and on the other hand four group (A-B-C-D) were used, which contain different training and testing percentages as displayed in Table 4.2. It is worth mentioning that the training and testing experiments were repeated ten times with different random sample for each sub dataset to guarantee that the full dataset represented in the training and testing sets in the correct way and the error rates on the different iterations are averaged to yield an overall error rate. The algorithms that did not perform well for all the training and testing and for the different attributes were excluded from the next step of the combination

process (Chapter six). To simplify the extensive list of experiments, the experiments of direct prediction models were classified in two groups as follows:

### 5.1.1 First Phase Experiments and Results

In this Section, six direct algorithms, namely Isotonic Regression, SMOReg, Kstar, IBK, ExtraTree and REPTree were implemented. Discussions of the algorithms are thoroughly documented in Chapter 2. Table 5.1.a and Table 5.1.b reports the empirical results illustrating MAE and Table 5.2 .a and Table 5.2. b presents the RMSE for the six algorithms. Figure 5.1 shows the performance of six algorithms in order to determine best approaches.

**Table 5.1.a** MAE for first phase experiment using sub dataset from (SBDS<sub>1</sub> to SBDS<sub>5</sub>)

Prediction Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>5</sub>
Isotonic Regression	A	2.220E-02	2.220E-02	2.200E-02	2.220E-02	2.220E-02
	B	2.420E-02	2.420E-02	2.400E-02	2.420E-02	2.420E-02
	C	2.780E-02	2.780E-02	2.800E-02	2.780E-02	2.780E-02
	D	3.250E-02	3.250E-02	3.300E-02	3.250E-02	3.250E-02
SMOReg	A	4.400E-02	6.600E-02	2.860E-02	4.940E-02	4.850E-02
	B	3.930E-02	7.010E-02	3.500E-02	4.770E-02	4.770E-02
	C	4.070E-02	6.990E-02	3.990E-02	5.200E-02	4.640E-02
	D	4.440E-02	6.990E-02	3.900E-02	5.160E-02	5.580E-02
Kstar	A	9.720E-01	4.560E-01	9.080E-01	7.100E-01	8.440E-01
	B	9.760E-01	4.690E-01	9.010E-01	7.110E-01	8.530E-01
	C	9.880E-01	4.870E-01	9.000E-01	7.230E-01	8.610E-01
	D	1.000E+00	5.160E-01	9.130E-01	7.390E-01	8.700E-01
IBk	A	3.740E-01	6.320E-01	2.520E-01	6.230E-01	6.040E-01
	B	3.960E-01	6.620E-01	2.670E-01	6.480E-01	6.290E-01
	C	4.180E-01	6.910E-01	2.820E-01	6.770E-01	6.550E-01
	D	4.470E-01	7.310E-01	3.050E-01	7.190E-01	6.910E-01
ExtraTree	A	8.320E-02	1.600E-01	1.010E-01	1.130E-01	1.000E-01
	B	8.670E-02	1.760E-01	1.090E-01	1.180E-01	1.100E-01
	C	1.040E-01	1.760E-01	1.220E-01	1.190E-01	1.090E-01
	D	1.190E-01	1.950E-01	1.320E-01	1.490E-01	1.310E-01
REPTree	A	8.320E-02	9.700E-02	8.300E-02	9.350E-02	8.420E-02
	B	9.670E-02	1.140E-01	9.700E-02	1.040E-01	9.760E-02
	C	2.220E-02	2.220E-02	2.200E-02	2.220E-02	2.220E-02
	D	2.420E-02	2.420E-02	2.400E-02	2.420E-02	2.420E-02

**Table 5.1.b** MAE for first phase experiment using sub dataset from (SBDS<sub>6</sub> to SBDS<sub>10</sub>)

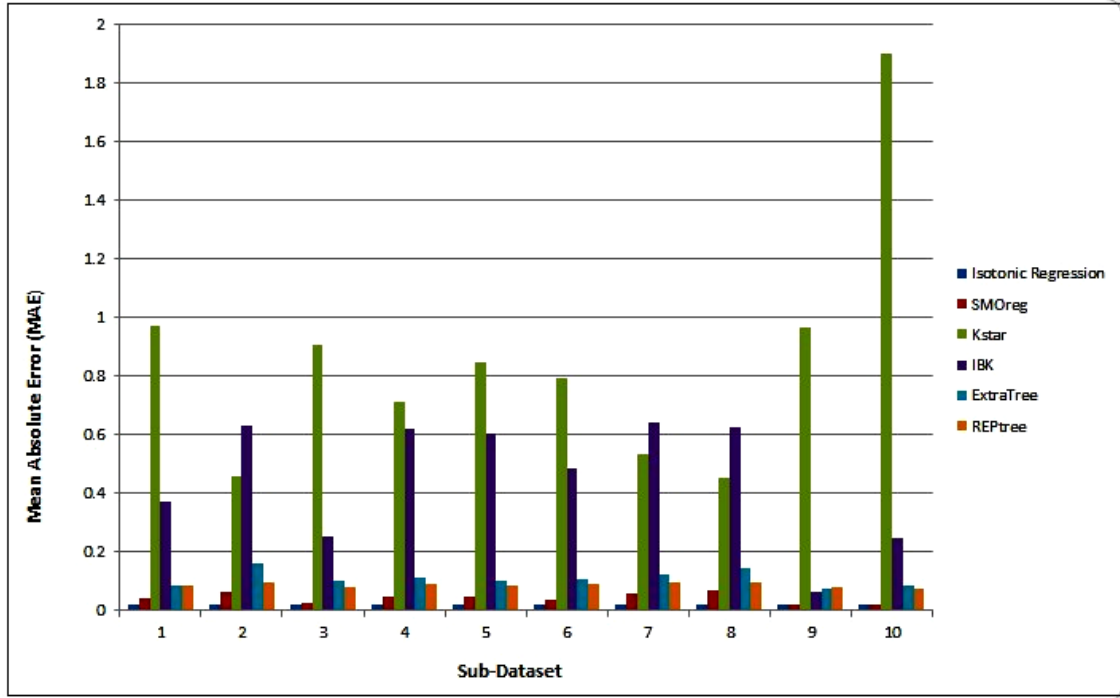
Prediction Model	Data	SBDS <sub>6</sub>	SBDS <sub>7</sub>	SBDS <sub>8</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
<b>Isotonic Regression</b>	<b>A</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>
	<b>B</b>	2.420E-02	2.420E-02	2.420E-02	2.420E-02	2.420E-02
	<b>C</b>	2.780E-02	2.780E-02	2.780E-02	2.780E-02	2.780E-02
	<b>D</b>	3.250E-02	3.250E-02	3.250E-02	3.250E-02	3.250E-02
<b>SMOreg</b>	<b>A</b>	3.770E-02	5.730E-02	6.790E-02	2.230E-02	2.320E-02
	<b>B</b>	3.840E-02	5.980E-02	7.060E-02	2.630E-02	2.440E-02
	<b>C</b>	3.680E-02	6.310E-02	7.320E-02	2.860E-02	2.650E-02
	<b>D</b>	4.090E-02	6.310E-02	7.520E-02	3.100E-02	<b>2.210E-02</b>
<b>Kstar</b>	<b>A</b>	7.940E-01	5.350E-01	<b>4.550E-01</b>	9.650E-01	1.900E+00
	<b>B</b>	7.920E-01	5.470E-01	4.740E-01	9.600E-01	1.900E+00
	<b>C</b>	7.980E-01	5.620E-01	4.910E-01	9.730E-01	1.910E+00
	<b>D</b>	8.110E-01	5.870E-01	5.200E-01	9.830E-01	1.930E+00
<b>IBk</b>	<b>A</b>	4.830E-01	6.420E-01	6.270E-01	<b>6.660E-02</b>	2.500E-01
	<b>B</b>	5.000E-01	6.720E-01	6.560E-01	6.910E-02	2.650E-01
	<b>C</b>	5.240E-01	7.020E-01	6.870E-01	7.350E-02	2.810E-01
	<b>D</b>	5.570E-01	7.450E-01	7.320E-01	8.110E-02	3.070E-01
<b>ExtraTree</b>	<b>A</b>	1.070E-01	1.230E-01	1.460E-01	<b>7.730E-02</b>	8.510E-02
	<b>B</b>	1.230E-01	1.320E-01	1.620E-01	8.290E-02	1.000E-01
	<b>C</b>	1.180E-01	1.540E-01	1.680E-01	9.400E-02	1.090E-01
	<b>D</b>	1.450E-01	1.730E-01	1.900E-01	1.050E-01	1.310E-01
<b>REPrece</b>	<b>A</b>	9.240E-02	9.570E-02	9.450E-02	8.140E-02	<b>7.540E-02</b>
	<b>B</b>	1.030E-01	1.090E-01	1.110E-01	9.390E-02	8.610E-02
	<b>C</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>
	<b>D</b>	2.420E-02	2.420E-02	2.420E-02	2.420E-02	2.420E-02

**Table 5.2.a** RMSE for first phase experiment using sub dataset from (SBDS<sub>1</sub> to SBDS<sub>5</sub>)

Prediction Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>5</sub>
<b>Isotonic Regression</b>	<b>A</b>	5.270E-02	5.270E-02	5.270E-02	5.270E-02	5.270E-02
	<b>B</b>	5.340E-02	5.340E-02	5.340E-02	5.340E-02	5.340E-02
	<b>C</b>	6.160E-02	6.160E-02	6.160E-02	6.160E-02	6.160E-02
	<b>D</b>	7.120E-02	7.120E-02	7.120E-02	7.120E-02	7.120E-02
<b>SMOreg</b>	<b>A</b>	6.710E-02	9.190E-02	5.320E-02	7.000E-02	7.060E-02
	<b>B</b>	6.990E-02	1.000E-01	6.620E-02	7.550E-02	7.540E-02
	<b>C</b>	7.130E-02	1.000E-01	6.990E-02	7.950E-02	7.430E-02
	<b>D</b>	7.970E-02	1.050E-01	7.470E-02	8.440E-02	8.750E-02
<b>Kstar</b>	<b>A</b>	1.850E+00	6.740E-01	1.640E+00	1.040E+00	1.310E+00
	<b>B</b>	1.890E+00	7.070E-01	1.660E+00	1.040E+00	1.350E+00
	<b>C</b>	1.950E+00	7.450E-01	1.690E+00	1.060E+00	1.390E+00
	<b>D</b>	1.960E+00	8.160E-01	1.710E+00	1.110E+00	1.390E+00
<b>IBk</b>	<b>A</b>	5.510E-01	8.760E-01	3.820E-01	8.760E-01	8.170E-01
	<b>B</b>	5.900E-01	9.500E-01	4.400E-01	9.330E-01	8.640E-01
	<b>C</b>	6.140E-01	9.930E-01	4.560E-01	9.770E-01	9.000E-01
	<b>D</b>	6.580E-01	1.070E+00	5.040E-01	1.050E+00	9.460E-01
<b>ExtraTree</b>	<b>A</b>	2.150E-01	3.680E-01	2.750E-01	2.630E-01	2.410E-01
	<b>B</b>	2.150E-01	3.680E-01	2.750E-01	2.630E-01	2.410E-01
	<b>C</b>	2.680E-01	4.170E-01	3.050E-01	2.980E-01	2.800E-01
	<b>D</b>	2.790E-01	4.660E-01	3.230E-01	3.770E-01	3.270E-01
<b>REPrece</b>	<b>A</b>	1.820E-01	2.290E-01	1.780E-01	2.380E-01	1.830E-01
	<b>B</b>	2.360E-01	2.930E-01	2.360E-01	2.750E-01	2.370E-01
	<b>C</b>	2.470E-01	3.060E-01	2.510E-01	2.810E-01	2.490E-01
	<b>D</b>	3.010E-01	3.670E-01	2.960E-01	3.560E-01	3.030E-01

**Table 5.2.b** RMSE for first phase experiment using sub-dataset from SBDS<sub>6</sub> to SBDS<sub>10</sub>

Prediction Model	Data	SBDS <sub>6</sub>	SBDS <sub>7</sub>	SBDS <sub>8</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
<b>Isotonic Regression</b>	<b>A</b>	<b>5.270E-02</b>	<b>5.270E-02</b>	<b>5.270E-02</b>	<b>5.270E-02</b>	<b>5.270E-02</b>
	<b>B</b>	5.340E-02	5.340E-02	5.340E-02	5.340E-02	5.340E-02
	<b>C</b>	6.160E-02	6.160E-02	6.160E-02	6.160E-02	6.160E-02
	<b>D</b>	7.120E-02	7.120E-02	7.120E-02	7.120E-02	7.120E-02
<b>SMOreg</b>	<b>A</b>	6.100E-02	8.060E-02	9.270E-02	4.490E-02	3.100E-02
	<b>B</b>	6.990E-02	8.930E-02	1.010E-01	5.630E-02	3.180E-02
	<b>C</b>	6.910E-02	9.100E-02	1.040E-01	5.850E-02	3.530E-02
	<b>D</b>	7.860E-02	9.480E-02	1.100E-01	6.630E-02	<b>3.030E-02</b>
<b>Kstar</b>	<b>A</b>	1.210E+00	7.880E-01	<b>6.710E-01</b>	2.200E+00	3.350E+00
	<b>B</b>	1.220E+00	8.050E-01	7.160E-01	2.250E+00	3.410E+00
	<b>C</b>	1.240E+00	8.320E-01	7.520E-01	2.310E+00	3.470E+00
	<b>D</b>	1.290E+00	9.020E-01	8.240E-01	2.310E+00	3.500E+00
<b>IBk</b>	<b>A</b>	7.200E-01	8.900E-01	8.740E-01	<b>1.280E-01</b>	4.110E-01
	<b>B</b>	7.650E-01	9.560E-01	9.370E-01	1.400E-01	4.380E-01
	<b>C</b>	8.090E-01	1.010E+00	9.850E-01	1.430E-01	4.560E-01
	<b>D</b>	8.790E-01	1.080E+00	1.070E+00	1.660E-01	4.930E-01
<b>ExtraTree</b>	<b>A</b>	2.700E-01	2.750E-01	3.470E-01	<b>1.940E-01</b>	2.340E-01
	<b>B</b>	2.700E-01	2.750E-01	3.470E-01	1.940E-01	2.340E-01
	<b>C</b>	2.880E-01	4.110E-01	3.810E-01	2.190E-01	3.010E-01
	<b>D</b>	3.900E-01	4.370E-01	4.490E-01	2.410E-01	3.670E-01
<b>REPre</b>	<b>A</b>	2.380E-01	2.470E-01	2.240E-01	1.760E-01	<b>1.660E-01</b>
	<b>B</b>	2.740E-01	2.820E-01	2.780E-01	2.320E-01	2.030E-01
	<b>C</b>	2.790E-01	2.930E-01	3.190E-01	2.400E-01	2.350E-01
	<b>D</b>	3.530E-01	3.610E-01	3.630E-01	2.930E-01	2.650E-01



**Figure 5.1** MAE for six prediction models with 10 sub-data sets

As illustrated in Figure 5.1, the K Star algorithm did not perform well for all the training and testing and for the different attributes. For this reason, it was excluded from the combination process in the next Chapter.

Time needed by the system to learn is another important criteria that may be considered in model selection [179], therefore in this Section comparisons between prediction models based on time are presented. According to Table 5.3, Isotonic regression and Extra Tree consumed less time, while Kstar failed to achieve suitable time when compared with other models and SMOreg succeeded to achieve less error but consumed long time. The recorded time in this Table represents the time required to by each algorithm for all 10 sub-data sets.

**Table 5.3** Time schedule for direct prediction models

Prediction Model	Data	Time (hour: min: sec)
Isotonic Regression	A	00:00:04
	B	00:00:04
	C	00:00:04
	D	00:00:04
SMOreg	A	00:10:15
	B	00:09:24
	C	00:07:58
	D	00:05:14
Kstar	A	00:25:19
	B	00:43:27
	C	00:54:58
	D	01:03:02
IBK	A	00:00:11
	B	00:00:19
	C	00:00:22
	D	00:00:25
ExtraTree	A	00:00:05
	B	00:00:04
	C	00:00:04
	D	00:00:04
REPtree	A	00:00:08
	B	00:00:11
	C	00:00:14
	D	00:00:15

Table 5.4 summarizes the important results as follows: SMOreg, Isotonic Regression, REPtree and ExtraTree achieved less MAE  $2.21E-02$ ,  $2.22E-02$ ,  $7.54E-02$  and  $7.73E-02$  respectively and IBK accomplished good results  $6.66E-02$  with sub dataset 9 only.

Training and testing (A) which represent 90% training and 10 testing achieved best results with most algorithms also the best results focused in SBDS<sub>9</sub> and SBDS<sub>10</sub> and poor results were posted in (SBDS<sub>5</sub>, SBDS<sub>7</sub> and SBDS<sub>8</sub>) thus were removed from further experiments. SMOreg surpassed other algorithms in RMSE ( $3.03E-02$ ). From the above discussions it is found that Kstar was not appropriate to solve the problem.



**Table 5.4** Summary of the results for direct prediction models

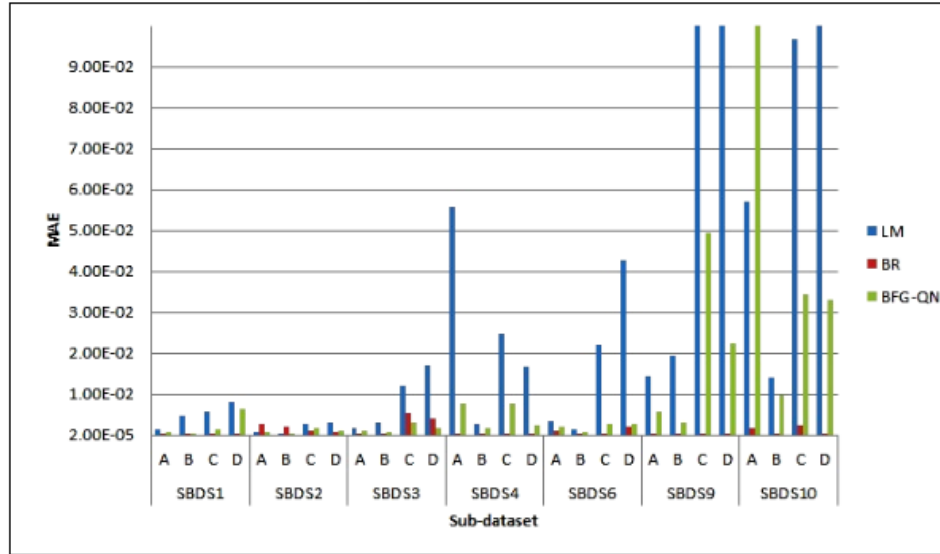
Prediction Model	Data	MAE	RMSE	Sub-dataset	Time (for 10-subdatasets)
<b>Isotonic Regression</b>	A	2.220E-02	5.270E-02	All sub-dataset	<b>00:00:04</b>
<b>SMOreg</b>	D	<b>2.210E-02</b>	<b>3.030E-02</b>	SBDS <sub>10</sub>	00:05:14
<b>Kstar</b>	A	4.546E-01	6.706E-01	SBDS <sub>8</sub>	00:25:19
<b>IBk</b>	A	6.660E-02	1.284E-01	SBDS <sub>9</sub>	00:00.11
<b>ExtraTree</b>	A	7.730E-02	1.936E-01	SBDS <sub>9</sub>	00:00.05
<b>REPre</b>	A	7.540E-02	1.661E-01	SBDS <sub>10</sub>	00:00.08

### 5.1.2 Second Phase Experiments and Results

Numerous important characteristics of neural networks make them proper and valuable for data mining and machine learning so the objective of this Section is to provide a variety of the training and testing percentages with a set of different inputs using several kinds of neural networks to get high accuracy for the model. Neural network experiments are accomplished in MATLAB. The architecture design of the network is described in 4.5.1.

#### Feed Forward Neural Network (FFN)

According to Table 5.5, FFN utilized the 7 sub-datasets, which were selected as the best sub-dataset based on previous experiments. The best results were obtained when using the Bayesian regularization (BR) back-propagation method with 80% training and 20% testing, and (SBDS<sub>1</sub>) achieved a MAE= 3.843E-05 with 90% training and 10% testing using 45 neurons. Figure 5.3 shows a comparison between the training algorithms for 7 sub-datasets and 4 groups of training and testing illustrating the superiority of BR. the performance using MAE and RMSE were measured.



**Figure 5.2** Comparison between training algorithms

### Recurrent Neural Network (RCN)

RCN was implemented using one hidden layer with 10 neurons and used three training algorithms: Levenberg –Marquardt (LM), Bayesian regularization (BR) and BFGS Quasi-Newton (BFG-QN). Bayesian regularization method outperformed other algorithms by 51.85%. It is noted from Table 5.6, for all the sub-datasets (80% training and 20% testing) is the best (shaded area), except in sub-dataset (SBDS<sub>6</sub>) 90% training and 10% testing is the best. On the other hand, the lowest value of the error is 3.941 E-05 when using 90% training and 10% testing with sub-dataset (SBDS<sub>6</sub>).

### Radial Basis Function Network (RBF)

The network was constructed until it reached a maximum number of neurons or the sum-squared error falls beneath an error goal. Table 5.7 shows the results obtained using the seven sub-datasets and different number of neurons: 40, 45, 50, 55 and 60. The shaded area indicates the best results when using 60 neurons in few sub-datasets then followed by 55 neurons. According to the percentage of training and testing sub-dataset (SBDS<sub>1</sub>– SBDS<sub>3</sub>-SBDS<sub>4</sub>- SBSD<sub>9</sub> and SBDS<sub>10</sub>) achieved the best results with 80% training and 20% testing. The best results over all sub-datasets is an MAE of 2.206 E-05 in SBDS<sub>9</sub> with 80% training and 20 % testing using 45 neurons.

**Table 5.5** Performance of FFN

Sub-datasets	Data	Mean Absolute Error			Hidden layer neurons
		LM	BR	BFG-QN	
SBDS <sub>1</sub>	A	1.48352E-03	<b>3.84294E-05</b>	8.00000E-04	45
	B	4.81400E-03	<b>1.35000E-04</b>	4.00000E-04	45
	C	5.77900E-03	<b>3.55000E-04</b>	1.50000E-03	45
	D	8.15200E-03	<b>4.38000E-04</b>	6.30000E-03	40
SBDS <sub>2</sub>	A	9.17000E-04	2.83700E-03	<b>9.00000E-04</b>	40
	B	4.48000E-04	2.18100E-03	<b>4.00000E-04</b>	40
	C	2.74700E-03	<b>1.02200E-03</b>	1.90000E-03	45
	D	3.15700E-03	<b>7.69000E-04</b>	1.00000E-03	60
SBDS <sub>3</sub>	A	1.83600E-03	<b>1.26594E-04</b>	1.10000E-03	50
	B	3.00400E-03	<b>1.24897E-04</b>	9.00000E-04	50
	C	1.21500E-02	5.31100E-03	<b>3.10000E-03</b>	45
	D	1.69940E-02	4.06200E-03	<b>1.80000E-03</b>	45
SBDS <sub>4</sub>	A	5.58850E-02	<b>5.74155E-05</b>	7.80000E-03	50
	B	2.86700E-03	<b>6.45000E-05</b>	1.60000E-03	50
	C	2.48740E-02	<b>3.04484E-04</b>	7.70000E-03	50
	D	1.67079E-02	<b>1.94000E-04</b>	2.40000E-03	50
SBDS <sub>6</sub>	A	3.50300E-03	<b>9.65000E-04</b>	2.00000E-03	55
	B	1.40900E-03	<b>6.80000E-05</b>	8.00000E-04	60
	C	2.19900E-02	<b>3.73327E-04</b>	2.80000E-03	40
	D	4.28700E-02	<b>2.10900E-03</b>	2.60000E-03	40
SBDS <sub>9</sub>	A	1.44560E-02	<b>6.23000E-05</b>	5.70000E-03	40
	B	1.94854E-02	<b>6.04640E-05</b>	3.10000E-03	60
	C	3.23723E-01	<b>3.49000E-04</b>	4.95000E-02	55
	D	1.07220E-01	<b>1.62000E-04</b>	2.25000E-02	55
SBDS <sub>10</sub>	A	5.69780E-02	<b>1.90200E-03</b>	1.92300E-01	40
	B	1.39500E-02	<b>1.97000E-04</b>	9.70000E-03	55
	C	9.65800E-02	<b>2.25700E-03</b>	3.45000E-02	40
	D	2.83030E-01	<b>4.50000E-04</b>	3.32000E-02	55

**Table 5.6** Performance of RCN

Sub-datasets	Data	Mean Absolute Error		
		LM	BR	BFG-QN
SBDS <sub>1</sub>	A	<b>1.14102E-03</b>	1.27500E-03	2.52400E-03
	B	1.17400E-03	2.48800E-03	<b>5.79000E-04</b>
	C	1.03650E-02	<b>6.98300E-03</b>	1.25780E-02
	D	1.29940E-02	<b>7.69800E-03</b>	1.29150E-02
SBDS <sub>2</sub>	A	4.60000E-04	<b>3.77000E-04</b>	2.18328E-02
	B	2.22000E-04	<b>1.74000E-04</b>	1.18390E-02
	C	<b>1.36700E-03</b>	1.44800E-03	7.03600E-02
	D	8.61000E-04	<b>3.32000E-04</b>	3.45920E-02
SBDS <sub>3</sub>	A	<b>5.22000E-04</b>	4.43900E-03	2.55500E-03
	B	3.57762E-04	<b>1.05000E-04</b>	6.40100E-03
	C	4.21100E-03	<b>2.82000E-04</b>	4.10660E-02
	D	6.82000E-04	<b>1.68000E-04</b>	1.67200E-02
SBDS <sub>4</sub>	A	7.68000E-04	<b>4.80285E-05</b>	1.58640E-02
	B	5.20102E-05	<b>3.94799E-05</b>	7.16800E-03
	C	4.73000E-04	2.17658E-04	<b>2.00530E-02</b>
	D	2.08400E-03	<b>1.81824E-04</b>	6.17200E-03
SBDS <sub>6</sub>	A	<b>3.94114E-05</b>	9.52860E-05	4.38700E-03
	B	1.16000E-04	<b>1.12000E-04</b>	4.13600E-03
	C	<b>4.41680E-04</b>	3.77708E-01	1.42909E-01
	D	4.95000E-04	<b>3.62000E-04</b>	9.70200E-03
SBDS <sub>9</sub>	A	<b>1.83283E-04</b>	1.51900E-03	3.33800E-03
	B	<b>1.72000E-04</b>	1.89800E-03	1.58400E-03
	C	<b>1.82716E-03</b>	6.89000E-03	6.86900E-03
	D	<b>9.69000E-04</b>	5.59500E-03	4.03800E-03
SBDS <sub>10</sub>	A	5.80000E-04	<b>2.48000E-04</b>	1.01690E-02
	B	<b>1.47000E-04</b>	2.19000E-04	2.20500E-03
	C	1.30400E-03	<b>1.01200E-03</b>	4.03600E-03
	D	<b>4.30824E-04</b>	6.10500E-03	1.03750E-02

**Table 5.7** Performance of RBF

Sub-datasets	Data	Mean Absolute Error based on the number of neurons				
		40	45	50	55	60
SBDS <sub>1</sub>	A	1.03146E-04	9.34926E-05	1.00340E-04	5.70070E-05	5.08729E-05
	B	1.53000E-04	1.48000E-04	1.09000E-04	2.58000E-04	<b>2.40010E-05</b>
	C	4.84160E-05	5.01626E-05	5.02828E-05	5.02436E-05	4.98570E-05
	D	<b>3.88153E-05</b>	<b>3.81595E-05</b>	<b>3.81548E-05</b>	<b>3.81543E-05</b>	3.81548E-05
SBDS <sub>2</sub>	A	5.05680E-03	<b>2.70885E-03</b>	<b>1.53678E-03</b>	<b>1.33255E-03</b>	<b>1.36009E-03</b>
	B	<b>4.09600E-03</b>	3.23000E-03	3.02800E-03	1.85800E-03	1.68000E-03
	C	6.13000E-03	6.04800E-03	5.16800E-03	4.47700E-03	2.85700E-03
	D	8.42200E-03	8.58200E-03	7.28400E-03	5.07200E-03	4.30200E-03
SBDS <sub>3</sub>	A	<b>2.97000E-04</b>	<b>2.93000E-04</b>	<b>2.91000E-04</b>	<b>2.86000E-04</b>	2.75000E-04
	B	7.31000E-04	7.51000E-04	7.21000E-04	3.81000E-04	<b>2.24000E-04</b>
	C	1.50700E-03	6.52000E-04	5.11000E-04	9.74410E-04	6.96864E-04
	D	1.60300E-03	1.56300E-03	1.51400E-03	1.52400E-03	1.54100E-03
SBDS <sub>4</sub>	A	2.11138E-04	2.09055E-04	2.09022E-04	2.09002E-04	2.08946E-04
	B	<b>6.08910E-05</b>	<b>6.23224E-05</b>	<b>6.23330E-05</b>	<b>6.23224E-05</b>	<b>6.35603E-05</b>
	C	1.25481E-04	1.10437E-04	1.13763E-04	1.16843E-04	1.16840E-04
	D	4.06000E-04	3.84000E-04	3.81000E-04	3.80000E-04	3.80000E-04
SBDS <sub>6</sub>	A	1.00800E-03	8.17000E-04	3.54000E-04	1.60000E-04	4.34000E-04
	B	<b>1.53000E-04</b>	<b>1.03000E-04</b>	<b>1.26000E-04</b>	<b>1.04000E-04</b>	1.21483E-04
	C	4.14000E-04	4.01671E-04	4.01344E-04	4.28000E-04	2.34880E-04
	D	2.54000E-04	2.16000E-04	1.97000E-04	1.85000E-04	<b>8.74530E-05</b>
SBDS <sub>9</sub>	A	9.04430E-02	9.04440E-02	9.04440E-02	9.04440E-02	9.04440E-02
	B	<b>2.21914E-05</b>	<b>2.20646E-05</b>	<b>2.21172E-05</b>	<b>2.21087E-05</b>	<b>2.21087E-05</b>
	C	4.51944E-05	4.51615E-05	4.49053E-05	4.48180E-05	4.47060E-05
	D	3.51792E-05	3.41134E-05	3.40980E-05	3.35330E-05	3.35330E-05
SBDS <sub>10</sub>	A	2.34000E-04	2.85000E-04	2.76000E-04	2.72000E-04	2.76000E-04
	B	<b>4.03840E-05</b>	<b>4.03361E-05</b>	<b>3.98230E-05</b>	<b>3.98158E-05</b>	<b>3.98230E-05</b>
	C	5.37473E-05	5.34101E-05	5.34009E-05	5.36361E-05	5.36107E-05
	D	1.21000E-04	1.20000E-04	1.20000E-04	1.16000E-04	1.16000E-04

Measure performance by RMSE is depicted in Table 5.8 for best results for each NN algorithms.

**Table 5.8** RMSE for FFN, RCN and RBF

Sub-datasets	Data	RMSE		
		FFN	RCN	RBF
SBDS <sub>1</sub>	A	<b>2.280E-03</b>	<b>1.237E-02</b>	2.613E-03
	B	3.606E-03	1.994E-02	1.521E-03
	C	4.494E-03	1.994E-02	1.677E-03
	D	4.669E-03	1.956E-02	<b>1.377E-03</b>
SBDS <sub>2</sub>	A	1.170E-01	7.114E-03	1.337E-02
	B	9.274E-02	4.125E-03	<b>1.273E-02</b>
	C	7.629E-03	8.823E-03	1.275E-02
	D	<b>6.181E-03</b>	<b>4.065E-03</b>	1.463E-02
SBDS <sub>3</sub>	A	4.123E-03	8.368E-03	6.071E-03
	B	3.464E-03	3.187E-03	<b>4.650E-03</b>
	C	5.568E-02	4.009E-03	5.396E-03
	D	<b>1.421E-02</b>	<b>2.889E-03</b>	8.677E-03
SBDS <sub>4</sub>	A	2.775E-03	2.539E-03	5.295E-03
	B	<b>2.490E-03</b>	<b>1.951E-03</b>	<b>2.423E-03</b>
	C	4.164E-03	3.521E-03	2.508E-03
	D	3.114E-03	3.007E-03	4.344E-03
SBDS <sub>6</sub>	A	1.138E-02	<b>2.300E-03</b>	4.632E-03
	B	<b>2.569E-03</b>	3.285E-03	3.166E-03
	C	4.615E-03	5.015E-03	3.657E-03
	D	1.024E-02	4.243E-03	<b>2.085E-03</b>
SBDS <sub>9</sub>	A	2.898E-03	4.959E-03	1.563E-03
	B	<b>2.408E-03</b>	<b>4.078E-03</b>	<b>1.291E-03</b>
	C	4.461E-03	1.020E-02	1.596E-03
	D	9.000E-02	6.941E-03	1.459E-03
SBDS <sub>10</sub>	A	1.597E-02	5.769E-03	5.598E-03
	B	<b>4.359E-03</b>	<b>3.764E-03</b>	1.960E-03
	C	1.134E-02	7.592E-03	<b>1.744E-03</b>
	D	1.493E-01	4.628E-03	2.398E-03

In order to investigate the effect of the algorithms the average of the training period for FFN, RCN, and RB was calculated when using BR as a training algorithm. Based on the results in Table 5.9, the training time depends on the sub dataset size for example SBDS<sub>2</sub> include 13 features consumes more time than other sub-datasets as well

as training rate and its impact, decrease the training ratio leads to increased speed, therefore always D is better than A. From Table 5.9, RBF is characterized by its speed followed by RCN and finally FFN.

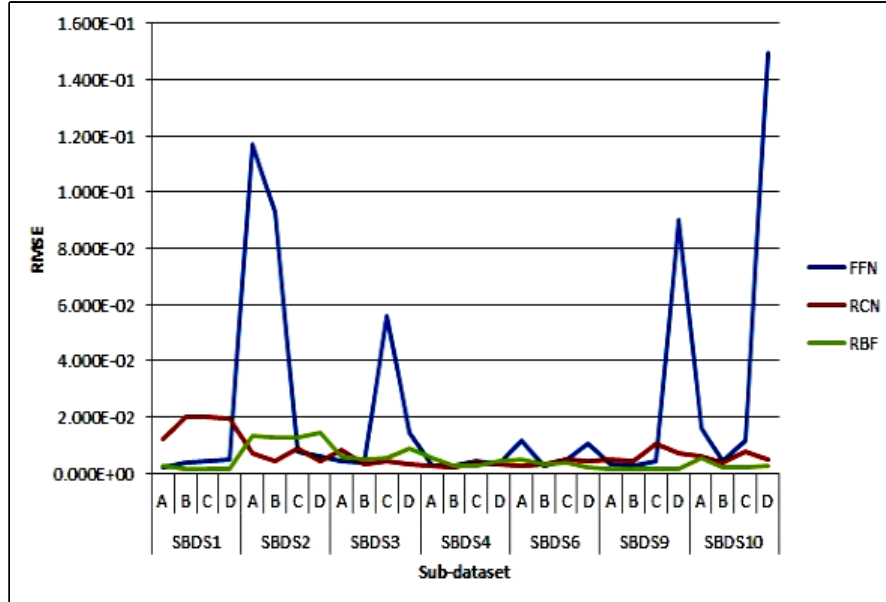
**Table 5.9** Comparison between NNs based on time

Sub-datasets	Data	TIME		
		FFN	RCN	RBF
SBDS <sub>1</sub>	A	00:03:20	00.04.54	00:00:20
	B	00:02:53	00.04.04	00:00:17
	C	00:01:50	00.03.23	00:00:14
	D	<b>00:01:24</b>	<b>00.03.21</b>	<b>00:00:13</b>
SBDS <sub>2</sub>	A	00:13:33	00.06.57	00.00.34
	B	00:13:29	00.06.42	00.00.19
	C	00:12:52	00.05.27	00.00.17
	D	<b>00:11:40</b>	<b>00.04:35</b>	<b>00:00:16</b>
SBDS <sub>3</sub>	A	00:03:53	00:05:00	00:00:28
	B	00:03:45	00:04:30	00:00:25
	C	00:04:15	00:03:53	00:00:21
	D	<b>00:02:48</b>	<b>00:03:15</b>	<b>00:00:20</b>
SBDS <sub>4</sub>	A	00:03:32	00:04:19	00:00:27
	B	00:02:31	00:03:49	00:00:24
	C	00:02:18	00:03:30	00:00:20
	D	<b>00:02:14</b>	<b>00:03:11</b>	<b>00:00:19</b>
SBDS <sub>6</sub>	A	00:03:39	00:04:28	00:00:28
	B	00:03:30	00:04:05	00:00:24
	C	00:03:14	00:03:06	00:00:21
	D	<b>00:02:39</b>	<b>00:03:04</b>	<b>00:00:19</b>
SBDS <sub>9</sub>	A	00:01:58	00:04:38	00:00:28
	B	00:01:29	00:04:05	00:00:24
	C	00:00:47	00:03:43	00:00:21
	D	<b>00:00:22</b>	<b>00:03:10</b>	<b>00:00:19</b>
SBDS <sub>10</sub>	A	00:02:01	00:04:49	00:00:26
	B	00:01:15	00:04:22	00:00:24
	C	00:01:38	00:03:09	00:00:21
	D	<b>00:00:32</b>	<b>00:02:46</b>	<b>00:00:19</b>

## 5.2 A Comparison Analysis of Direct Prediction Models

From the previous results, the following comparisons were drawn: Based on experiments in Section 5.1.1, SMOreg outperformed other algorithms but it suffers from the consumed time. Kstar gets high error therefore it was excluded from further experiments. To further improve the results, five previous algorithms were used with meta-learning models, which is discussed in Chapter 6.

The results of three different types of neural networks were compared as shown in Figure 5.3 using RMSE and observed that the RBF network outperformed other methods in obtaining the lowest error (MAE= 2.206 E-05 and RMSE= 1.291E-03). Also, the data set using training 80% and testing 20% accomplished the best results for RCN and RBF neural network methods. In the FFN and RBF networks, the best results were obtained when using 45 neurons and RBF proved its superiority. RBF networks outperformed again in the time factor, as it was faster than feed-forward and recurrent neural networks. Table 5.10 shows this comparison.



**Figure 5.3** Comparison among 3 type of NNs



**Table 5.10** Summary for the results which explain the comparison between NNs

Prediction Model	Data	MAE	RMSE	Sub-dataset	Time
<b>FFN</b>	A	3.84294E-05	2.280E-03	SBDS <sub>1</sub>	00:03:20
<b>RCN</b>	B	3.94114E-05	2.300E-03	SBDS <sub>6</sub>	00:04:05
<b>RBF</b>	B	<b>2.20646E-05</b>	<b>1.291E-03</b>	<b>SBDS<sub>9</sub></b>	00:00:24

More results in detail for NNs experiments are attached in Appendix B.

### 5.3 Conclusions

In this research, simple machine-learning approaches were applied to predict the daily WTI price for every barrel of crude oil in USD. List of features used as the input factors were divided into ten sub-datasets resulting in numerous attribute selection algorithms and four data sets with different percentages of training and testing. Experiments start with six direct prediction models namely isotonic regression, SMOReg, Kstar, IBK, ExtraTree and REPTree followed by several types of NNs including FFN, RCN and RBF. This Chapter provides successful comparisons of direct prediction models, sub- datasets, and different group of training and testing data sets.

# CHAPTER SIX

## 6. COMBINED PREDICTION MODELS

### Overview

In order to improve the results of direct models that was explored in the previous Chapter, several types of combined models are illustrated in this Chapter. Experiments are done according to the previous combined predictor schemes as discussed in Chapter 4. The Chapter starts with Meta prediction followed by hybrid prediction and finally Ensemble model. Various comparisons are done between different combined models and finally with the results of the final model. By the end of the Chapter, the two research questions of this thesis, mentioned below were answered:

*Which set of features can better describe the performance?*

*What is the best percentage for training data and testing data?*

The results of the experiments are also published in [180], [181], [124], and [182]

### 6.1 Meta Prediction Experiments

Meta prediction models depending on two parts were grouped: Bagging and Random subspace, which separate data into subparts and each part is trained by the same predictor. Another part including Ensemble selection, Voting, and Stacking, which provides the same input to a number of predictors and combine their output using a given decision logic. First Bagging and Random-subspace with five direct prediction models were implemented.

In this experiment, improve prediction model results by using the Bagging model. Best 7 sub-datasets with four categories of training and testing were used, and finally calculated the error by using MAE and RMSE for all prediction models as illustrated in Tables 6.1 and 6.2 respectively. All direct models improve their results after the implementation of Bagging and the results are displayed in Tables 5.1 and 5.2.

SMOreg achieved best results with MAE = 1.840E-02 and RMSE = 2.610E-02 among all the overall algorithms. Based on the sub datasets, best results were focused again on SBDS<sub>9</sub> and SBDS<sub>10</sub> for all the algorithms except Isotonic Regression, as the factors did not have any effect on this algorithm. On other hand, best results existing in-group (A) except SMOreg also achieved best results with (B).

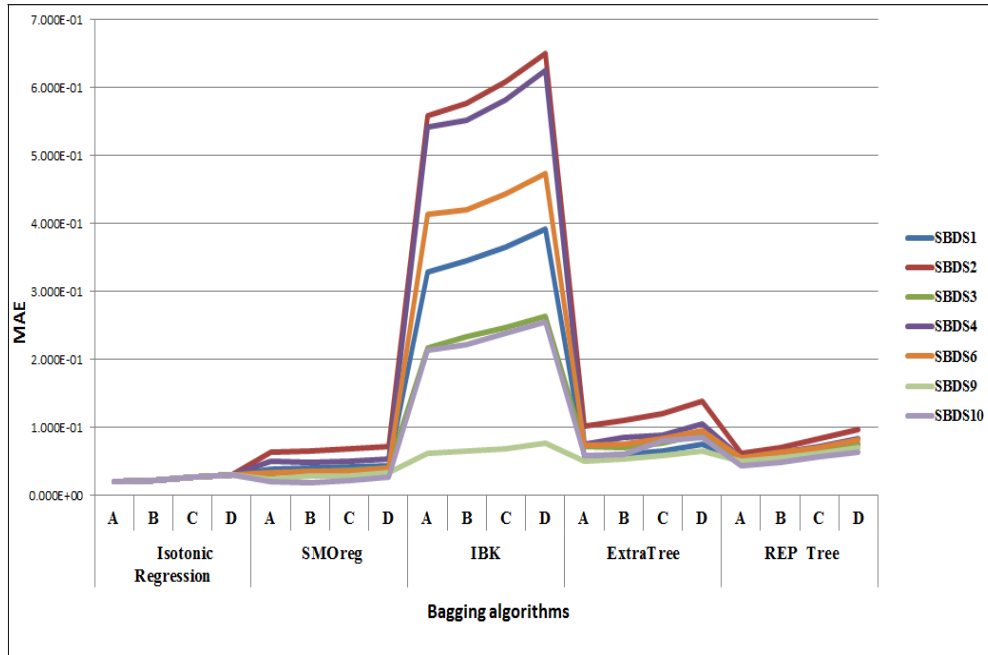
**Table 6.1** Bagging algorithms with 5 base prediction models using MAE

Prediction Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
<b>Bagging Isotonic Regression</b>	<b>A</b>	<b>2.060E-02</b>	<b>2.060E-02</b>	<b>2.060E-02</b>	<b>2.060E-02</b>	<b>2.060E-02</b>	<b>2.060E-02</b>	<b>2.060E-02</b>
	<b>B</b>	2.230E-02	2.230E-02	2.230E-02	2.230E-02	2.230E-02	2.230E-02	2.230E-02
	<b>C</b>	2.600E-02	2.600E-02	2.600E-02	2.600E-02	2.600E-02	2.600E-02	2.600E-02
	<b>D</b>	3.020E-02	3.020E-02	3.020E-02	3.020E-02	3.020E-02	3.020E-02	3.020E-02
<b>Bagging SMOreg</b>	<b>A</b>	3.840E-02	6.400E-02	3.200E-02	4.930E-02	3.330E-02	2.390E-02	1.970E-02
	<b>B</b>	3.970E-02	6.570E-02	3.450E-02	4.880E-02	3.550E-02	2.820E-02	<b>1.840E-02</b>
	<b>C</b>	4.180E-02	6.830E-02	3.600E-02	5.050E-02	3.590E-02	2.830E-02	2.250E-02
	<b>D</b>	4.380E-02	7.190E-02	3.950E-02	5.410E-02	3.930E-02	3.330E-02	2.620E-02
<b>Bagging IBK</b>	<b>A</b>	3.290E-01	5.592E-01	2.169E-01	5.416E-01	4.131E-01	<b>6.160E-02</b>	2.141E-01
	<b>B</b>	3.451E-01	5.764E-01	2.327E-01	5.526E-01	4.208E-01	6.450E-02	2.212E-01
	<b>C</b>	3.650E-01	6.080E-01	2.468E-01	5.817E-01	4.433E-01	6.870E-02	2.391E-01
	<b>D</b>	3.927E-01	6.501E-01	2.638E-01	6.249E-01	4.738E-01	7.660E-02	2.560E-01
<b>Bagging Extra Tree</b>	<b>A</b>	5.630E-02	1.023E-01	7.120E-02	7.500E-02	7.360E-02	<b>4.950E-02</b>	5.910E-02
	<b>B</b>	6.010E-02	1.105E-01	6.930E-02	8.440E-02	7.450E-02	5.360E-02	6.050E-02
	<b>C</b>	6.520E-02	1.209E-01	7.710E-02	8.900E-02	8.390E-02	5.810E-02	7.980E-02
	<b>D</b>	7.490E-02	1.381E-01	8.970E-02	1.049E-01	9.510E-02	6.510E-02	8.450E-02
<b>Bagging REP Tree</b>	<b>A</b>	5.230E-02	6.160E-02	5.110E-02	5.570E-02	5.520E-02	4.980E-02	<b>4.330E-02</b>
	<b>B</b>	5.800E-02	6.980E-02	5.840E-02	6.380E-02	6.290E-02	5.540E-02	4.850E-02
	<b>C</b>	6.690E-02	8.310E-02	6.480E-02	7.140E-02	7.040E-02	6.160E-02	5.690E-02
	<b>D</b>	7.590E-02	9.640E-02	7.520E-02	8.300E-02	8.160E-02	7.040E-02	6.290E-02

**Table 6.2** Bagging algorithms with 5 base prediction models using RMSE

Prediction Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
<b>Bagging Isotonic Regression</b>	<b>A</b>	<b>4.560E-02</b>	<b>4.560E-02</b>	<b>4.560E-02</b>	<b>4.560E-02</b>	<b>4.560E-02</b>	<b>4.560E-02</b>	<b>4.560E-02</b>
	<b>B</b>	4.910E-02	4.910E-02	4.910E-02	4.910E-02	4.910E-02	4.910E-02	4.910E-02
	<b>C</b>	6.160E-02	6.160E-02	6.160E-02	6.160E-02	6.160E-02	6.160E-02	6.160E-02
	<b>D</b>	7.160E-02	7.160E-02	7.160E-02	7.160E-02	7.160E-02	7.160E-02	7.160E-02
<b>Bagging SMOreg</b>	<b>A</b>	6.060E-02	9.050E-02	5.380E-02	7.040E-02	5.640E-02	4.530E-02	2.700E-02
	<b>B</b>	6.910E-02	9.910E-02	6.380E-02	7.690E-02	6.670E-02	5.570E-02	<b>2.610E-02</b>
	<b>C</b>	7.070E-02	9.980E-02	6.480E-02	7.780E-02	6.680E-02	5.660E-02	3.070E-02
	<b>D</b>	7.970E-02	1.078E-01	7.500E-02	8.740E-02	7.770E-02	6.700E-02	3.510E-02
<b>Bagging IBK</b>	<b>A</b>	4.920E-01	7.841E-01	3.486E-01	7.776E-01	6.475E-01	<b>1.310E-01</b>	3.673E-01
	<b>B</b>	5.310E-01	8.412E-01	4.121E-01	8.091E-01	6.679E-01	1.447E-01	3.757E-01
	<b>C</b>	5.491E-01	8.831E-01	4.278E-01	8.476E-01	6.976E-01	1.485E-01	4.030E-01
	<b>D</b>	5.936E-01	9.560E-01	4.635E-01	9.324E-01	7.707E-01	1.701E-01	4.309E-01
<b>Bagging Extra Tree</b>	<b>A</b>	1.383E-01	2.129E-01	1.888E-01	1.622E-01	1.724E-01	<b>1.231E-01</b>	1.472E-01
	<b>B</b>	1.514E-01	2.342E-01	1.661E-01	1.964E-01	1.735E-01	1.444E-01	1.510E-01
	<b>C</b>	1.610E-01	2.581E-01	1.825E-01	2.003E-01	2.061E-01	1.521E-01	2.035E-01
	<b>D</b>	1.882E-01	3.129E-01	2.099E-01	2.441E-01	2.195E-01	1.572E-01	2.225E-01
<b>Bagging REP Tree</b>	<b>A</b>	1.401E-01	1.621E-01	1.353E-01	1.527E-01	1.509E-01	1.343E-01	<b>1.204E-01</b>
	<b>B</b>	1.556E-01	1.926E-01	1.641E-01	1.775E-01	1.773E-01	1.468E-01	1.354E-01
	<b>C</b>	1.848E-01	2.301E-01	1.724E-01	1.949E-01	1.937E-01	1.648E-01	1.661E-01
	<b>D</b>	1.949E-01	2.648E-01	1.903E-01	2.189E-01	2.152E-01	1.746E-01	1.662E-01

According to Figure 6.1, five prediction models with Bagging were compared. As evident, bagging is significantly more accurate than the prediction models except IBK, which is also less accurate comparing with another direct prediction models, therefore IBK was removed from next set of experiments.



**Figure 6.1** Bagging algorithm with 5 base prediction models

Similar to the Bagging experiments, 7 sub-data sets with 4 categories was exposed to Random subspace method. Number of single predictors was squeezed to four algorithms when Random subspace is used as a result of the exclusion of IBK algorithm due to their poor performance. MAE and RMSE results are illustrated in Tables 6.3 and 6.4 respectively.

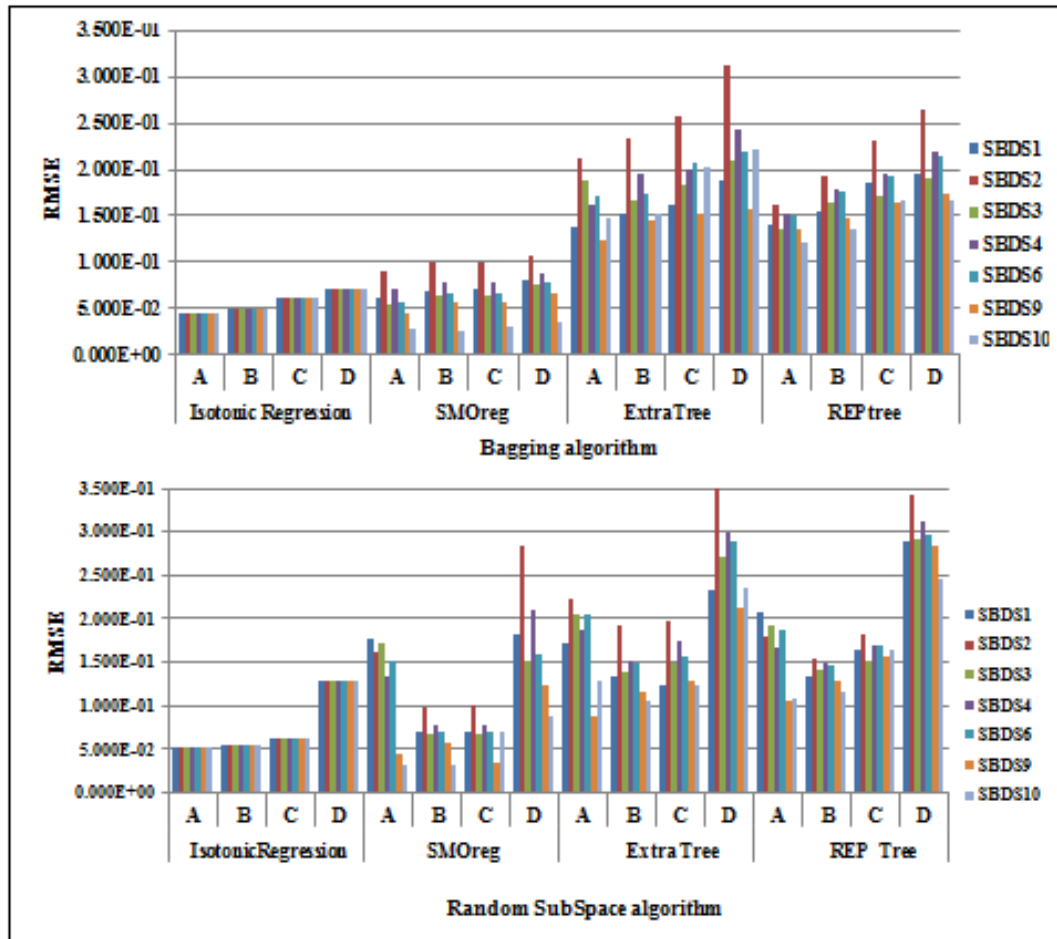
The main goal for this experiment is to investigate the best results for modeling oil prices. The four prediction models were combined as mentioned above using random subspace and then compared their results with bagging methods. Figure 6.2 shows that Random subspace works better than bagging with Extra Tree and REP Tree models, while bagging better than Random subspace with Isotonic Regression and SMOreg.

**Table 6.3** MAE for Random Subspace

Prediction Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
Random subspace Isotonic Regression	A	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>	<b>2.220E-02</b>
	B	2.420E-02	2.420E-02	2.420E-02	2.420E-02	2.420E-02	2.420E-02	2.420E-02
	C	2.780E-02	2.780E-02	2.780E-02	2.780E-02	2.780E-02	2.780E-02	2.780E-02
	D	3.250E-02	3.250E-02	3.250E-02	3.250E-02	3.250E-02	3.250E-02	3.250E-02
Random subspace SMOreg	A	4.350E-02	6.510E-02	3.040E-02	4.850E-02	3.520E-02	<b>2.210E-02</b>	2.320E-02
	B	4.000E-02	6.770E-02	3.660E-02	4.930E-02	3.680E-02	2.630E-02	2.440E-02
	C	3.960E-02	6.880E-02	3.820E-02	4.970E-02	3.760E-02	2.860E-02	2.650E-02
	D	4.540E-02	7.120E-02	3.780E-02	5.270E-02	4.010E-02	3.100E-02	2.230E-02
Random subspace Extra Tree	A	4.120E-02	7.550E-02	5.120E-02	5.670E-02	5.360E-02	<b>3.870E-02</b>	4.010E-02
	B	4.790E-02	8.280E-02	5.420E-02	6.090E-02	5.830E-02	4.120E-02	4.320E-02
	C	4.910E-02	8.970E-02	5.890E-02	6.810E-02	6.310E-02	4.700E-02	5.060E-02
	D	5.830E-02	1.028E-01	6.830E-02	7.540E-02	7.270E-02	5.340E-02	5.930E-02
Random subspace REP Tree	A	4.710E-02	5.370E-02	4.610E-02	5.130E-02	4.880E-02	4.610E-02	<b>4.290E-02</b>
	B	5.330E-02	6.090E-02	5.490E-02	5.710E-02	5.720E-02	5.340E-02	4.630E-02
	C	6.480E-02	7.080E-02	6.270E-02	6.590E-02	6.540E-02	6.030E-02	5.580E-02
	D	7.230E-02	8.600E-02	7.350E-02	7.860E-02	7.450E-02	7.120E-02	6.190E-02

**Table 6.4** RMSE for Random Subspace

Prediction Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
Random subspace Isotonic Regression	A	<b>5.270E-02</b>	<b>5.270E-02</b>	<b>5.270E-02</b>	<b>5.270E-02</b>	<b>5.270E-02</b>	<b>5.270E-02</b>	<b>5.270E-02</b>
	B	5.340E-02	5.340E-02	5.340E-02	5.340E-02	5.340E-02	5.340E-02	5.340E-02
	C	6.160E-02	6.160E-02	6.160E-02	6.160E-02	6.160E-02	6.160E-02	6.160E-02
	D	1.294E-01	1.294E-01	1.294E-01	1.294E-01	1.294E-01	1.294E-01	1.294E-01
Random subspace SMOreg	A	1.760E-01	1.618E-01	1.714E-01	1.332E-01	1.512E-01	<b>4.490E-02</b>	3.100E-02
	B	6.990E-02	9.870E-02	6.650E-02	7.720E-02	6.890E-02	5.630E-02	3.180E-02
	C	7.020E-02	9.910E-02	6.770E-02	7.730E-02	6.910E-02	3.530E-02	7.020E-02
	D	1.809E-01	2.834E-01	1.505E-01	2.102E-01	1.597E-01	1.233E-01	8.870E-02
Random subspace Extra Tree	A	1.711E-01	2.217E-01	2.053E-01	1.865E-01	2.041E-01	<b>8.770E-02</b>	1.290E-01
	B	1.345E-01	1.916E-01	1.382E-01	1.506E-01	1.482E-01	1.149E-01	1.064E-01
	C	1.228E-01	1.976E-01	1.483E-01	1.740E-01	1.572E-01	1.292E-01	1.228E-01
	D	2.322E-01	4.093E-01	2.719E-01	3.003E-01	2.897E-01	2.127E-01	2.363E-01
Random subspace REP Tree	A	2.084E-01	1.787E-01	1.933E-01	1.666E-01	1.874E-01	1.051E-01	<b>1.078E-01</b>
	B	1.327E-01	1.546E-01	1.416E-01	1.498E-01	1.457E-01	1.286E-01	1.154E-01
	C	1.643E-01	1.826E-01	1.509E-01	1.700E-01	1.690E-01	1.555E-01	1.643E-01
	D	2.879E-01	3.428E-01	2.927E-01	3.130E-01	2.967E-01	2.836E-01	2.469E-01



**Figure 6.2** Comparison between Bagging and Random SubSpace using 7 sub datasets and 4 categories of training and testing.

Another part of Meta prediction experiments contains Ensemble selection, Voting, and Stacking. These algorithms combine several prediction models using different ways as explained in Chapter 4.

The main idea behind this Section is to promote diversity among prediction models and then combine them by different methods to obtain better predictive performance. In this experiment, four different prediction models were used: SMOreg, Isotonic Regression, Extra-Tree and REPTree. Then three different combining techniques including stacking, voting, and ensemble selection were implemented. the results based on the different percentages of training and testing were illustrated. MAE and

RMSE were used as a performance measure for all the prediction models. Tables 6.5 and 6.6 show results for each technique.

**Table 6.5** MAE for meta learning (Stacking, Voting and Ensemble selection)

Combination Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
Stacking	A	<b>2.710E-02</b>	3.870E-02	2.760E-02	2.870E-02	7.230E-02	3.870E-02	3.870E-02
	B	3.740E-02	4.250E-02	3.080E-02	3.320E-02	4.170E-02	4.250E-02	4.250E-02
	C	4.440E-02	5.890E-02	2.990E-02	3.140E-02	9.640E-02	4.920E-02	4.920E-02
	D	4.920E-02	8.150E-02	3.680E-02	3.890E-02	3.730E-02	5.680E-02	5.680E-02
Voting	A	5.650E-02	7.070E-02	5.790E-02	6.180E-02	6.080E-02	<b>5.400E-02</b>	5.490E-02
	B	5.940E-02	7.340E-02	6.230E-02	6.490E-02	6.090E-02	5.770E-02	5.830E-02
	C	6.260E-02	7.630E-02	6.440E-02	7.180E-02	6.800E-02	6.370E-02	5.620E-02
	D	6.400E-02	7.940E-02	7.060E-02	7.380E-02	7.000E-02	6.690E-02	6.400E-02
Ensemble selection	A	3.320E-02	1.470E-02	<b>1.420E-02</b>	1.730E-02	1.470E-02	3.040E-02	2.570E-02
	B	3.600E-02	1.670E-02	1.460E-02	1.780E-02	1.610E-02	3.260E-02	2.840E-02
	C	3.790E-02	1.640E-02	1.690E-02	1.880E-02	1.710E-02	3.540E-02	3.230E-02
	D	4.230E-02	1.700E-02	1.840E-02	2.700E-02	1.730E-02	4.320E-02	3.210E-02

**Table 6.6** RMSE for meta learning (Stacking, Voting and Ensemble selection)

Combination Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
Stacking	A	<b>5.510E-02</b>	8.240E-02	5.550E-02	5.630E-02	5.550E-02	1.086E-01	6.160E-02
	B	7.040E-02	8.730E-02	5.730E-02	5.860E-02	8.340E-02	8.730E-02	8.730E-02
	C	9.060E-02	1.128E-01	6.240E-02	6.330E-02	1.421E-01	1.096E-01	1.096E-01
	D	9.28E-02	1.34E-01	7.35E-02	7.47E-02	8.65E-02	7.37E-02	1.06E-01
Voting	A	8.550E-02	1.195E-01	8.610E-02	9.620E-02	9.930E-02	<b>7.670E-02</b>	8.510E-02
	B	9.730E-02	1.289E-01	1.041E-01	1.136E-01	1.020E-01	9.500E-02	1.132E-01
	C	1.091E-01	1.345E-01	1.069E-01	1.191E-01	1.159E-01	1.047E-01	9.650E-02
	D	2.548E-01	3.164E-01	2.814E-01	2.938E-01	2.787E-01	2.663E-01	2.550E-01
Ensemble selection	A	4.990E-02	2.600E-02	<b>2.420E-02</b>	2.660E-02	2.530E-02	5.380E-02	4.390E-02
	B	5.240E-02	2.810E-02	2.690E-02	2.920E-02	2.900E-02	6.250E-02	4.930E-02
	C	5.420E-02	3.010E-02	3.010E-02	3.040E-02	2.860E-02	6.370E-02	5.540E-02
	D	6.280E-02	3.050E-02	3.350E-02	4.100E-02	3.020E-02	7.760E-02	5.220E-02

According to Section 5.1.2, NNs accomplished distinguished results and hence NNs with Meta prediction models were used in order to improve results. Bagging does not



work with NNs and the error was increased with Random subspace when compared with direct NNs results, as shown in Table 6.7. It is important to note that most Meta prediction models consumed a long time up to 4 hours or more. Therefore, hybrid model was used as another type of combined prediction models and their results are illustrated in Section 6.2.

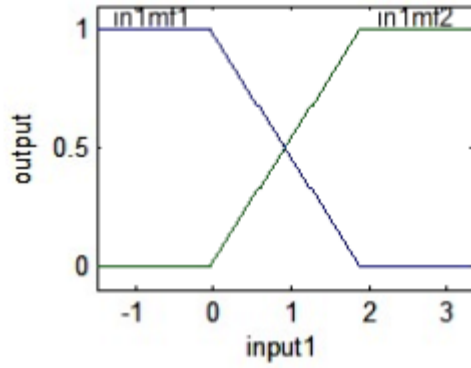
**Table 6.7** NNs results with Bagging and Random subspace

Combination Model	Data	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
Bagging Multilayer Perceptron	A	5.210E+01	2.228E+01	5.218E+01	5.765E+01	5.166E+01	5.039E+01	5.013E+01
	B	5.195E+01	2.556E+01	5.234E+01	4.942E+01	5.197E+01	5.054E+01	5.027E+01
	C	5.224E+01	2.510E+01	5.234E+01	5.154E+01	5.193E+01	5.055E+01	5.027E+01
	D	5.204E+01	2.342E+01	5.239E+01	4.849E+01	5.227E+01	5.068E+01	5.035E+01
Random-subspace Multilayer Perceptron	A	2.270E-02	4.450E-02	9.600E-03	6.600E-03	2.640E-02	2.066E-01	1.926E-01
	B	2.570E-02	5.070E-02	1.020E-02	8.700E-03	1.970E-02	2.670E-01	2.211E-01
	C	2.130E-02	4.900E-02	9.300E-03	7.900E-03	4.150E-02	2.301E-01	2.182E-01
	D	3.260E-02	5.050E-02	8.200E-03	1.220E-02	2.330E-02	2.351E-01	2.588E-01

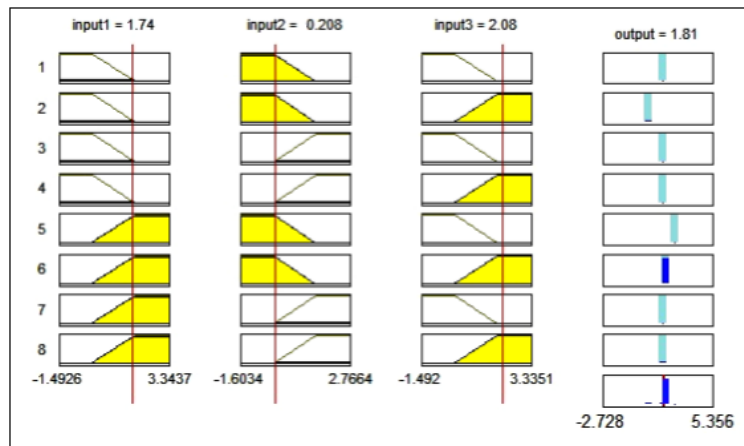
## 6.2 Hybrid Prediction Models

In this Section, a neuro-fuzzy (ANFIS) model is developed using the MATLAB Fuzzy Logic Toolbox to predict the crude oil price. As mentioned in Chapter 4, the neural network learning method is used for building a fuzzy model and used 7 sub datasets and 100 learning epochs. The membership function, rule base and ANFIS structure are displayed in Figures 6.3, 6.4 and 6.5 respectively for SBDS7-B (three inputs). The learned eight *if-then* rules appear as follows:

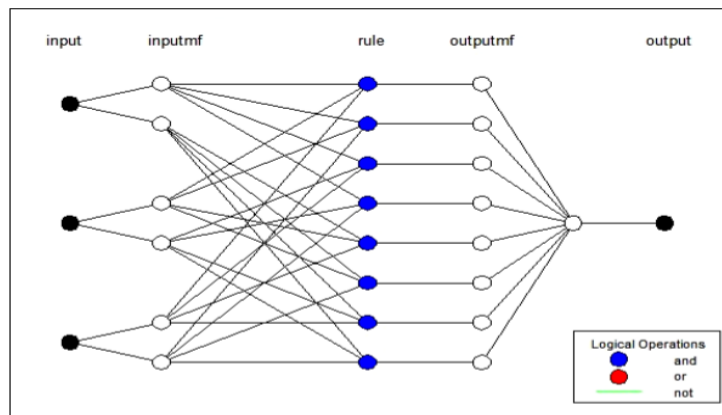
1. *If (WTI is in1mf1) and (GPNY is in2mf1) and (FC1 is in3mf1) then (WTI is out1mf1)*
2. *If (WTI is in1mf1) and (GPNY is in2mf1) and (FC1 is in3mf2) then (WTI is out1mf2)*
3. *If (WTI is in1mf1) and (GPNY is in2mf2) and (FC1 is in3mf1) then (WTI is out1mf3)*
4. *If (WTI is in1mf1) and (GPNY is in2mf2) and (FC1 is in3mf2) then (WTI is out1mf4)*
5. *If (WTI is in1mf2) and (GPNY is in2mf1) and (FC1 is in3mf1) then (WTI is out1mf5)*
6. *If (WTI is in1mf2) and (GPNY is in2mf1) and (FC1 is in3mf2) then (WTI is out1mf6)*
7. *If (WTI is in1mf2) and (GPNY is in2mf2) and (FC1 is in3mf1) then (WTI is out1mf7)*
8. *If (WTI is in1mf2) and (GPNY is in2mf2) and (FC1 is in3mf2) then (WTI is out1mf8)*



**Figure 6.3** Trapezoidal-shaped membership function for the first Input



**Figure 6.4** Developed TSK FIS using 3 inputs



**Figure 6.5** Developed ANFIS structure with 3 inputs

According to Figure 6.4, a sample rule would appear as follows:

*If (WTI (t) is 1.74) and (GPNY is 0.208) and (FCI is 2.08) then (WTI(t+1) is 1.81)*

Tables 6.8 and 6.9 show that the best results for each sub-datasets based on Datasets A and B (bold font) and overall data SBDS<sub>10</sub>. Dataset (B) achieved best results with MAE= 4.70906E-07 and RMSE=7.382E-07

**Table 6.8** ANFIS results (MAE) for 7 sub-data sets

DATA	Mean Absolute Error (MAE)						
	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
A	<b>1.235E-05</b>	<b>8.941E-06</b>	3.752E-02	3.444E-04	<b>9.969E-04</b>	<b>2.392E-06</b>	7.953E-06
B	1.566E-05	1.328E-05	<b>2.892E-02</b>	3.692E-04	1.503E-03	6.789E-06	<b>4.709E-07</b>
C	5.812E-05	2.874E-05	2.125E-01	8.191E-04	7.596E-03	9.509E-06	1.773E-06
D	4.847E-05	2.902E-01	1.129E-01	<b>1.904E-04</b>	1.233E-02	1.902E-05	2.470E-02

**Table 6.9** ANFIS results (RMSE) for 7 sub-datasets

DATA	Root Mean Square Error (RMSE)						
	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
A	<b>2.799E-05</b>	<b>1.175E-05</b>	8.680E-02	5.860E-01	<b>2.193E-03</b>	<b>4.699E-06</b>	7.626E-07
B	3.037E-05	2.826E-05	<b>7.802E-02</b>	1.646E-03	6.549E-03	1.497E-05	<b>7.382E-07</b>
C	2.813E-04	1.415E-04	4.394E-01	2.506E-02	3.378E-02	2.116E-05	3.955E-06
D	8.016E-05	2.540E-03	5.860E-01	<b>6.648E-04</b>	1.210E-01	2.390E-05	4.942E-02

Training by ANFIS completed in about 3 seconds and Table 6.10 illustrates ANFIS time for training.

**Table 6.10** ANFIS training time

DATA	Time (hh: mm: ss)						
	SBDS <sub>1</sub>	SBDS <sub>2</sub>	SBDS <sub>3</sub>	SBDS <sub>4</sub>	SBDS <sub>6</sub>	SBDS <sub>9</sub>	SBDS <sub>10</sub>
A	00:00:10	00:01:55	00:02:17	00:00:09	00:00:28	00:00:05	00:00:05
B	00:00:08	00:01:46	00:02:02	00:00:08	00:00:25	00:00:04	00:00:04
C	00:00:07	00:01:40	00:01:42	00:00:07	00:00:22	00:00:04	00:00:04
D	00:00:06	00:01:10	00:01:33	00:00:06	00:00:20	00:00:03	00:00:03

Experimental results illustrated that the ANFIS model performed well. As mentioned in Section 4.5.2, that main motivation behind the Ensemble methodology is to weigh

numerous single predictors and combine them to obtain a predictor that outperforms them all [145].

### 6.3 Ensemble Prediction Models

In this research, the basic ensemble method and generalized ensemble method were employed, which are the most popular techniques used in this phase:

#### 6.3.1. The Basic Ensemble Method

Basic ensemble (average) method with NNs were used, according to Table 5.10. Best results for RBF and RCN were achieved with dataset (B) and for FFN when using (A). To combine three models, dataset (B) for all NNs was used and then combined them by using the average method for creating ensembles and the results are depicted in Table 6.11.

**Table 6.11** MAE for basic ensemble results for neural networks

Neural Networks	MAE	RMSE
<b>RCN</b>	3.9411E-05	2.300E-03
<b>RBF</b>	<b>2.2065E-05</b>	<b>1.291E-03</b>
<b>FFN</b>	6.0465E-05	2.569E-03
<b>Ensemble Average</b>	3.1780E-05	2.172E-03

It is important to notice that Average method improved the individual results for RCN and FFN but RBF still had the best overall results. In order to improve ANFIS results again Ensemble Average with ANFIS using datasets (A) and (B) were applied. The MAE and RMSE results are displayed in Tables 6.12 and 6.13 for group (A) and for group (B) respectively.

**Table 6.12** Ensemble using Average method for Data (A)

Data (A)	MAE	RMSE
SBDS <sub>1</sub>	1.2352E-05	2.799E-05
SBDS <sub>2</sub>	8.9407E-06	1.1752E-05
SBDS <sub>6</sub>	9.9686E-04	2.193E-03
SBDS <sub>9</sub>	<b>2.3922E-06</b>	<b>4.69999E-06</b>
Ensemble Average	2.5011E-04	1.097E-03

**Table 6.13** Ensemble using Average method for Data (B)

Data (B)	MAE	RMSE
SBDS <sub>3</sub>	2.89157E-02	7.802E-02
SBDS <sub>10</sub>	<b>4.70906E-07</b>	<b>7.382E-07</b>
Ensemble Average	1.44578E-02	3.901E-02

Likewise ANFIS Ensemble average results are better than some sub-dataset such as SBDS<sub>6</sub> in DATA (A) and SBDS<sub>3</sub> in DATA (B). However, there was no clear superiority on the other sub datasets such as SBDS<sub>1</sub>, SBDS<sub>2</sub> and SBDS<sub>9</sub> in data (A) and SBDS<sub>10</sub> in data (B).

### 6.3.2. The Generalized Ensemble Method (GEM)

Another important concept regarding the performance of a predictive model is the GEM method. Based on equations (4.5) and (4.6) we need to find the optimal values of weight  $\alpha$  to minimize the MAE or RMSE between the outputs and the desired values. PSO algorithm was used to determine the optimal weights. PSO algorithm is described in Chapter 2. As evident from Tables 6.14 and 6.15 the GEM model is better in predicting ANFIS results than the average method for Data (A) and Data (B) respectively.

**Table 6.14** Ensemble of PSO-ANFIS for Data (A)

Data (A)	MAE	RMSE
SBDS <sub>1</sub>	1.2352E-05	2.7993E-05
SBDS <sub>2</sub>	8.9407E-06	1.1752E-05
SBDS <sub>6</sub>	9.9686E-04	2.1932E-03
SBDS <sub>9</sub>	2.3922E-06	4.69999E-06
<b>Ensemble –PSO</b>	<b>2.39215E-06</b>	<b>4.69295E-06</b>

**Table 6.15** Ensemble of PSO-ANFIS for Data (B)

Data (B)	MAE	RMSE
SBDS <sub>3</sub>	2.89157E-02	7.8021E-02
SBDS <sub>10</sub>	4.70906E-07	7.3868E-07
<b>Ensemble –PSO</b>	<b>4.62053E-07</b>	<b>7.2736E-07</b>

## 6.4 A Comparison Analysis of Combined Prediction Models

The results of the previous experiments in Section 6.1 were compared to determine the best Meta learning for the prediction of crude oil prices. The results are summarized in Table 6.16. Ensemble selection achieved the best results with MAE 1.420E-02 and RMSE 2.42E-02 when compared to Bagging using SMOreg.

**Table 6.16** Comparison among Meta learning models

Meta learning model	Data	MAE	RMSE	Sub-dataset	Time
Bagging SMOreg	B	1.840E-02	2.610E-02	SBDS10	00:00:10
Random Subspace SMOreg	A	2.210E-02	4.490E-02	SBDS9	00:01:06
Stacking	A	2.710E-02	5.510E-02	SBDS1	00:05:41
Voting	A	5.400E-02	7.670E-02	SBDS9	00:01:02
<b>Ensemble selection</b>	<b>A</b>	<b>1.420E-02</b>	<b>2.420E-02</b>	SBDS3	00:06:17

By comparing all the results obtained from Meta learning and ANFIS, the optimal result with the minimum MAE and RMSE values were derived from ANFIS with data (B). This set produced MAE = **4.70906E-07** and RMSE = **7.382E-07** for its RMSE with a competitive training time of 04 sec. Based on the performance comparison in Table 6.17, the Ensemble-PSO-ANFIS leads the other Ensemble approaches in terms lowest MAE and RMSE values.

**Table 6.17** Performance comparison between Ensemble prediction models

Ensemble prediction models	Data	MAE	RMSE
<b>Ensemble Average</b>	(A)	2.5011E-04	1.0970E-03
	(B)	1.4458E-02	3.9010E-02
<b>Ensemble-PSO-ANFIS</b>	(A)	2.3922E-06	4.6930E-06
	<b>(B)</b>	<b>4.6205E-07</b>	<b>7.2736E-07</b>

### 6.4.1 Comparison of the (Ensemble –PSO-ANFIS) Prediction Model Results with Other Machine Learning Approaches

In order to measure the performance of the GEM model using ANFIS-PSO method, a comparison is made with other machine learning methods presented in Chapter 3 and is shown in Table 6.18.

**Table 6.18** Comparison of models used in the literature to predict WTI crude oil price using the ANFIS-PSO Ensemble

Reference	Algorithm	Task	Performance
[93]	Hybrid AI method	Prediction of WTI crude oil price	RMSE =2.040E+00
[35]	Support Vector Machine	Prediction of WTI crude oil price	RMSE =2.192E+00
[177]	EMD-SVM-ELM	Prediction of WTI crude oil price	RMSE =4.770E-01 MAE =3.630E-01
[34]	Orthogonal wavelet support vector machine	Prediction of WTI crude oil price	MAE= 3.560E-04
[183]	GA-NN	Prediction of WTI crude oil price	RMSE =1.15E-03
[184]	Co-active neuro fuzzy inference system	Prediction of WTI crude oil price	RMSE= 2.0864E-04 MAE =8.5321E-04
[182]	<b>Ensemble –ANFIS-PSO</b>	<b>Prediction of WTI crude oil price</b>	<b>RMSE 7.2736E-07</b> <b>MAE =4.6205E-07</b>

## 6.5 Conclusions

This Chapter presented the main experiments using combined models and the contributions are summarized, as follows:

1. Meta-learning empirical results were derived from two parts: Bagging and Random subspace . Further, Ensemble selection, Voting, and Stacking which combines the outputs from several predictors were also presented.
2. ANFIS achieved better results than Meta-learning models and NNs approaches in terms of accuracy and training time.
3. In order to improve the results, all the four training and testing datasets were operated and used for the basic ensemble and generalized ensemble methods. The best

training result was obtained from the data that were trained using 80% training and 20% for testing and obtained a mean absolute error (MAE) value of 4.62053E-07, and root mean squared error (RMSE) value of 7.2736E-07 using Ensemble PSO -ANFIS for SBDS<sub>3</sub> and SBDS<sub>10</sub>.

From the implementation of combined prediction models, it is evident that VIX, WTI, GPNY, ER, and FC1 are the most important factors to determine the crude oil price and ANFIS is a good interpretable model to explore and explain crude oil market's *if-then* rules. Finally dataset (B) has the best percentages of training and testing.

Comparison with different results from the literature as presented in Table 6.18 further illustrates the effectiveness and superiority of the Ensemble method using ANFIS PSO for the prediction of WTI crude oil price.



# CHAPTER SEVEN

## 7. INFORMATION ENTERPRISE ARCHITECTURE FOR CRUDE OIL PRICING AND PREDICTION

### Overview

In Chapter 3, the enterprise architecture framework was discussed, its significance and objectives. This Chapter illustrates a novel information enterprise architecture for crude oil pricing and prediction based on Zachman framework. Accordingly, the main research question of this thesis is mentioned below:

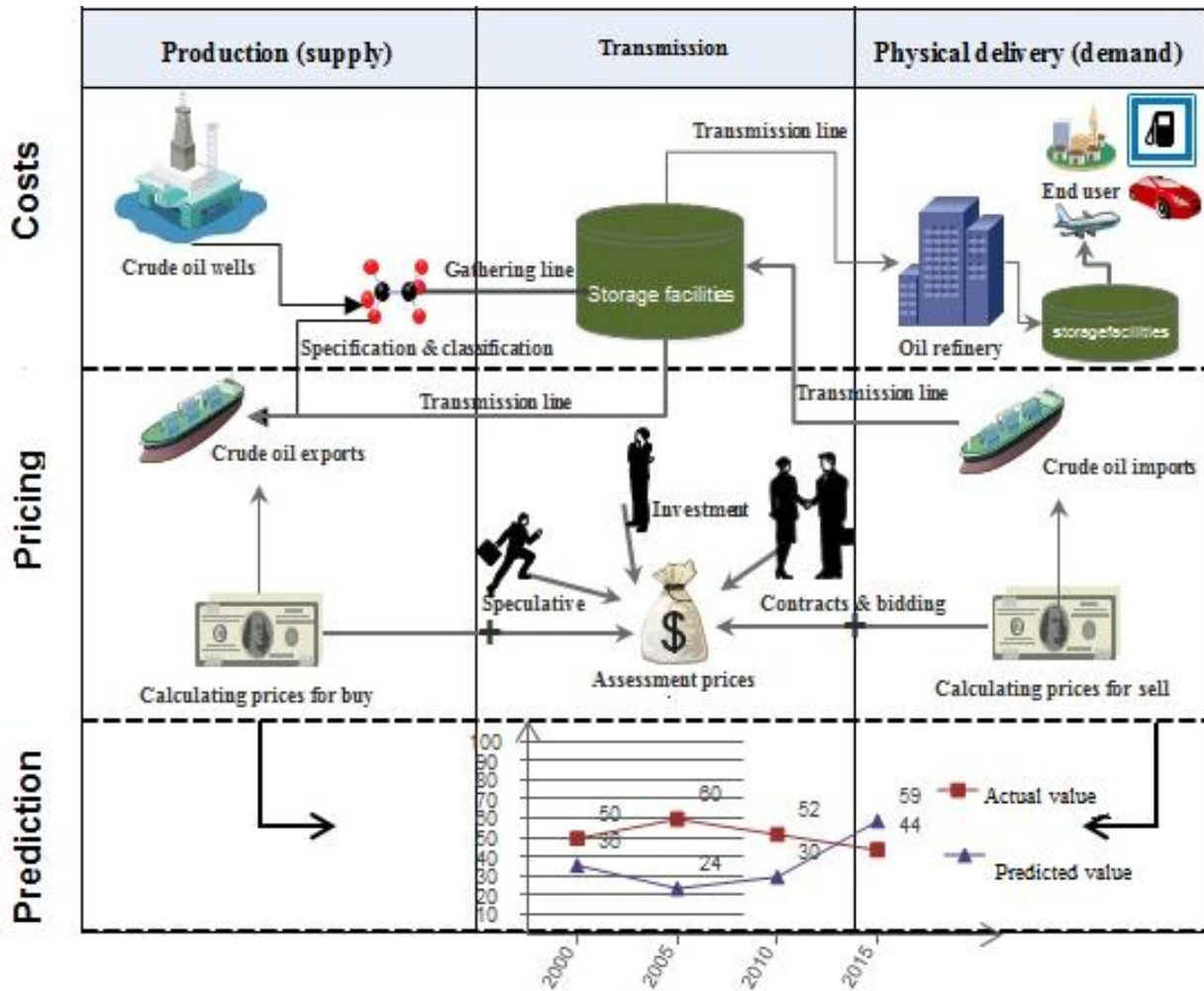
*How should crude oil pricing and prediction functions and tasks be organized under a common architecture?*

The results from this work helps us to understand the concept of the crude oil price under the umbrella of complicated factors, organizing varied processes, follow up the overlap between them and provides guideline to researchers in two areas of pricing and prediction for crude oil prices. An ensemble of adaptive neuro-fuzzy inference system using particle swarm optimization approach were employed, which is detailed in Sections 6.2 and 6.3. Empirical results depict that the architectural representation guided by the selected framework can provide a holistic view to the management of the crude oil prediction system. Moreover the identified patterns in a form of *if-then* rules can be easily interpreted when compared to other available models.

### 7.1 Information Enterprise Architecture for Crude Oil Pricing and Prediction

Crude oil consists of the remains of prehistoric animals and plants, which is exposed to extreme pressures and temperatures in the Earth's crust. Crude oil was adopted dramatically in the modern society and is affected by many factors and linked by an international network of thousands of producers, refiners, marketers, brokers, traders, and consumers for buying and selling physical volumes of crude oil. So there is great

need to organize, analyze and explain the systematic view of crude oil pricing and prediction. For this reason, we will develop EAF using the Zachman framework for crude oil pricing and prediction. For simplicity, we divided crude oil pricing and prediction system to three variety stages: **costs, pricing and prediction**. Also, we shall not focus on the economic details in each stage as each stage strongly correlated with each other. Figure 7.1 illustrates the different stages.



**Figure 7.1** Stages for crude oil pricing and prediction system

## 7.1.1 Stages of Crude Oil Pricing and Prediction System

### A. Costs Stage

This involves numerous activities such as:

- **Production:** A measure of the total cost to produce crude oil is the upstream costs. The upstream costs include the costs to operate and maintain oil wells and related equipment and services to get and extracting crude oil from well to the surface and finding costs of discovering and developing reserves of oil and the costs to purchase properties or acquire leases that might contain oil reserves [185] .
- **Transporting Crude Oil:** This focuses on moving crude oil from production to distribution and marketing their products to consumers or citizens, including pipelines, ports, tankers, barges, trucks, crude oil storage, retail storage, tanks and pumps [186] .
- **Refining:** The process by which crude oil is turned into products such as gasoline classification crude oil after extraction it from well to determine the type of crude to light or heavy usually traders characterize a crude by citing its source from which the crude was produced. API gravity (a measure of density) is a rough indication of distillation properties, which determine how much gasoline, kerosene, etc., can be distilled from the crude along with other factors and sulfur content which affects processing costs [187].

### B. Crude Oil Pricing

Oil price is influenced by many factors, which leads to making it a highly complex issue passing through different times by a number of the international crude oil pricing methods and theories as shown in the Table 7.1. Recently, benchmark oil works as a reference price for buyers and sellers of crude oil. There are several international benchmarks of pricing system such as in the North Sea (Brent), in US West Texas Intermediate (WTI), in the Middle East to Asia (Dubai) and the Far East is priced according to the (Tapis) from Malaysia and (Minas) of Indonesia. Our study focused on WTI, Brent and Dubai, because WTI is light-weight and has low sulfur content as these

properties make it excellent for making gasoline [43] and it is the mostly used for crude oil price in the United States and the oil industry in the US has a significant role to preserve the stability of the oil market and prices as a major power for production and consumption in the world. Brent 2/3 of the world (Europe or Africa) uses it as a benchmark for pricing the crude oil.

**Table 7.1** Development of international oil pricing methods [188]

Year	Oil pricing method	Price maker	Price basis
Before 1960s	Fixed Price	Multinational oil companies	Market price
1960-1970s	Fixed Price	Governments of Oil producing countries	Official Fixed Price
1970s -1980s	Market Linked	OPEC	Benchmark Oil
1990s and still	Formula pricing	Market	Benchmark Oil

In addition, commodity traders registered with the Commodity Futures Trading Commission (CFTC) are playing significant role in oil prices through the establishment of agreements to buy or sell oil at a specified date in the future for an agreed price by traders, these agreements known as bidding on oil futures contracts [189]. Commodities traders are divided into two categories:

1. Representatives of companies who actually use the oil.
2. Speculators who want to make money from changes in the price of oil.

Speculators invest in oil futures, essentially bets on how much oil will cost at a later date, These futures are traded in the New York Mercantile Exchange (or NYMEX), as well as the International Petroleum Exchange [190]. Government regulation for exporting and importing countries also has a big effect on oil prices. Recently the Government continues to find ways for people switching to power sources like wind and solar energy keeping the speculators from going completely out of control [190] .

While this topic is linked to the role of speculation in the determination of the oil price, it goes outside to world crisis, policy agendas etc, which recently dominated and the world disasters. Oil prices are determined largely by the activities of traders, however, there are potential factors related to market, These factors are supply-demand and oil reserves [49]. Other controllers or key players in the oil market are the Organization of the Petroleum Exporting Countries (OPEC) members, non-OPEC suppliers and non-

Organization for Economic Cooperation and Development (OECD) consumers [42]. Usually the price of crude oil is determined based on the characteristics of different types of crude oil, which is traded internationally by using the *formula pricing*, which is characterized as flexible. It can be written as:

$$Px = PR \pm D \quad (7.1)$$

Where  $Px$  is the price of crude  $x$ ;  $PR$  is the benchmark crude price; and  $D$  is the value of the price differential. These price differentials are linked to factors such as the difference in quality between the contracted and benchmark crude oils, transportation costs and so on determine the price of a certain type of crude oil at a discount or premium to the reference price (benchmarks) [42], that was previously mentioned. Here lies the importance of these indicators, such as West Texas Intermediate (WTI), Brent and Dubai- in the use as a basis for pricing oil the system, to price cargoes under long-term contracts or in the spot market transactions, for the settlement of financial contract settlement of derivative financial instruments such as swaps and for tax purposes by both oil companies, traders, banks and Governments [42].

Another integral part of the pricing system is to assess a price by oil pricing reporting agencies (PRAs). The two most significant PRAs in the oil market are Platts and Argus. The above discussion explained that the number of participants contribute in oil price behavior, which makes it interlinked and a complex market structure, which needs to be analyzed, described and explained using an IT framework as we aim for the development in this Section.

### **C. Crude Oil Prediction**

Monitoring the prices and the predicting of its price movement has been critically concerned and represent an integrated part of the decision-making process for the production, export, development and transport for the owners of industries and investors. For Government's the issue of predicting oil prices in the short term and long term has a great significance impact on the public policy of the state and national decision-making and to build a local budget. For researchers and academics, oil price prediction represents a prominent role and deeper understanding of the issues in the economy, the

financial theories, market hypotheses and pricing of consumer goods to the citizen [62]. There are many different theories and models as discussed in Chapter three, such as traditional statistical and econometric techniques [15] [16] have been commonly applied to crude oil price prediction [17]. However, several experiments have proved that the prediction performance might be very poor if one continued using these traditional statistical and econometric models [18].

This specific work is part of an ongoing research project focusing on explaining crude oil prediction using data mining techniques and the experiments identified and investigated effective predictive models that integrated to the architectural model for crude oil pricing and prediction as per the architecture. Thus, developing a prediction model in this framework means, reporting the best results of the data mining experiments presented in Sections 6.2 and 6.3.

In Section 4.6.2 the techniques used for the design of international framework for crude oil price prediction were displayed .

## 7.2 Zachman Framework for Crude Oil Pricing and Prediction

According to Zachman framework, the rows of this architectural model describe different perspectives and the concepts are explained in Table 7.2. Similarly, the columns of this architectural model describe several dimensions of the model, which are referred to as abstractions. They are described in Table 7.3

**Table 7.2** Rows of the crude oil pricing and prediction information architectural

Viewpoint	Description
<b>Scope (Planner's View)</b>	Describes the strategies, content and constraints of the crude oil pricing and prediction.
<b>Enterprise Model (Owner's View)</b>	Define Policies, procedures, goals, structure and processes that are used in crude oil pricing and prediction with an enterprise.
<b>System Model (Designer's View)</b>	Contains system requirements, objects, activities and functions, network that implements the crude oil pricing and prediction model. The system model states how the system is to perform its functions.
<b>Technology Model(Builder's View)</b>	Includes Database management system (DBMS) type requirements, Specifications of applications that operate on particular technology platforms and Specification of network devices and their relationships.
<b>As Built ( Sub-contractor View)</b>	Crude oil pricing and prediction rules constrained by specific components of the technology model that can be allocated to contractors for implementation

**Table 7.3** Dimension of the crude oil pricing and prediction information architectural

Viewpoint	Description
<b>Why(Motivation)</b>	It interprets crude oil pricing and prediction strategies and objectives into specific meaning.
<b>How (Processes)</b>	Processes to translate crude oil pricing and prediction requirements into more detailed implementation and operation definitions.
<b>What (Data)</b>	It is data to define and understand crude oil pricing and prediction requirements.
<b>Who (Entities)</b>	It explains who is related to crude oil pricing and prediction data and information management
<b>Where (Network)</b>	Places and locations related to crude oil pricing and prediction
<b>When (Time)</b>	It describes cycles and events related to crude oil pricing and prediction

### 7.2.1 Contextual View- Scope/ Planners Perspective

This perspective is concerned with the major components of the Information System to the planning of the crude oil pricing and prediction system and addresses its financial capability, constraints, and scope (what will be part of the Information System and what will not) [191]. In this case, the system is being described from the perspective of the Planner. Interview results presented as guided by the 6 motivating questions of Zachman framework, were used to populate the cells of this row. In addition, analysis of crude oil pricing and prediction, strategic plan was also used to define the content of cells at this viewpoint.

**A. Planner/Why:** The first cell lists motivation for this system in terms of Government’s policy, citizens’ services and economy aspect.

**B. Planner/How:** Cell two includes all processes that play role in crude oil pricing/prediction, including extraction and search operations, specification and classification crude oil, loading and shipping Cargoes, Physical delivery, Calculation and assessment of prices, predicting future prices and the Buy/sell (bid or offer) decisions and also the cross-functional processes that oversee and interconnect all the processes.

**C. Planner/What:** The third cell of scope addresses detailed planning of the list of data (things) that affects the direction and purpose of crude oil pricing/prediction

system. May take into consideration an array of issues including, but not limited to fundamentals of supply and demand, Stock Movements, Political issues, financial issues, environmental issues, Incoterm's used, transportation and Storage, legal and contract / agreement issues, swaps and so on.

**D. Planner/Who:** Cell four addresses users of the different departments including human resources, facilities, finance, sales, marketing, technical support, partners and legal contracts, customers and suppliers and the government authorities and users of these departments who operate on the data, such as laboratory analyst, investors, traders, brokers, buyers, consumers, suppliers, speculators, engineers, geologists, legal, researcher and workers.

**E. Planner/Where:** Cell five addresses the location and places that have a relationship with crude oil pricing/prediction system, which includes the oil-exporting country and oil-importing country with their institutions, companies and banks which operating in the field of crude oil, delivery and loading ports, international exchanges and price reporting agencies (PRAs), which participate in price assessment and tracks activities of price market value and market information including, firm bids and offers from companies, identities of buyers and sellers, confirmed prices, volumes, location, and stated trading terms during the entire day, and publishes a wide range of data relating to market value [42, 192-194]. This also includes the sites that have the control and effect of crude oil pricing system OPEC and OCED.

**F. Planner/When:** Cell six addresses the list of events, sequencing the timing of the processes and flows that significant to the crude oil pricing/prediction system. This includes Crude oil production cycle, Shipping events, timing of loading and delivery (spot-contract), market fluctuations such as time submit for bids and offers and market on Close (MOC) system [195], and the holiday calendar is annual compilation of the world's regional and national holidays, whose observation will impact the publishing of assessments and market such as new year's day, labor day, national day and so on [195]. Table 7.4 presents the Contextual View- Scope/Planners Perspective.



**Table 7.4** Contextual view-scope/planners perspective

<b>Scope (planner's perspective)</b>	<b>Motivation (Why)</b>	<b>Function (How)</b>	<b>Content (What)</b>
	<ul style="list-style-type: none"> <li>• Governments to secure the supply of oil to their nations.</li> <li>• Citizens in transportation – manufacturing-agriculture-heating –lightering.</li> <li>• Economic growth depends on energy</li> <li>• Building the government’s budget.</li> <li>• Hedge by strategic oil stocks</li> </ul>	<ul style="list-style-type: none"> <li>• Extraction and search</li> <li>• Specification and classification crude oil</li> <li>• Loading and shipping Cargoes</li> <li>• Physical delivery (spot and long term)</li> <li>• Calculating and Assessment prices</li> <li>• Buy/sell (bid or offer) and withdrawn</li> <li>• Prediction future prices</li> <li>• Legal process</li> </ul>	<ul style="list-style-type: none"> <li>• Fundamentals of supply and demand</li> <li>• Stock Movements</li> <li>• Political aspect</li> <li>• Financial issues</li> <li>• Environmental issues</li> <li>• Incoterms used</li> <li>• Transportation and Storage</li> <li>• Future prices</li> </ul>
	<b>People (Who)</b>	<b>Network (Where)</b>	<b>Time (When)</b>
	<ul style="list-style-type: none"> <li>• Exploration Crude oil department</li> <li>• Chemicals Laboratories</li> <li>• Transportation and Shipping department</li> <li>• Finance department</li> <li>• Marketing department</li> <li>• Legal contracts</li> </ul>	<ul style="list-style-type: none"> <li>• Price Reporting Agencies and International exchanges</li> <li>• Oil- exporting and importing country                             <ul style="list-style-type: none"> <li>- Banks</li> <li>- Petroleum ministry</li> <li>- Oil companies</li> <li>- Delivery and Loading port</li> </ul> </li> <li>• OPEC</li> <li>• OCED</li> </ul>	<ul style="list-style-type: none"> <li>• Crude oil production cycle</li> <li>• Timing of loading</li> <li>• Shipping events</li> <li>• Time of delivery</li> <li>• Time of loading</li> <li>• Time of prediction</li> <li>• Market fluctuations                             <ul style="list-style-type: none"> <li>- Timely submit for Bids and offers</li> <li>- Market on Close (MOC) system</li> <li>- Holiday calendar</li> </ul> </li> </ul>

### 7.2.2 Conceptual View- Owners Perspective

The aim of this perspective is to determine the full circumstances surrounding the crude oil pricing/prediction information, including details of crude oil quality and specifications, system sizes and dimensions, any location and loading/delivery information and lead times, motivation and people participating in crude oil pricing and prediction.

**A. Owner/Why:** In the first cell the major purpose for this specific system from an enterprise point of view is presented in terms of objectives.

**B. Owner/How:** In Column two within this perspective, workflow of the stakeholders interacting in the process of the crude oil pricing/prediction. Activity diagram in Figure 7.2 is used to represent the content of this cell [196]. It clearly explains the

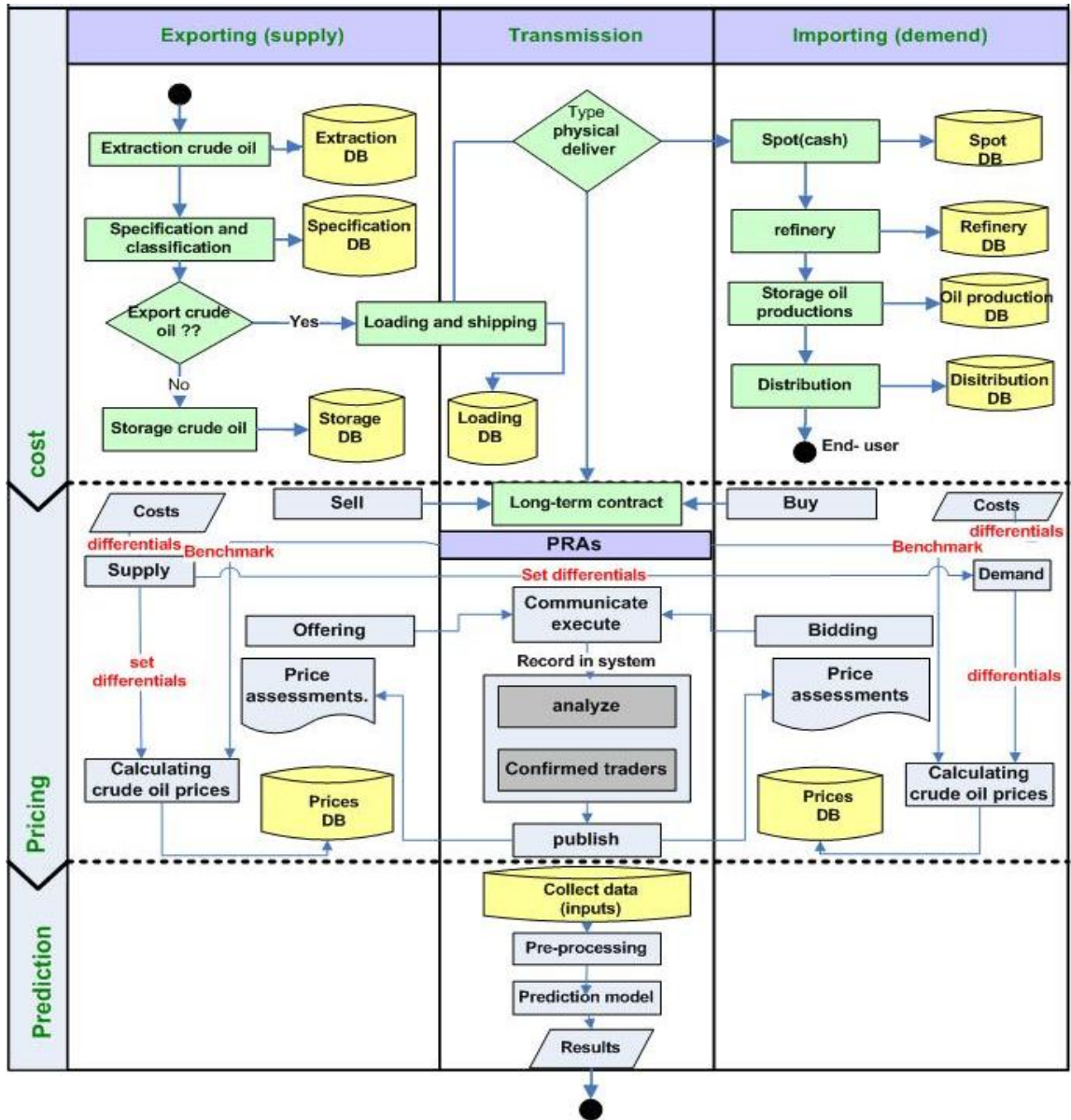
workflow in between the locations identified earlier. Its content is derived from the cell in the above row.

**C. Owner/What:** In this perspective, in the third column, the owner dictates his requirements of data list, important business entities, their attributes, relationship and factors which are applied or affect in the crude oil pricing and prediction, important data entities and their relations can be extracted based on scope/planner's perspective outputs [197] and entity dictionary can be used to represent this cell [196].

**D. Owner/Who:** The fourth cell as showed in Table 7.5 displays the general workflow in between and within crude oil pricing and prediction. As explained in [88, 196] organization chart or processes vs. organization Matrix can be used to model the content of this cell.

**E. Owner/Where:** Cell five defines the location of business nodes and place where stakeholders use from the system [196, 198]. This cell uses stereotypes of UML packages associated [196, 199] to modeling organizational units within location and as shown in Figure 7.3. Five important crude oil pricing and prediction locations are represented with their dependency relationship.

**F. Owner/When:** Finally, the sixth cell describes the time dimension of an enterprise and focuses on sequencing of the timing of process, events and flows significant to the crude oil pricing and prediction. According to [198], time dimension may be of two forms one of the forms represents the snapshot of a point in time and the other defines a period. A list of events with their suggested period is a preferred approach to represent this cell for easy understanding.

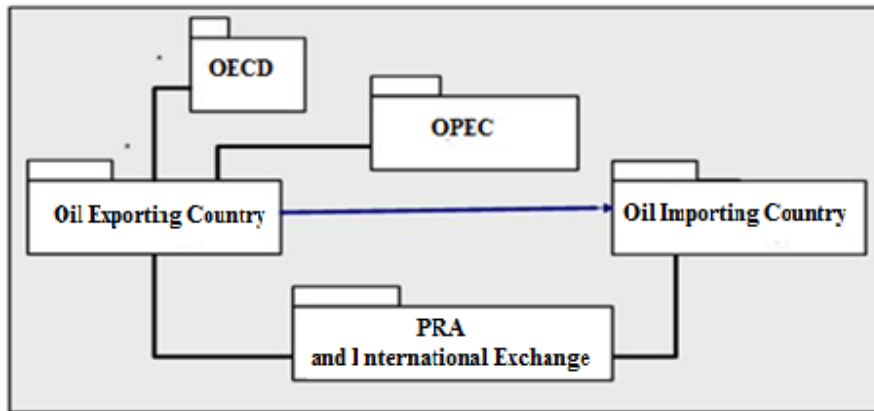


**Figure 7.2** Activity diagram representing the conceptual-function cell

**Table 7.5** Conceptual – people using process Vs organization matrix

	ES	SC	LS	PD	CA	B/S	LP	PP
Exploration crude oil department								
Chemicals laboratories								
Transportation and shipping department								
Finance department								
Marketing department								
Legal contracts								

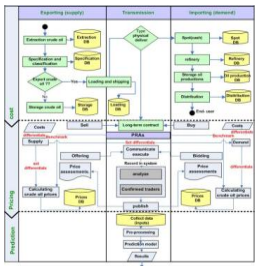
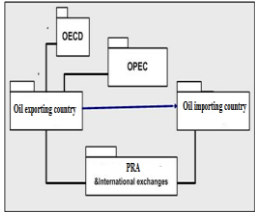
**Key:** *ES:* Extraction and search; *SC:* Specification and classification crude oil; *LS:* loading and shipping Cargoes; *PD:* Physical delivery (spot and long term); *CA:* Calculating and assessment prices; *BS:* Buy/sell (bid or offer) and withdrawn; *LP:* Legal process; *PP:* prediction crude oil prices.



**Figure 7.3** UML package representing conceptual-network cell

Based on the above discussion, architectural description of the second row of the crude oil pricing and prediction system is presented in Table 7.6

**Table 7.6** Conceptual view-enterprise/owners perspective

<b>Conceptual View /owners perspective</b>	<b>Motivation (Why)</b>	<b>Function (How)</b>	<b>Content (What)</b>																																																
	<ul style="list-style-type: none"> <li>• Single complete master source of Information includes settlement of their financial contracts, settlement of derivative instruments by banks and companies such as swap contracts.</li> <li>• Multi-view and advanced data analysis of the behavior and determinants of crude oil prices, as well as the linkages between physical and financial markets.</li> <li>• Facilitated information sharing.</li> <li>• Supporting market stability and security of supply and demand</li> </ul>	 <p>See Figure 7.2</p>	<ul style="list-style-type: none"> <li>• <b>World crude oil demand:</b> The amount of consumption; population, Future contracts.</li> <li>• <b>World crude oil supply:</b> OPEC production, NON-OPEC production, Future contracts, cost of production.</li> <li>• <b>Stock movement:</b> Reserves, Country</li> <li>• <b>Political issues:</b> War, world crisis</li> <li>• <b>Financial issues:</b> Exchange rate, investment banks, Hedge funds, Institutional investors, Retail investors, VIX</li> <li>• <b>Environmental issues:</b> Hurricanes, earthquakes, floods, explosions</li> <li>• <b>Incoterms used:</b> Ex-Works; (Free carrier; carriage Paid to; carriage and insurance; delivered at the terminal; delivered at Place; delivery duty paid)</li> <li>• <b>Transportation and storage:</b> Vessels classification, storage tanks capacities, quality pipelines, transportation and shipping cost.</li> <li>• <b>Prediction future prices:</b> Actual prices, predict the value (output), attributes (inputs)</li> </ul>																																																
	<b>People (Who)</b>	<b>Network (Where)</b>	<b>Time (When)</b>																																																
See Table 7.5	 <p>See Figure 7.3</p>	<ul style="list-style-type: none"> <li>• Crude oil production cycle (Daily, weekly, monthly, yearly)</li> <li>• Timing of loading (Daily, weekly, monthly, yearly)</li> <li>• Shipping events (daily, weekly, monthly)</li> <li>• Timing of delivery (daily, weekly, monthly)</li> <li>• Timing of prediction (daily, weekly, monthly, yearly)</li> <li>• Market fluctuations</li> <li>• Timely submit for Bids and offers (Immediately, hour, min, Sec)</li> <li>• Market on Close (MOC) (Immediately, hour, min, Sec)</li> <li>• Holiday calendar (day)</li> </ul>																																																	
	<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th></th> <th>E</th> <th>S</th> <th>L</th> <th>P</th> <th>C</th> <th>B</th> </tr> </thead> <tbody> <tr> <td>Exploration Crude oil department</td> <td style="background-color: #cccccc;"></td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Chemicals Laboratories</td> <td></td> <td style="background-color: #cccccc;"></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Transportation and Shipping department</td> <td></td> <td></td> <td style="background-color: #cccccc;"></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Finance department</td> <td></td> <td></td> <td></td> <td style="background-color: #cccccc;"></td> <td></td> <td></td> </tr> <tr> <td>Marketing department</td> <td></td> <td></td> <td></td> <td></td> <td style="background-color: #cccccc;"></td> <td></td> </tr> <tr> <td>Legal contracts</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td style="background-color: #cccccc;"></td> </tr> </tbody> </table>		E	S	L	P	C	B	Exploration Crude oil department							Chemicals Laboratories							Transportation and Shipping department							Finance department							Marketing department							Legal contracts							
	E	S	L	P	C	B																																													
Exploration Crude oil department																																																			
Chemicals Laboratories																																																			
Transportation and Shipping department																																																			
Finance department																																																			
Marketing department																																																			
Legal contracts																																																			

### 7.2.3 System Model: Logical View- Designer’s Perspective

The system model is placed in the third row of the crude oil pricing and prediction Framework, as the functionality of this fully attributed model is to reflect the

enterprise model of the above row. This perspective models helps to the specific the requirements, human-system interface issue, logical data model, geographical location and timing of the crude oil pricing and prediction system from the viewpoint of a system. The system analyst (Designer) represents the business in a disciplined form. The designer is an engineer or architect of the final product or service. The designer represents the laws of nature, the system or the logical constraints of the product or service design . Architecture of this perspective are presented in Table 7.7

**1. Designer/Why:** The first cell presents the reason of the system in terms of functional requirements. Using data from the interviews with crude oil experts to consider in the analysis of the cells above in defining the content of this cell.

**2. Designer/How:** Column two represents a layered architectural design of crude oil pricing/prediction and discusses the processes that have contribution and an impact on crude oil pricing/prediction. This includes extraction/search, Specification/classification, crude oil loading/shipping cargoes, calculation and assessment of prices, buy/sell (bid or offer) and predict future prices, Figure 7.4 illustrates the data flow diagram, which includes three presentation or end users view, layers business logic and data storage respectively.

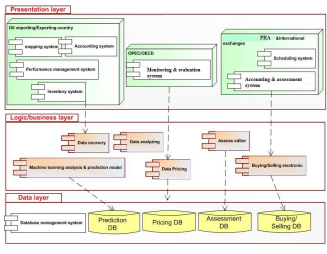
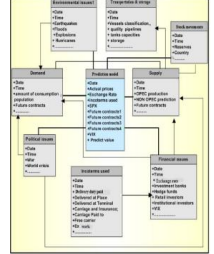
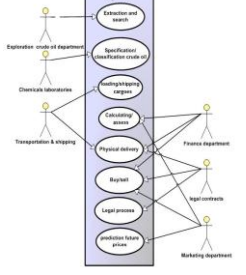
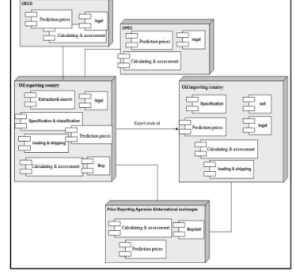
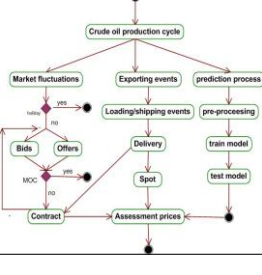
**3. Designer/What:** Column three (Logical-Content) describes the business structure, including business entities and their relationship. Example models of this cell could be a business entity relationship diagram that models the business concepts, entities and attributes as shown in Figure 7.5. The content of this cell could be associated with other entities mentioned in the owner's perspective [197].

**4. Designer/Who:** Cell four describes the structure and the contents of user interactions with the system. This cell defines the “Roles”, Roles are assigned to users who perform a job with the responsibilities. It can be modeled using Systems vs. Roles Matrix or UML Use Case diagram. Accordingly, Case model in Figure 7.6 is used to represent the proposed user interaction in crude oil pricing and prediction situation. Its content is derived from the above cell and analysis of the function perspective of the systems view.

5. **Designer/Where:** Cell four in this perspective addresses the available nodes of a whole system/enterprise and logical links in between them. In the crude oil pricing and prediction system the exporting country, importing country, OPEC, OECD and Price Reporting Agencies and International exchanges linkage could be a satellite link connecting two countries or countries with their respective modules. System diagram and UML component are also proposed, Deployment diagram, as displayed in Figure 7.7. Using location stereotype of packages is a preferred modeling technique to represent this cell [199].

6. **Designer/When:** In column six, designer addresses the impact of time on the system and events to be scheduled in a timely manner depending on the crude oil pricing and prediction information and the events of the business, which causes specific data transformations and entity state changes to take place in this cell by using UML. Figure 7.8 shows a State Diagram to capture the relationship of time from the Designers perspective [200].

**Table 7.7** System model: View-crude oil prediction /pricing designer’s perspective

		Motivation (Why)	Function (How)	Content (What)
System model - crude oil pricing and prediction system	Designer’s perspective	<ul style="list-style-type: none"> <li>• Provide a platform for information sharing.</li> <li>• Enable periodic descriptive, exploratory and predictive analysis.</li> <li>• Effective communication channels and interaction with the world via technology and media to broaden and expand global energy community.</li> </ul>	 <p style="text-align: center;">See Figure 7.4</p>	 <p style="text-align: center;">See Figure 7.5</p>
		<p style="text-align: center;"><b>People (Who)</b></p>  <p style="text-align: center;">See Figure 7.6</p>	<p style="text-align: center;"><b>Network (Where)</b></p>  <p style="text-align: center;">See Figure 7.7</p>	<p style="text-align: center;"><b>Time (When)</b></p>  <p style="text-align: center;">See Figure 7.8</p>



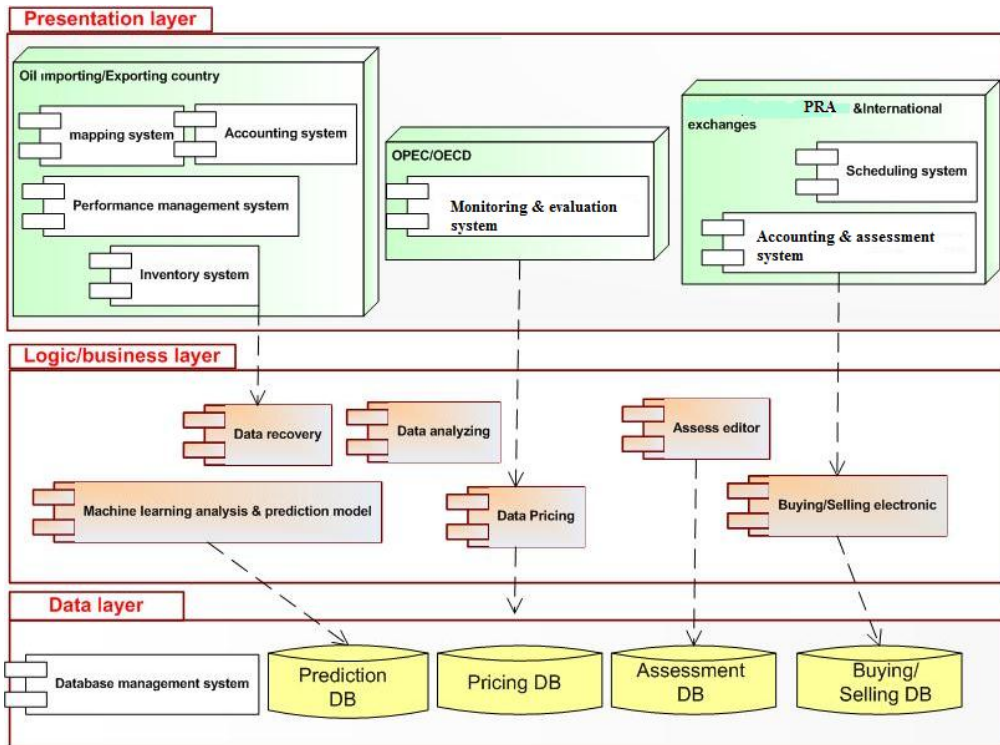


Figure 7.4 Layer architecture representing logical –function cell

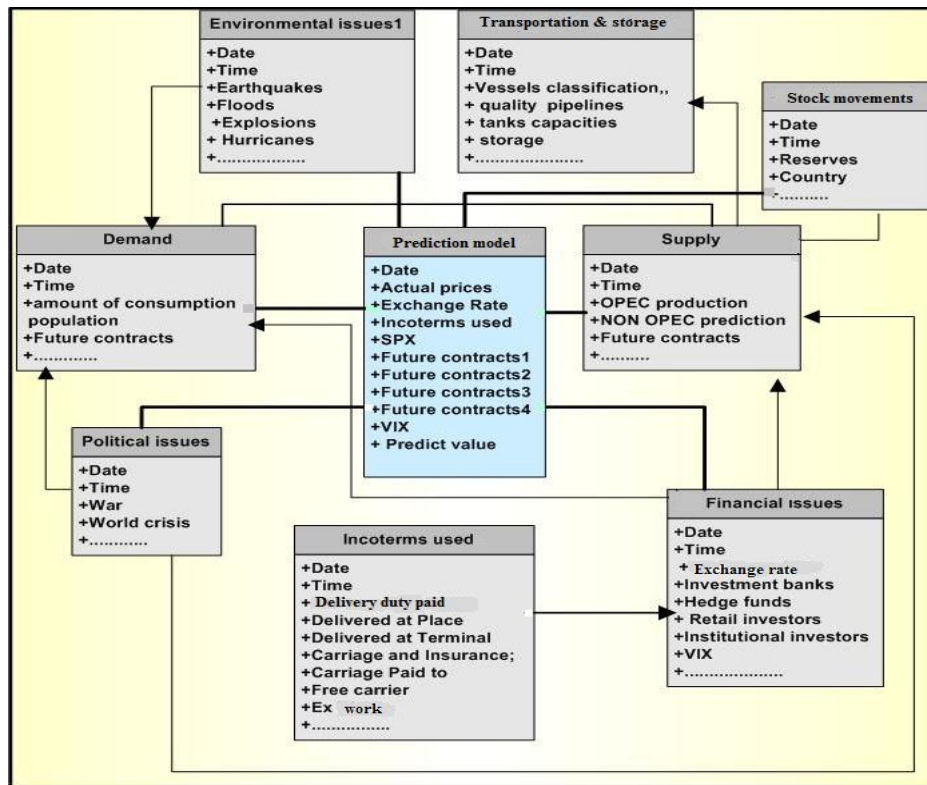
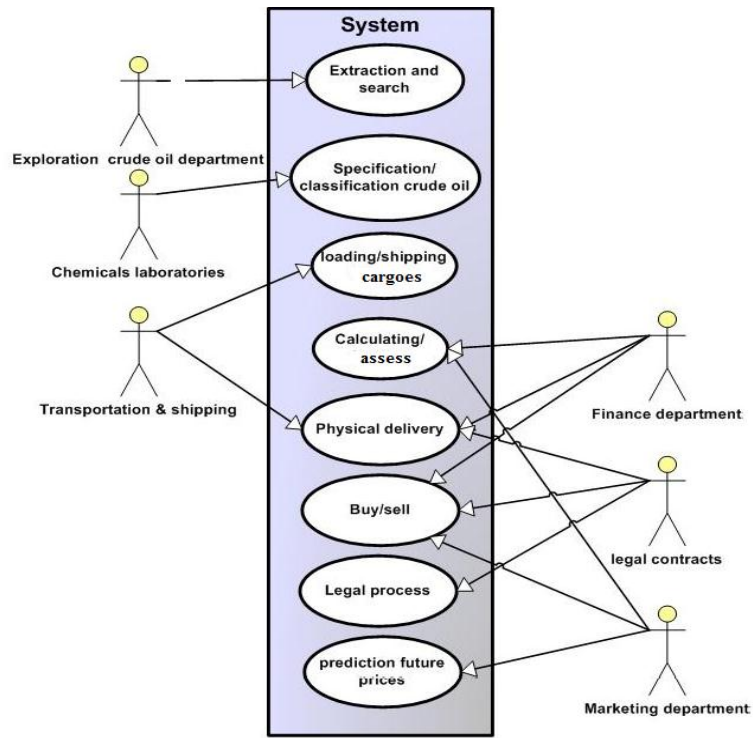
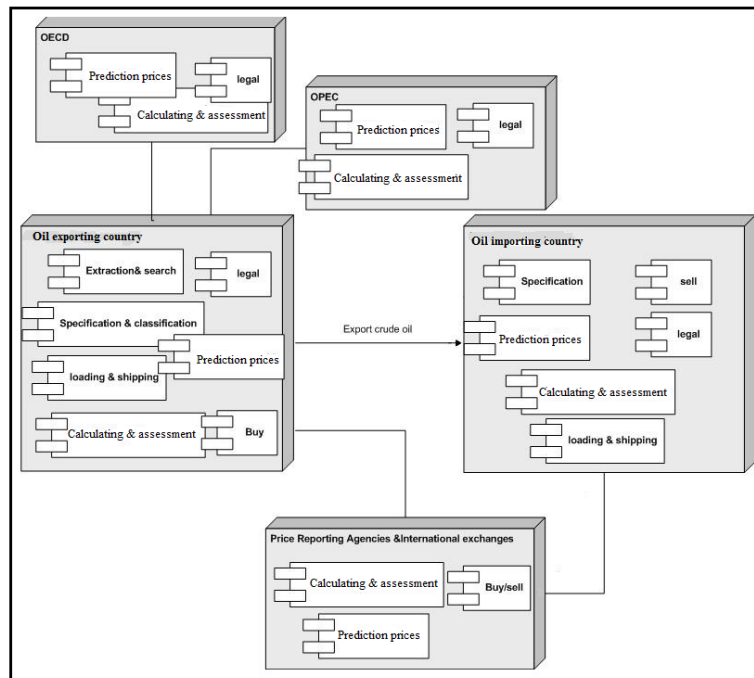


Figure 7.5 Crude oil pricing and prediction model representing logical-content cell

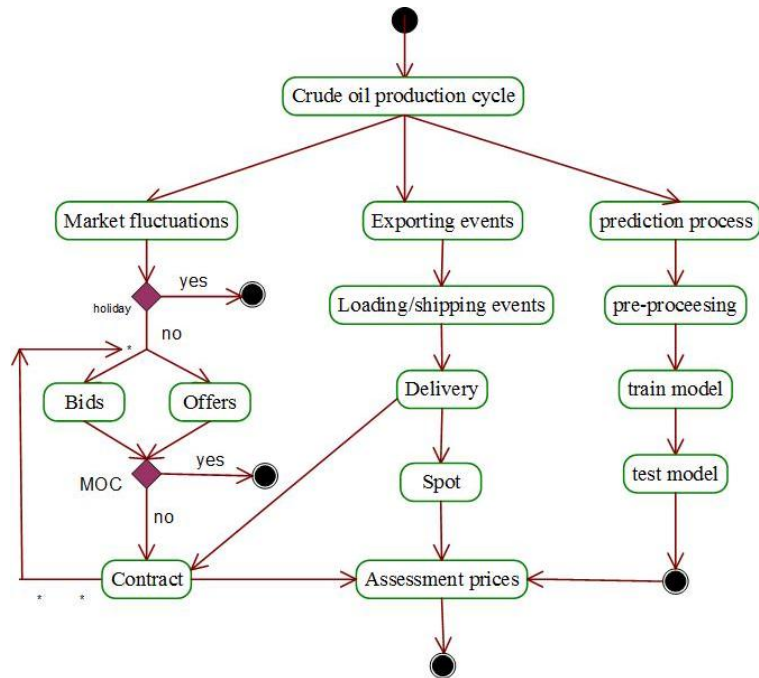




**Figure 7.6** Use case model representing logical-people cell



**Figure 7.7** Deployment diagram representing logical-network



**Figure 7.8** State diagram representing logical– time cell

## 7.2.4 Technology Model - Builder’s Perspective

The technology model of the Zachman Framework represents the physical representation of things of the enterprise. The builder of this model depends on the technology chosen for implementation by understanding its environment and integrates all the data sources developed from different platforms and operating systems for providing a common enterprise wide view [198].

**A. Builder/Why:** Column one deals with the specification of the crude oil pricing and prediction rules. Decision Table or FOL (First Order Logic) can represent this cell [201] according to previous results in prediction using hybrid techniques (ANFIS) as presented in Section 6.3. The formal representation of the knowledge is “*if-then*” statement when attributes are used in antecedent parts and crude oil prices are used in consequent. Another concept in our system is pricing, where builder can address the relation between many factors to determine the situation of crude oil price such as the basic economics principles, if the supply is high and demand is low, then the crude oil price will be low; the inverse, is also true [190]. Finally through this cell, other objec-

tives and strategies for crude oil pricing and prediction are represented in this decision table.

**B. Builder /How:** Column two answers the question about what are the processes that is needed for computerization of the crude oil pricing/prediction system depending on the sensitivity of the crude oil price data, and the importance of operations, which contribute in this system. The answer will include secret operations such as encryption, decryption, monitoring process, recovery process, scheduling, evaluation process, accounting process and the machine learning analysis and prediction model.

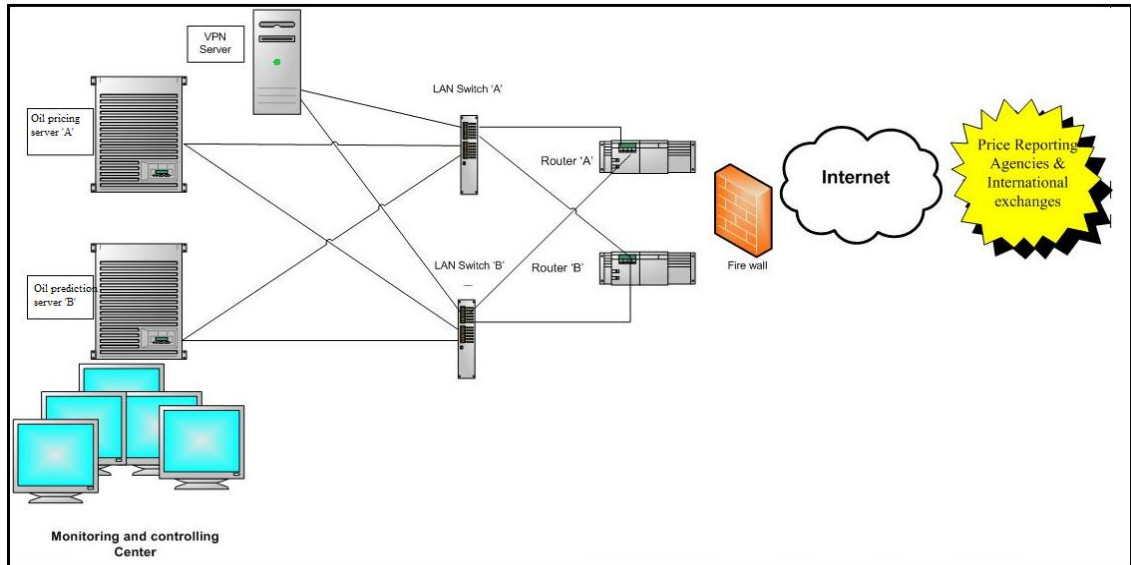
**C. Builder /What:** Column three prepares a physical data model that includes a variety of technological description according to previous data presented in row 3 such as database schema, storage management, data encryption, Multimedia.

**D. Builder /Who:** Column 4 is achieved by the “Users”, the “Roles” from Row 3, the physical expression of the workflow of the enterprise, including specific individual and presentation architecture such as work flow, client interface and presentation format.

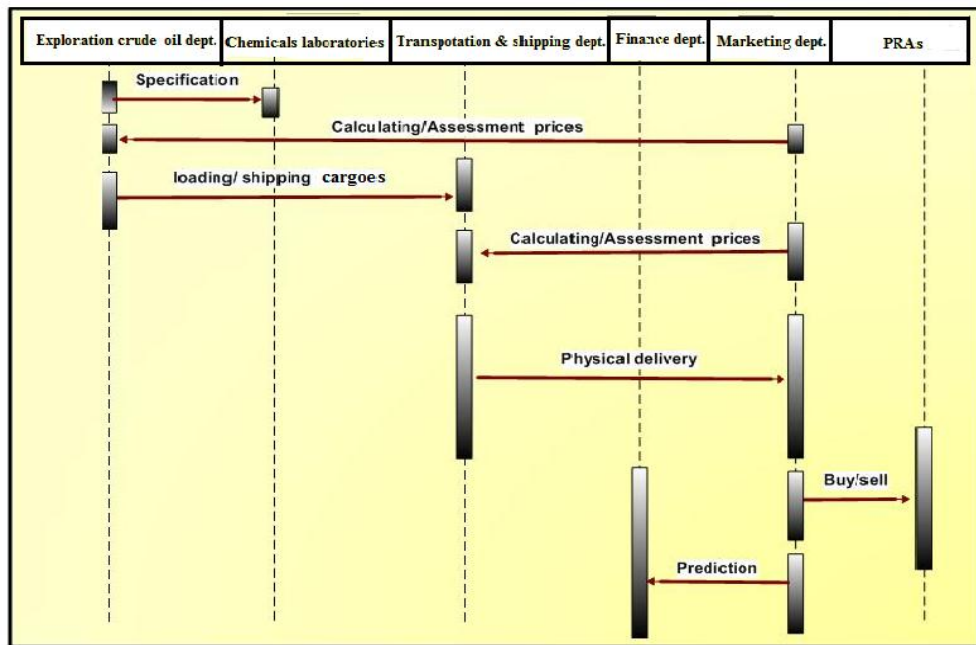
**E. Builder /Where:** Column five secures the business nodes by implementing access control devices such as various crude oil pricing and prediction servers to provide database, software application and machine learning prediction model. The communication links between the server nodes handles the data of column 3 by implementing protocols for connection between the client and the server, firewall to protect the network from any attack and VPN server to provide the secure connection over the internet in order to login to the oil market Figure displays network structure, the equipment’s have been duplicated in order to have full redundancy. Figure 9 depicts a diagram of the sample system showing the hardware and software requirements that would be listed as a deliverable to satisfy the requirements for this cell of the Zachman Framework [83].

**F. Builder /When:** Finally, in the sixth cell, the Builder specifies events and timing at the logical level and a specification for event sequence in a more detail depending on the crude oil pricing and prediction information. Sequence diagram as shown in Figure

7.10 was used to represent timing in crude oil pricing and prediction system, Sequence diagrams are recommended in the literature to determine the content of this cell [88, 200]. Technology Model for Crude oil pricing/prediction Builder's Perspective is illustrated in Table 7.8.

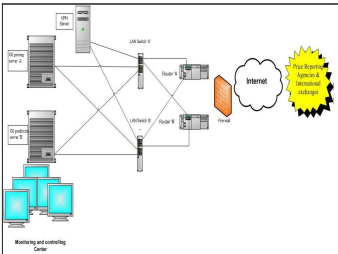
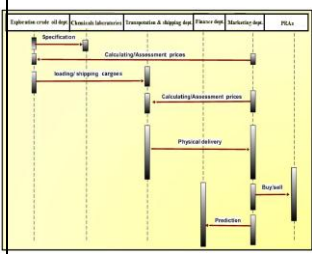


**Figure 7.9** Network architecture describing Technology –network



**Figure 7.10** Sequence diagram representing logical– function

**Table 7.8** Technology model-Crude oil prediction /pricing builder’s perspective

Technology model - crude oil pricing/prediction builder’s perspective	<b>Motivation (Why)</b>	<b>Function (How)</b>	<b>Content (What)</b>
	Decision table: <ul style="list-style-type: none"> <li>• <i>If-then</i> rules for prediction crude oil prices.</li> <li>• <i>If-then</i> rules for constructing crude oil pricing/prediction objectives and strategies.</li> </ul>	<ul style="list-style-type: none"> <li>• Geologic mapping</li> <li>• Monitoring process</li> <li>• Scheduling</li> <li>• Evaluation process</li> <li>• Accounting process</li> <li>• Performance management</li> <li>• Inventory process</li> <li>• Recovery process</li> <li>• Assess editor</li> <li>• Machine learning analysis</li> <li>• Prediction model</li> </ul>	<ul style="list-style-type: none"> <li>• Database schema</li> <li>• Storage management</li> <li>• Data encryption</li> <li>• Multimedia</li> </ul>
	<b>People (Who)</b>	<b>Network (Where)</b>	<b>Time (When)</b>
	<ul style="list-style-type: none"> <li>• Work flow</li> <li>• Client interface</li> <li>• Presentation format</li> </ul>	 <p style="text-align: center;">See Figure 7.9</p>	 <p style="text-align: center;">See Figure 7.10</p>

### 7.2.5 Detailed Representations - Subcontractor’s Perspective

Present specifications components of the technology model that can be allocated to contractors for implementation, as shown in Table 7.9

**A. Subcontractor /Why:** The first cell describes the technical requirements for crude oil pricing and prediction and specification of the business rules.

**B. Subcontractor /How:** Column two of the subcontractor row defines the programs and the codes that are derived from row 3, 4.

**C. Subcontractor /What:** Column three addresses the definition of all data languages specified in the “Physical Data Model” of Row 4 and required for implementation such as Data base modules and Firewall backup to protect the network and software from any attack.

**D. Subcontractor /Who:** column four from the subcontractor perspective identifies the user access permissions [200], and determine restrictions to specific functions provided by the system. Crude oil pricing and prediction model is like a financial trading system and must be designed to distinguish the type of user entering the system. At the program level, the system would present user functionality to those who are logging as system administrators according to their tasks.

**E. Subcontractor /Where:** Column five addresses the physical data network components, specification of node addresses and protocols for communicating among nodes such as LAN switches, routers and servers having the oil pricing database and software applications, monitoring and controlling the center.

**F. Subcontractor /When:** Column five addresses the “Timing Definition”, which defines the market fluctuations time cycle by calculating the time taken to buy, sell and Time stamp.

**Table 7.9** Detailed representations of crude oil prediction and pricing

<b>Detailed representations crude oil prediction and pricing subcontractor’s perspective (Planner’s Perspective)</b>	<b>Motivation (Why)</b>	<b>Function (How)</b>	<b>Content (What)</b>
	<ul style="list-style-type: none"> <li>• Integrate multiple application and infrastructure platforms for different of the universal crude oil pricing/prediction systems</li> <li>• Provide electro editors on the day to assess the market daily.</li> <li>• Provide the majority of market data used by the assess editor.</li> </ul>	<ul style="list-style-type: none"> <li>• Code</li> <li>• Procedures</li> <li>• DBMS stored</li> <li>• Program language</li> <li>• Machine learning program such as Matlab and WEKA</li> <li>• Backup devices</li> </ul>	<ul style="list-style-type: none"> <li>• Database modules</li> <li>• Firewall backup</li> <li>• Data language</li> <li>• Documentation</li> <li>• Configuration files</li> <li>• Data definition script</li> <li>• Software application for pricing prediction</li> <li>• SQL</li> </ul>
	<b>People (Who)</b>	<b>Network (Where)</b>	<b>Time (When)</b>
	<ul style="list-style-type: none"> <li>• System operation permissions</li> <li>• Screens</li> <li>• Security architecture</li> </ul>	<ul style="list-style-type: none"> <li>• LAN switches</li> <li>• Routers</li> <li>• Servers having the oil pricing database</li> <li>• Monitoring and controlling center</li> </ul>	<ul style="list-style-type: none"> <li>• Holiday schedule</li> <li>• Time stamps</li> <li>• Time order setting conditions</li> </ul>

### 7.2.6 Functioning System-Real System Crude Oil Prediction and Pricing

Finally, the functioning enterprise depicts the operational system that is under consideration. Row 6 is therefore the reality and it is what the users of the enterprise’s product or service experience physically [202]. Table 7.10 presents this perspective.

**Table 7.10** Real system crude oil pricing/prediction

<b>Real system (crude oil prediction/ pricing)</b>	<b>Motivation (Why)</b>	<b>Function (How)</b>	<b>Content (What)</b>
	Strategies and objectives for crude oil prediction/ pricing	Functions and methods of crude oil pricing and predict future prices	Data effects in crude oil prices and prediction.
	<b>People (Who)</b>	<b>Network (Where)</b>	<b>Time (When)</b>
	Major players and organizations Crude oil price	Exporting, importing countries Network between them	Crude oil prices spot and future Schedules

## 7.3 Conclusions

This Chapter developed information enterprise architecture for crude oil pricing and prediction using Zachman framework. Six perspectives and discussed different aspects of international crude oil pricing and prediction were presented, which leads to a deeper understanding of the comprehensive structure of the crude oil market and constructing information technology based infrastructure in this field.

# CHAPTER EIGHT

## 8. CONCLUSIONS AND RECOMMENDATIONS

### Overview

This Chapter provides a summary of the main achievement and major findings of this research. It also identifies the main contributions of the novel information enterprise architecture for crude oil pricing and prediction developed in this thesis. Finally, future research topics are identified that might be of interest for further investigation.

### 8.1 Summary

Development of the economy depends on the different resources of energy that even most of economic sectors such as commercial, industrial and transportation are impossible to operate without energy. Among the different energy sources, oil plays a significant role to become the most efficient and important source of energy. Oil embodies a vital role in the national and international economies as the backbone and the source of indispensable raw inputs for numerous industries and represents a major component in many manufacturing processes such as plastics and chemicals.

Crude oil price prediction is a challenging task due to its complex nonlinear and chaotic behavior. There is a great need for oil price volatility measuring and modeling of oil price chaotic behavior. During the last couple of decades, both academicians and practitioners have devoted proactive knowledge to address this issue.

The basic concepts in the oil industry as a basis for information retrieval to understand the circumstances and conditions surrounding of the crude oil prediction were explained. Then the crude oil chaotic behavior was explained and the evidence for the presence of chaos and non-linear dynamics in oil prices according to global events. Four fundamental factors that had played an important role in oil prices were presented. The basic concepts of machine learning/data mining technology was presented and their applications to investigate the prediction of crude oil prices practically. Further Zach-



man framework approach was presented, which is used to develop an information enterprise architecture for crude oil pricing and prediction.

In our empirical research, a dataset that comprises of 3337 records as instances and 14 attributes were used, which are intended to represent the information of economic factors that play an important role in the oil market. The dataset is daily data and covers the period from 4 January 1999 to 10 October 2012. Using daily price is another challenge as it might be more complex in terms of higher level of noise and represent the real oil markets.

Pre-processing steps, which were absent from previous research, were discussed in the literature review. Several aspects of initial preparation of data before starting to construct models were used such as feature selection algorithm, data partition, and normalization, which lead to creating 10 sub datasets containing different factors.

Numerous machine-learning methods were experimented based on their performance for the prediction of crude oil prices. In order to improve the results of direct prediction models, different styles of combined models were implemented such as Meta learning followed by hybrid prediction and finally Ensemble model. From the implementation of combined prediction models, it is evident that VIX, WTI, GPNY, ER, and FC1 are the most important factors to determine the crude oil price and ANFIS is a good interpretable model to explore and explain crude oil market's *if-then* rules. The best training result was obtained from the data that were trained using 80% training and 20% for testing and obtained a mean absolute error (MAE) value of 4.62053E-07, and root mean squared error (RMSE) value of 7.2736E-07 using Ensemble PSO –ANFIS with a competitive training time of 04 sec.

It is standard to include time lags in oil market studies; for example, Golombek, et al. [97] used 5 lags for their model whereas Kilian [203] used 12 lags. Time lags from 1 to 12 were used but all of them failed to improve results

## 8.2 Contributions of the Research

- The main contribution of this research is the development of a new information enterprise architecture framework for crude oil pricing and prediction to organize, analyze and explain the behavior of crude oil pricing and prediction.
- Investigation of the suitability Zachman Framework in oil price prediction domain is also another aspect. Zachman Framework has been used in areas like network security planning, education services delivery, determining the content of digital libraries etc. This work extends the use of the framework for oil price prediction domain, which will improve understanding and facilitate communication among oil sectors and organizations.
- Experimental evidence and illustrations showing the effectiveness and superiority of the proposed Ensemble method using ANFIS PSO for the prediction of WTI crude oil price.
- A novel comparison was conducted on the one hand between direct models and on the other hand among combined prediction models in order to investigate the appropriate algorithm for oil price prediction problem. This comparison is useful for a better understanding of the performance of various models and can be used to create effective models for oil price prediction and other similar problems.

## 8.3 Future Research

1. The proposed model focused on economic factors and future oil prices as inputs to predict crude oil price because until recently, central banks and international organizations tend to rely exclusively on the oil futures curve to predict the price of oil. In addition, recent research demonstrates that models include the economic determinants of the price of oil, such as changes in oil inventories, oil production for example gasoline or heating oil and global real economic activity, may provide more accurate out-of-sample forecasts [204] [205]. Whereas evidence indicates that crude oil prices were

significantly affected by events of uncertainties such as during the First Gulf War, the Venezuelan unrest, the Iraq War, the Asian financial crises, and the world financial recession [206]. The future research could examine the impact of uncertainties in an integrated manner with this study.

**2.** Anomaly detection in the data refers to detecting any unexpected changes in a subsequence of a set of adjacent time series [207]. This problem occurs in wide variety application areas such as climate patterns, public health, and oil prices. As we mentioned early, in June 2014 oil prices suffered from the declining price in an anomaly form until now. Therefore the objective, in this case, to find a subsequence in the time series showing the highest difference from all other subsequences to avoid insufficiently representative in the future, specifically in normal and an anomalous boundary which is often not precise.

**3.** Information enterprise architecture for crude oil pricing and prediction can be extended in future to include petroleum ministries, finance institutions and all companies working in the crude oil field in detail to provide complete description of the crude oil pricing and prediction system and to develop a framework that supports strongly with visualization of prediction model (results) and information enterprise architecture for crude oil price, so that the end-user can fully understand and deal with the changes and dynamics within the data.

## REFERENCES

- [1] E.L. (2014). *Why the oil price is falling*. Available: <http://www.economist.com/blogs/economist-explains/2014/12/economist-explains-4> [Accessed 23/2/2015]
- [2] S. Dunn, "Hydrogen futures: toward a sustainable energy system," *International journal of hydrogen energy*, vol. 27, pp. 235-264, 2002.
- [3] J. Tuo and S. Yanbing, "Summary of World Oil Price Forecasting Model," in *Knowledge Acquisition and Modeling (KAM), 2011 Fourth International Symposium on*, 2011, pp. 327-330.
- [4] J. L. Williams. (2011). *Oil Price History and Analysis*. Available: <http://www.wtrg.com/prices.htm> [Accessed 10/9/2014]
- [5] H. Pan, I. Haidar, and S. Kulkarni, "Daily prediction of short-term trends of crude oil prices using neural networks exploiting multimarket dynamics," *Frontiers of Computer Science in China*, vol. 3, pp. 177-191, 2009.
- [6] T. Brahmairene, J.-C. Huang, and Y. Sissoko, "Crude oil prices and exchange rates: Causality, variance decomposition and impulse response," *Energy Economics*, 2014.
- [7] F. Wang and S. Wang, "Analysis on impact factors of oil price fluctuation in China," in *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on*, 2011, pp. 6146-6150.
- [8] P. R. B. E. Sector. ( 2010). *Review of Issues Affecting the Price of the Crude Oil* Available: <http://www.nrcan.gc.ca/files/energy/pdf/eneene/pdf/pcopdp-eng.pdf> [Accessed]
- [9] L. Kilian and D. P. Murphy, "The role of inventories and speculative trading in the global market for crude oil," *Journal of Applied Econometrics*, vol. 29, pp. 454-478, 2014.
- [10] R. A. Lizardo and A. V. Mollick, "Oil price fluctuations and US dollar exchange rates," *Energy Economics*, vol. 32, pp. 399-408, 2010.
- [11] C. Morana, "Oil price dynamics, macro-finance interactions and the role of financial speculation," *Journal of banking & finance*, vol. 37, pp. 206-226, 2013.
- [12] X. Zhang, Q. Wu, and J. Zhang, "Crude oil price forecasting using fuzzy time series," in *Knowledge Acquisition and Modeling (KAM), 2010 3rd International Symposium on*, 2010, pp. 213-216.
- [13] M. Sompui and W. Wongsinlatam, "Prediction Model for Crude Oil Price Using Artificial Neural Networks," *Applied Mathematical Sciences*, vol. 8, pp. 3953-3965, 2014.
- [14] L. Yu, S. Wang, and K. Lai, "A rough-set-refined text mining approach for crude oil market tendency forecasting," *International Journal of Knowledge and Systems Sciences*, vol. 2, pp. 33-46, 2005.
- [15] B. Abramson and A. Finizza, "Probabilistic forecasts from probabilistic models: a case study in the oil market," *International Journal of forecasting*, vol. 11, pp. 63-72, 1995.

- [16] A. Coppola, "Forecasting oil price movements: Exploiting the information in the futures market," *Journal of Futures Markets*, vol. 28, pp. 34-56, 2008.
- [17] L. Yu, S. Wang, and K. K. Lai, "Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm," *Energy Economics*, vol. 30, pp. 2623-2635, 2008.
- [18] A. S. Weigend, "Time series prediction: forecasting the future and understanding the past," *Santa Fe Institute Studies in the Sciences of Complexity*, 1994.
- [19] S. Kulkarni and I. Haidar, "Forecasting model for crude oil price using artificial neural networks and commodity futures prices," *arXiv preprint arXiv:0906.4838*, 2009.
- [20] Y. Fan, Q. Liang, and Y.-M. Wei, "A generalized pattern matching approach for multi-step prediction of crude oil price," *Energy Economics*, vol. 30, pp. 889-904, 2008.
- [21] W. Jun, L. Zhi-bin, and S. Qiong, "Oil Price Forecasting based on Hierarchical Support Vector Machine [J]," *Computer Applications of Petroleum*, vol. 63, pp. 5-8, 2009.
- [22] A. Azadeh, M. Moghaddam, M. Khakzad, and V. Ebrahimipour, "A flexible neural network-fuzzy mathematical programming algorithm for improvement of oil price estimation and forecasting," *Computers & Industrial Engineering*, vol. 62, pp. 421-430, 2012.
- [23] A. Bredenberg. (2012). *The Damage Done in Transportation -- Which Energy Source Will Lead to the Greenest Highways?* Available: <http://news.thomasnet.com/imt/2012/04/30/the-damage-done-in-transportation-which-energy-source-will-lead-to-the-greenest-highways> [Accessed]
- [24] B. Graham. (2012). *Joy Global: A Misunderstood Cyclical?* Available: <http://www.rationalwalk.com/?p=12709> [Accessed 18/8/2014]
- [25] K. He, L. Yu, and K. K. Lai, "Crude oil price analysis and forecasting using wavelet decomposed ensemble model," *Energy*, vol. 46, pp. 564-574, 2012.
- [26] L. E. Svensson, "Oil prices and ECB monetary policy," *Briefing Paper for the Committee on Economic and Monetary Affairs of the European Parliament*, 2005.
- [27] P. K. Narayan, S. Sharma, W. C. Poon, and J. Westerlund, "Do oil prices predict economic growth? New global evidence," *Energy economics*, vol. 41, pp. 137-146, 2014.
- [28] S. Rafiq, R. Salim, and H. Bloch, "Impact of crude oil price volatility on economic activities: An empirical investigation in the Thai economy," *Resources Policy*, vol. 34, pp. 121-132, 2009.
- [29] D. W. Jones, P. N. Leiby, and I. K. Paik, "Oil price shocks and the macroeconomy: what has been learned since 1996," *The Energy Journal*, pp. 1-32, 2004.
- [30] S. H. Saghaian, "The impact of the oil sector on commodity prices: correlation or causation?," *Journal of Agricultural & Applied Economics*, vol. 42, p. 477, 2010.
- [31] J. Wang, W. Xu, X. Zhang, Y. Bao, Y. Pang, and S. Wang, "Data Mining Methods for Crude Oil Market Analysis and Forecast," *Data Mining in Public and Private Sectors: Organizational and Government Applications*, vol. 184, 2010.

- [32] J. D. Hamilton, "Understanding crude oil prices," National Bureau of Economic Research 2008.
- [33] L. Yu, Y. Zhao, and L. Tang, "A compressed sensing based AI learning paradigm for crude oil price forecasting," *Energy Economics*, vol. 46, pp. 236-245, 2014.
- [34] H. Chiroma, S. Abdul-Kareem, A. Abubakar, A. M. Zeki, and M. J. Usman, "Orthogonal Wavelet Support Vector Machine for Predicting Crude Oil Prices," in *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, 2014, pp. 193-201.
- [35] W. Xie, L. Yu, S. Xu, and S. Wang, "A New Method for Crude Oil Price Forecasting Based on Support Vector Machines," in *Computational Science—ICCS 2006*, ed: Springer, 2006, pp. 444-451.
- [36] J. Liu, Y. Bai, and B. Li, "A new approach to forecast crude oil price based on fuzzy neural network," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, 2007, pp. 273-277.
- [37] A. Abraham, "Adaptation of fuzzy inference system using neural learning," in *Fuzzy systems engineering*, ed: Springer, 2005, pp. 53-83.
- [38] Petroleum. (2013). *Petroleum Formation*. Available: <http://www.petroleum.co.uk/formation> [Accessed 27/12/2013]
- [39] P. R. I. D. OPEC. (2013). *I Need to Know- an Introduction to the Oil Industry & OPEC*. Available: [http://www.opec.org/opec\\_web/static\\_files\\_project/media/downloads/publications/ChildrenBook2013.pdf](http://www.opec.org/opec_web/static_files_project/media/downloads/publications/ChildrenBook2013.pdf) [Accessed 24/5/2013]
- [40] U. S. E. I. A. EIA. (2013). *Defintions, Sources and Explanatory Notes -Crude Oil Production*. Available: [http://www.eia.gov/dnav/pet/TblDefs/pet\\_crd\\_crpdn\\_tbldef2.asp](http://www.eia.gov/dnav/pet/TblDefs/pet_crd_crpdn_tbldef2.asp) [Accessed 23/11/2014]
- [41] S. Dunn and J. Holloway, "The Pricing of Crude Oil," *RBA Bulletin*, pp. 65-74, 2012.
- [42] B. Fattouh, *An anatomy of the crude oil pricing system*: Oxford Institute for Energy Studies Oxford, England, 2011.
- [43] K. Amadeo. (2014). *Crude Oil Prices Definition*. Available: [http://useconomy.about.com/od/economicindicators/p/Crude\\_Oil.htm](http://useconomy.about.com/od/economicindicators/p/Crude_Oil.htm) [Accessed 11/10/2014]
- [44] B. Fattouh, "The dynamics of crude oil price differentials," *Energy Economics*, vol. 32, pp. 334-342, 2010.
- [45] E. Watch. (2010). *Crude Oil Benchmarks, Oil Markers*. Available: <http://www.economywatch.com/world-industries/oil/crude-oil/benchmarks.html> [Accessed]
- [46] A. Charles and O. Darné, "Volatility persistence in crude oil markets," *Energy Policy*, vol. 65, pp. 729-742, 2014.
- [47] E. Panas and V. Ninni, "Are oil markets chaotic? A non-linear dynamic analysis," *Energy economics*, vol. 22, pp. 549-568, 2000.
- [48] A. Jazeera. (2008). *Interview: Oil prices to rise again*. Available: <http://www.aljazeera.com/focus/2008/09/200898133143509358.html> [Accessed]

- [49] X. Zeng, "Machine Learning Approach for Crude Oil Price Prediction," Doctor of Philosophy Doctoral level ETD - final, Faculty of Engineering and Physical Sciences, School of Computer Science, The University of Manchester Manchester, UK, 2014.
- [50] EIA. *Cushing, OK WTI spot price FOB*. Available: <http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=rwtc&f=a> [Accessed]
- [51] Katchum. (2012). *Crude Oil: Supply Exceeds Demand For The First Time In A Decade*. Available: <http://seekingalpha.com/author/katchum> [Accessed 23/7/2013]
- [52] R. D. Knabb. (2005). *Tropical Cyclone Report Hurricane Katrina*. Available: [http://www.nhc.noaa.gov/data/tcr/AL122005\\_Katrina.pdf](http://www.nhc.noaa.gov/data/tcr/AL122005_Katrina.pdf) [Accessed 20/8/2013]
- [53] R. Gibbons. (2011). *Japan's Earthquake Leads To Falling Oil Prices On Fears Of Decreased Demand*. Available: [http://www.huffingtonpost.com/2011/03/11/oil-prices-fall-in-early-n\\_834404.html](http://www.huffingtonpost.com/2011/03/11/oil-prices-fall-in-early-n_834404.html) [Accessed 11/11/2013]
- [54] K. Amadeo. (2014). *How Are Oil Prices Determined?* Available: [http://useconomy.about.com/od/commoditiesmarketfaq/f/oil\\_prices.htm](http://useconomy.about.com/od/commoditiesmarketfaq/f/oil_prices.htm) [Accessed 7/7/2014]
- [55] J. Han and M. Kamber, *Data Mining, Southeast Asia Edition: Concepts and Techniques*: Morgan kaufmann, 2006.
- [56] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2005.
- [57] M. Venkatadri and L. C. Reddy, "A review on data mining from past to the future," *International Journal of Computer Applications*, vol. 15, pp. 19-22, 2011.
- [58] F. Gorunescu, *Data Mining: Concepts, models and techniques* vol. 12: Springer, 2011.
- [59] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*: MIT press, 2001.
- [60] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, *et al.*, "Data mining techniques for cancer detection using serum proteomic profiling," *Artificial intelligence in medicine*, vol. 32, pp. 71-83, 2004.
- [61] H. Chiroma, S. Abdulkareem, A. ABUBAKAR, and J. U. MOHAMMED, "Computational Intelligence Techniques with Application to Crude Oil Price Projection: A literature Survey from 2001-2012," *Neural Network World*, vol. 23, pp. 523-551, 2013.
- [62] L. A. Gabralla and A. Abraham, "Computational Modeling of Crude Oil Price Forecasting: A Review of Two Decades of Research," *International Journal of Computer Information Systems and Industrial Management Applications*. , vol. 5 pp. 729-740, 2013.
- [63] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, pp. 1-37, 2008.

- [64] L. A. Gabralla, R. Jammazi, and A. Abraham, "Oil price prediction using ensemble machine learning," in *Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on*, 2013, pp. 674-679.
- [65] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, pp. 199-222, 2004.
- [66] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, pp. 37-66, 1991.
- [67] J. G. Cleary and L. E. Trigg, "K<sup>\*</sup>: An Instance-based Learner Using an Entropic Distance Measure," in *ICML*, 1995, pp. 108-114.
- [68] W. B. Wu, M. Woodroffe, and G. Mentz, "Isotonic regression: Another look at the changepoint problem," *Biometrika*, vol. 88, pp. 793-804, 2001.
- [69] F. Pedregosa. (2013). *isotonic regression* Available: <http://fa.bianp.net/blog/2013/isotonic-regression/> [Accessed 2/3/2014]
- [70] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3-42, 2006.
- [71] W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A comparative study of Reduced Error Pruning method in decision tree algorithms," in *Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on*, 2012, pp. 392-397.
- [72] H. Chiroma, S. Abdulkareem, and A. Y. u. Gital, "An Intelligent Model Framework for Handling Effects of Uncertainty Events for Crude Oil Price Projection: Conceptual Paper," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2014.
- [73] Y. Chauvin and D. E. Rumelhart, *Backpropagation: theory, architectures, and applications*: Psychology Press, 1995.
- [74] H. Demuth, M. Beale, and M. Hagan, "Neural network toolbox™ 6," *User's guide*, 2008.
- [75] A. Abraham, "Artificial neural networks," *Handbook of Measuring System Design*, vol. 0-470-02143-8, pp. 901-908, 2005.
- [76] J.-S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 23, pp. 665-685, 1993.
- [77] L. A. Zadeh, "Roles of soft computing and fuzzy logic in the conception, design and deployment of information/intelligent systems," in *Computational intelligence: soft computing and fuzzy-neuro integration with applications*, ed: Springer, 1998, pp. 1-9.
- [78] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *Systems, Man and Cybernetics, IEEE Transactions on*, pp. 116-132, 1985.
- [79] R. C. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proceedings of the sixth international symposium on micro machine and human science*, 1995, pp. 39-43.
- [80] N. Zarvić and R. Wieringa, "An integrated enterprise architecture framework for business-IT alignment," *Designing Enterprise Architecture Frameworks: Integrating Business Processes with IT Infrastructure*, p. 63, 2014.



- [81] L. Urbaczewski and S. Mrdalj, "A comparison of enterprise architecture frameworks," *Issues in Information Systems*, vol. 7, pp. 18-23, 2006.
- [82] R. Sessions, "Comparison of the top four enterprise architecture methodologies," 2007.
- [83] C. L. Thompson, "Scaling the Zachman Framework: A Software Development Methodology for Non-enterprise Applications," Regis University, 2006.
- [84] Z. Mahmood, "Architectural representations for describing enterprise information and data," in *Proceedings 10th WSEAS conference on computers*, 2006, pp. 728-733.
- [85] S. S. Ostadzadeh, J. Habibi, and S. A. Ostadzadeh, "A Framework for Decision Support Systems Based on Zachman Framework," in *Advanced Techniques in Computing Sciences and Software Engineering*, ed: Springer, 2010, pp. 497-502.
- [86] A. Ramadan and M. Hefnawi, "A Network Security Architecture Using The Zachman Framework," in *Managing Critical Infrastructure Risks*, ed: Springer, 2007, pp. 133-143.
- [87] S. S. Ostadzadeh, F. S. Aliee, and S. A. Ostadzadeh, "A method for consistent modeling of Zachman Framework cells," in *Advances and Innovations in Systems, Computing Sciences and Software Engineering*, ed: Springer, 2007, pp. 375-380.
- [88] C. M. Pereira and P. Sousa, "A method to define an Enterprise Architecture using the Zachman Framework," in *Proceedings of the 2004 ACM symposium on Applied computing*, 2004, pp. 1366-1371.
- [89] F. Goethals, W. Lemahieu, M. Snoeck, and J. Vandenbulcke, "An overview of enterprise architecture framework deliverables," *sICFAI University Press*, 2006.
- [90] J. A. Zachman, "A framework for information systems architecture," *IBM systems journal*, vol. 26, pp. 276-292, 1987.
- [91] J. Zachman, "The zachman framework for enterprise architecture," *Zachman International*, 2002.
- [92] Q. Ji, "System analysis approach for the identification of factors driving crude oil prices," *Computers & Industrial Engineering*, vol. 63, pp. 615-625, 2012.
- [93] S. Wang, L. Yu, and K. K. Lai, "A novel hybrid AI system framework for crude oil price forecasting," in *Data Mining and Knowledge Management*, ed: Springer, 2005, pp. 233-242.
- [94] B. Fattouh, L. Kilian, and L. Mahadeva, "The role of speculation in oil markets: What have we learned so far?," 2012.
- [95] A. Bénassy-Quéré, V. Mignon, and A. Penot, "China and the relationship between the oil price and the dollar," *Energy Policy*, vol. 35, pp. 5795-5805, 2007.
- [96] Y. Huang, G. H. Huang, Z. Hu, I. Maqsood, and A. Chakma, "Development of an expert system for tackling the public's perception to climate-change impacts on petroleum industry," *Expert systems with Applications*, vol. 29, pp. 817-829, 2005.
- [97] R. Golombek, A. Irarrazabal, and L. Ma, "OPEC's market power: An empirical dominant firm model for the oil market," 2014.
- [98] C. Yang, M.-J. Hwang, and B.-N. Huang, "An analysis of factors affecting price volatility of the US oil market," *Energy Economics*, vol. 24, pp. 107-119, 2002.

- [99] A.-S. Chen, M. T. Leung, and L.-H. Wang, "Application of polynomial projection ensembles to hedging crude oil commodity risk," *Expert Systems with Applications*, vol. 39, pp. 7864-7873, 2012.
- [100] L. Weiqi, M. Linwei, D. Yaping, and L. Pei, "An econometric modeling approach to short-term crude oil price forecasting," in *Control Conference (CCC), 2011 30th Chinese*, 2011, pp. 1582-1585.
- [101] X. Zhang, K. K. Lai, and S.-Y. Wang, "A new approach for crude oil price analysis based on empirical mode decomposition," *Energy Economics*, vol. 30, pp. 905-918, 2008.
- [102] A. Khashman and N. I. Nwulu, "Support vector machines versus back propagation algorithm for oil price prediction," in *Advances in Neural Networks—ISNN 2011*, ed: Springer, 2011, pp. 530-538.
- [103] H. G. Huntington, "Oil price forecasting in the 1980s: what went wrong?," *The Energy Journal*, pp. 1-22, 1994.
- [104] G. Barone-Adesi, Bourgoin, F., Giannopoulos, K.: , "Don't look back. Risk " *Energy policy*, vol. 14, 1995.
- [105] G. Barone-Adesi and F. Bourgoin, "K. Giannopoulos,(1998)." Don't look back", " *Risk*, vol. 11, pp. 100-103, 1998.
- [106] C. Morana, "A semiparametric approach to short-term oil price forecasting," *Energy Economics*, vol. 23, pp. 325-338, 2001.
- [107] A. Lanza, M. Manera, and M. Giovannini, "Modeling and forecasting cointegrated relationships among heavy oil and product prices," *Energy Economics*, vol. 27, pp. 831-848, 2005.
- [108] C.-l. Zhao and B. Wang, "Forecasting Crude Oil Price with an Autoregressive Integrated Moving Average (ARIMA) Model," in *Fuzzy Information & Engineering and Operations Research & Management*, ed: Springer, 2014, pp. 275-286.
- [109] H. Xie, M. ZHOU, Y. HU, and M. YU, "Forecasting the Crude Oil Price with Extreme Values," *Journal of Systems Science and Information*, vol. 2, pp. 193-205, 2014.
- [110] E. M. Azoff, *Neural network time series forecasting of financial markets*, 1st ed.: John Wiley & Sons, Inc., 1994.
- [111] I. Haidar, S. Kulkarni, and H. Pan, "Forecasting model for crude oil prices based on artificial neural networks," in *Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008. International Conference on*, 2008, pp. 103-108.
- [112] A. Alizadeh and K. Mafinezhad, "Monthly Brent oil price forecasting using artificial neural networks and a crisis index," in *Electronics and Information Engineering (ICEIE), 2010 International Conference On*, 2010, pp. V2-465-V2-468.
- [113] T. Mingming and Z. Jinliang, "A multiple adaptive wavelet recurrent neural network model to analyze crude oil prices," *Journal of Economics and Business*, vol. 64, pp. 275-286, 2012.
- [114] A. A. Godarzi, R. M. Amiri, A. Talaei, and T. Jamasb, "Predicting oil price movements: A dynamic Artificial Neural Network approach," *Energy Policy*, vol. 68, pp. 371-382, 2014.

- [115] A. Khashman and N. I. Nwulu, "Intelligent prediction of crude oil price using Support Vector Machines," in *Applied Machine Intelligence and Informatics (SAMI), 2011 IEEE 9th International Symposium on*, 2011, pp. 165-169.
- [116] H. Chiroma, S. Abdulkareem, A. I. Abubakar, and T. Herawan, "Kernel Functions for the Support Vector Machine: Comparing Performances on Crude Oil Price Data," in *Recent Advances on Soft Computing and Data Mining*, ed: Springer, 2014, pp. 273-281.
- [117] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, and S. Yang Bang, "Constructing support vector machine ensemble," *Pattern recognition*, vol. 36, pp. 2757-2767, 2003.
- [118] M. Kaboudan, "Compumetric forecasting of crude oil prices," in *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, 2001, pp. 283-287.
- [119] S. C. Huang and L. F. Chang, "Oil Price Forecasting with Hierarchical Multiple Kernel Machines," in *Computer, Consumer and Control (IS3C), 2014 International Symposium on*, 2014, pp. 260-263.
- [120] H. S. Hasanabadi, S. Khan, and R. V. Mayorga, "Forecasting Return Volatility of Crude Oil Future Prices Using Artificial Neural Networks ; Based on Intra Markets Variables and Focus on the Speculation Activity " 2014.
- [121] M. Panella, F. Barcellona, and R. L. D'Ecclesia, "Forecasting energy commodity prices using neural networks," *Advances in Decision Sciences*, vol. 2012, 2012.
- [122] A. Ghaffari and S. Zare, "A novel algorithm for prediction of crude oil price variation based on soft computing," *Energy Economics*, vol. 31, pp. 531-536, 2009.
- [123] L. Yu, S. Wang, and K. K. Lai, "A generalized Intelligent-agent-based fuzzy group forecasting model for oil price prediction," in *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, 2008, pp. 489-493.
- [124] L. A. Gabralla, H. Mahersia, and A. Abraham, "Ensemble Neurocomputing Based Oil Price Prediction," in *Afro-European Conference for Industrial Advancement*, 2015, pp. 293-302.
- [125] X. Zhang and Z. Qiu, "Crude Oil Price Forecasting Based on the ARIMA and BP Neural Network Combinatorial Algorithm," *Bridges*, vol. 10, p. 9780784412602.0075, 2014.
- [126] D. Xu, Y. Zhang, C. Cheng, W. Xu, and L. Zhang, "A Neural Network-Based Ensemble Prediction Using PMRS and ECM," in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, 2014, pp. 1335-1343.
- [127] A. Shabri and R. Samsudin, "Crude Oil Price Forecasting Based on Hybridizing Wavelet Multiple Linear Regression Model, Particle Swarm Optimization Techniques, and Principal Component Analysis," *The Scientific World Journal*, vol. 2014, 2014.
- [128] X. Li, K. He, K. K. Lai, and Y. Zou, "Forecasting Crude Oil Price with Multiscale Denoising Ensemble Model," *Mathematical Problems in Engineering*, vol. 2014, 2014.
- [129] L. K. Griffin, "Analysis and comparison of DoDAF and ZACHMAN framework for use as the architecture for the United States Coast Guard's maritime patrol (WPC)," Monterey, California. Naval Postgraduate School, 2005.

- [130] A. T. Bahill, R. Botta, and J. Daniels, "The Zachman framework populated with baseball models," *Journal of Enterprise Architecture*, vol. 2, pp. 50-68, 2006.
- [131] A. Abdullah, "Applying the Zachman Framework data dimension to determine content of a digital library," in *Building an Information Society for All, Proceedings of the International Conference on Libraries, Information & Society ICoLIS2007*, 2007, pp. 26-27.
- [132] Q.-A. Fazil, Z. Abdullah, and S. Noah, "Applying Zachman Framework to determine the content of semantic theses digital library," in *Information Technology (ITSim), 2010 International Symposium in*, 2010, pp. 1596-1600.
- [133] G. Piho, J. Tepandi, and M. Roost, "Domain analysis with archetype patterns based Zachman Framework for enterprise architecture," in *Information Technology (ITSim), 2010 International Symposium in*, 2010, pp. 1351-1356.
- [134] M. M. Gaber, *Scientific data mining and knowledge discovery: Principles and foundations*: Springer, 2009.
- [135] P. Brazdil, C. G. Carrier, C. Soares, and R. Vilalta, *Metalearning: applications to data mining*: Springer, 2008.
- [136] H. S. Own and A. Abraham, "A Novel-weighted Rough Set-based Meta Learning for Ozone Day Prediction," *Acta Polytechnica Hungarica*, vol. 11, 2014.
- [137] A. Verikas, Z. Kalsyte, M. Bacauskiene, and A. Gelzinis, "Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey," *Soft Computing*, vol. 14, pp. 995-1010, 2010.
- [138] F. A. Anifowose, "Advances in Hybrid Computational Intelligence Application in Oil and Gas Reservoir Characterization," in *SPE Saudi Arabia section Young Professionals Technical Symposium*, 2012.
- [139] C. Grosan and A. Abraham, *Intelligent systems: A modern approach* vol. 17: Springer, 2011.
- [140] I. Partalas, G. Tsoumakas, I. Katakis, and I. Vlahavas, "Ensemble pruning using reinforcement learning," in *Advances in Artificial Intelligence*, ed: Springer, 2006, pp. 301-310.
- [141] R. Vilalta, C. G. Giraud-Carrier, P. Brazdil, and C. Soares, "Using Meta-Learning to Support Data Mining," *IJCSA*, vol. 1, pp. 31-45, 2004.
- [142] R. Maclin and D. Opitz, "Popular ensemble methods: An empirical study," *arXiv preprint arXiv:1106.0257*, 2011.
- [143] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?," *Machine learning*, vol. 54, pp. 255-273, 2004.
- [144] M. Blachnik, "Ensembles of instance selection methods based on feature subset," *Procedia Computer Science*, vol. 35, pp. 388-396, 2014.
- [145] E. Menahem, L. Rokach, and Y. Elovici, "Combining one-class classifiers via meta learning," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 2013, pp. 2435-2440.
- [146] M. M. Inc. (2012). *Growing gold and silver producer in the Americas*. Available: <http://www.mcewenmining.com/> [Accessed 15/5/2013]
- [147] EIA. (2012). *PETROLEUM & OTHER LIQUIDS*. Available: <http://www.eia.gov/petroleum/data.cfm> [Accessed 23/4/2013]

- [148] A. Alexandridis and E. Livanis, "Forecasting crude oil prices using wavelet neural networks," *Published in the proc. of 5th FSDET, Athens, Greece* vol. 8, 2008.
- [149] investopedia. (2012). *Federal Funds Rate*. Available: <http://www.investopedia.com/terms/f/federalfundrate.asp> [Accessed 7/8/2014]
- [150] I. El-Sharif, D. Brown, B. Burton, B. Nixon, and A. Russell, "Evidence on the nature and extent of the relationship between oil prices and equity values in the UK," *Energy Economics*, vol. 27, pp. 819-830, 2005.
- [151] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence*, vol. 17, pp. 375-381, 2003.
- [152] H. Liu and R. Setiono, "A probabilistic approach to feature selection-a filter solution," in *ICML*, 1996, pp. 319-327.
- [153] M. Robnik-Šikonja and I. Kononenko, "An adaptation of Relief for attribute estimation in regression," in *Machine Learning: Proceedings of the Fourteenth International Conference (ICML '97)*, 1997, pp. 296-304.
- [154] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.
- [155] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*: John Wiley & Sons, 2004.
- [156] H. Blokeel and J. Struyf, "Efficient algorithms for decision tree cross-validation," *The Journal of Machine Learning Research*, vol. 3, pp. 621-650, 2003.
- [157] K. K. Lai, K. He, and J. Yen, "Modeling VaR in crude oil market: a multi scale nonlinear ensemble approach incorporating wavelet analysis and ANN," in *Computational Science-ICCS 2007*, ed: Springer, 2007, pp. 554-561.
- [158] O. Kaynar, I. Yilmaz, and F. Demirkoparan, "Forecasting of natural gas consumption with neural network and neuro fuzzy system," *Energy Education Science and Technology Part A-Energy Science and Research*, vol. 26, pp. 221-238, 2011.
- [159] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, pp. 10-18, 2009.
- [160] S. S. Aksenova. (2004). *Machine Learning with WEKA WEKA Explorer Tutorial for WEKA Version 3.4.3* Available: <http://csed.sggs.ac.in/csedsites/default/files/WEKA%20Explorer%20Tutorial.pdf> [Accessed 9/2/2014]
- [161] A. Jović, K. Brkić, and N. Bogunović, "An overview of free software tools for general data mining," in *37th International Convention MIPRO 2014*, 2014.
- [162] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, *et al.*, "Weka-a machine learning workbench for data mining," in *Data Mining and Knowledge Discovery Handbook*, ed: Springer, 2010, pp. 1269-1277.
- [163] D. Houcque, "Introduction to MATLAB for Engineering Students," *Northwestern University*, 2005.
- [164] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123-140, 1996.

- [165] E. Rashedi and A. Mirzaei, "A hierarchical clusterer ensemble method based on boosting theory," *Knowledge-Based Systems*, vol. 45, pp. 83-93, 2013.
- [166] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, pp. 296-308, 2010.
- [167] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine learning*, vol. 36, pp. 105-139, 1999.
- [168] T. K. Ho, "The random subspace method for constructing decision forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, pp. 832-844, 1998.
- [169] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 18.
- [170] C. W. Wang, "New ensemble machine learning method for classification and prediction on gene expression data," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, 2006, pp. 3478-3481.
- [171] M. Panda and M. R. Patra, "Ensemble voting system for anomaly based network intrusion detection," *International journal of recent trends in engineering*, vol. 2, pp. 8-13, 2009.
- [172] H.-B. Shen and K.-C. Chou, "Ensemble classifier for protein fold pattern recognition," *Bioinformatics*, vol. 22, pp. 1717-1722, 2006.
- [173] Z. Zhao and Y. Zhang, "Design of ensemble neural network using entropy theory," *Advances in Engineering Software*, vol. 42, pp. 838-845, 2011.
- [174] D. West, P. Mangiameli, R. Rampal, and V. West, "Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application," *European Journal of Operational Research*, vol. 162, pp. 532-551, 2005.
- [175] I. Maqsood, M. R. Khan, and A. Abraham, "An ensemble of neural networks for weather forecasting," *Neural Computing & Applications*, vol. 13, pp. 112-122, 2004.
- [176] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, p. 79, 2005.
- [177] R. A. Ahmed and A. B. Shabri, "A Hybrid of EMD-SVM Based on Extreme Learning Machine for Crude Oil Price Forecasting," *Australian Journal of Basic and Applied Sciences*, vol. 8(15), pp. 341-351, 2014.
- [178] M. Flower. (2010). *Sudan conflict and oil exploration* Available: <http://www.oil-price.net/en/articles/sudan-conflict-and-oil-exploration.php> [Accessed 31/7/2013]
- [179] C. Giraud-Carrier, "Beyond predictive accuracy: what?," in *Proceedings of the ECML-98 Workshop on Upgrading Learning to Meta-Level: Model Selection and Data Transformation*, 1998, pp. 78-85.
- [180] L. A. Gabralla and A. Abraham, "Prediction of Oil Prices Using Bagging and Random Subspace," in *Proceedings of the Fifth International Conference on*

- Innovations in Bio-Inspired Computing and Applications IBICA 2014*, 2014, pp. 343-354.
- [181] L. A. Gabralla and A. Abraham, "Hybrid soft computing methods for prediction of oil prices," in *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, 2014 b, pp. 140-144.
- [182] L. A. Gabralla, T. M. Wahby, V. K. Ojha, and A. Abraham, "Ensemble of Adaptive Neuro-Fuzzy Inference System Using Particle Swarm Optimization for Prediction of Crude Oil Prices," presented at the International Conference on Hybrid Intelligent Systems (HIS), Kuwait, 2015 b.
- [183] H. Chiroma, S. Abdulkareem, and T. Herawan, "Evolutionary Neural Network model for West Texas Intermediate crude oil price prediction," *Applied Energy*, vol. 142, pp. 266-273, 2015.
- [184] H. Chiroma, S. Abdulkareem, A. Abubakar, A. Zeki, A. Gital, and M. Usman, "Co—Active neuro-fuzzy inference systems model for predicting crude oil price based on OECD inventories," in *Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on*, 2013, pp. 232-235.
- [185] EIA. (2014). *How much does it cost to produce crude oil and natural gas?* Available: <http://www.eia.gov/tools/faqs/faq.cfm?id=367&t=6> [Accessed 23/10/2014]
- [186] G. T. Steven Levine , Daniel Arthur ,Michael Tolleth. (2014). *Understanding Crude Oil and Product Markets*. Available: <http://www.api.org/oil-and-natural-gas-overview/~-/media/Files/Oil-and-Natural-Gas/Crude-Oil-Product-Markets/Crude-Oil-Primer/Understanding-Crude-Oil-and-Product-Markets-Primer-Low.pdf> [Accessed 12/11/2014]
- [187] P. R. Robinson, "Petroleum processing overview," in *Practical Advances in Petroleum Processing*, ed: Springer, 2006, pp. 1-78.
- [188] R. McNamara. (2012). *Edwin Drake Drilled the First Oil Well in Pennsylvania in 1859*. Available: <http://history1800s.about.com/od/oil/a/first-oil-well.htm> [Accessed 19/8/2014]
- [189] K. Amadeo. (2013). *How Are Oil Prices Determined?* Available: [http://useconomy.about.com/od/commoditiesmarketfaq/f/oil\\_prices.htm](http://useconomy.about.com/od/commoditiesmarketfaq/f/oil_prices.htm) [Accessed 14/10/2014]
- [190] P. E. George. (2008). *How the Crude Oil Market Works*. Available: <http://science.howstuffworks.com/environmental/energy/crude-oil-market.htm> [Accessed 15/9/2014]
- [191] J. Zachman. (2006). *Enterprise Architecture and Legacy Systems*. Available: <http://www.ies.aust.com/PDF-papers/zachman1.pdf> [Accessed 30/9/2013]
- [192] ARGUS. (2014). *Methodology and specifications guide*. Available: [http://www.argusmedia.com/~-/media/Files/PDFs/Meth/argus\\_crude.pdf?la=en](http://www.argusmedia.com/~-/media/Files/PDFs/Meth/argus_crude.pdf?la=en) [Accessed 13/1/2015]
- [193] I. IEA, OPEC and IOSCO. (2010). *Oil Price Reporting Agencies*. Available: [http://www.iea.org/media/g20/4\\_2011\\_Oil\\_Price\\_Reporting\\_Agencies.pdf](http://www.iea.org/media/g20/4_2011_Oil_Price_Reporting_Agencies.pdf) [Accessed]
- [194] PLATTS. (2014). *Crude Oil*. Available: <http://www.platts.com/IM.Platts.Content/MethodologyReferences/MethodologySpecs/Crude-oil-methodology.pdf> [Accessed 02/09/2014]



- [195] PLATTS. (2014). *Holiday Schedule*. Available: <http://www.platts.com/holiday> [Accessed 27/11/2014]
- [196] A. Radwan and M. Aarabi, "Study of implementing Zachman framework for modeling information systems for manufacturing enterprises aggregate planning," *Simulation*, vol. 16, p. 18, 2011.
- [197] R. Rezaei and F. Shams, "A methodology to create data architecture in Zachman framework," *World Applied Science Journal*, vol. 3, pp. 343-349, 2008.
- [198] L. Ertaul and R. Sudarsanam, "Security Planning Using Zachman Framework for Enterprises," in *Proceedings of EURO mGOV*, 2005.
- [199] A. Fatolahi and F. Shams, "An investigation into applying UML to the Zachman framework," *Information Systems Frontiers*, vol. 8, pp. 133-143, 2006.
- [200] D. C. Hay, "The Zachman Framework: an Introduction. Essential Strategies," *Inc. The data Administration Newsletter*, 1997.
- [201] A. Källgren, "Towards a Framework for Enterprise Architecture at Vattenfall," 2008.
- [202] C. O'Rourke, N. Fishman, and W. Selkow, *Enterprise architecture using the Zachman framework*: Course Technology Ptr, 2003.
- [203] L. Kilian, "Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market," 2006.
- [204] C. Baumeister, "The Art and Science of Forecasting the Real Price of Oil," *Bank of Canada Review*, vol. 2014, pp. 21-31, 2014.
- [205] C. Baumeister, L. Kilian, and X. Zhou, "Are product spreads useful for forecasting? An empirical evaluation of the Verleger hypothesis," *Bank of Canada* 2013.
- [206] J. D. Hamilton, "Historical oil shocks," National Bureau of Economic Research 2011.
- [207] H. Izakian, "Cluster-Centric Anomaly Detection and Characterization in Spatial Time Series," Doctor of Philosophy, Electrical and Computer Engineering, University of Alberta, 2014.



# APPENDICES

## Appendix A: Attribute selection methods and their features using WEKA

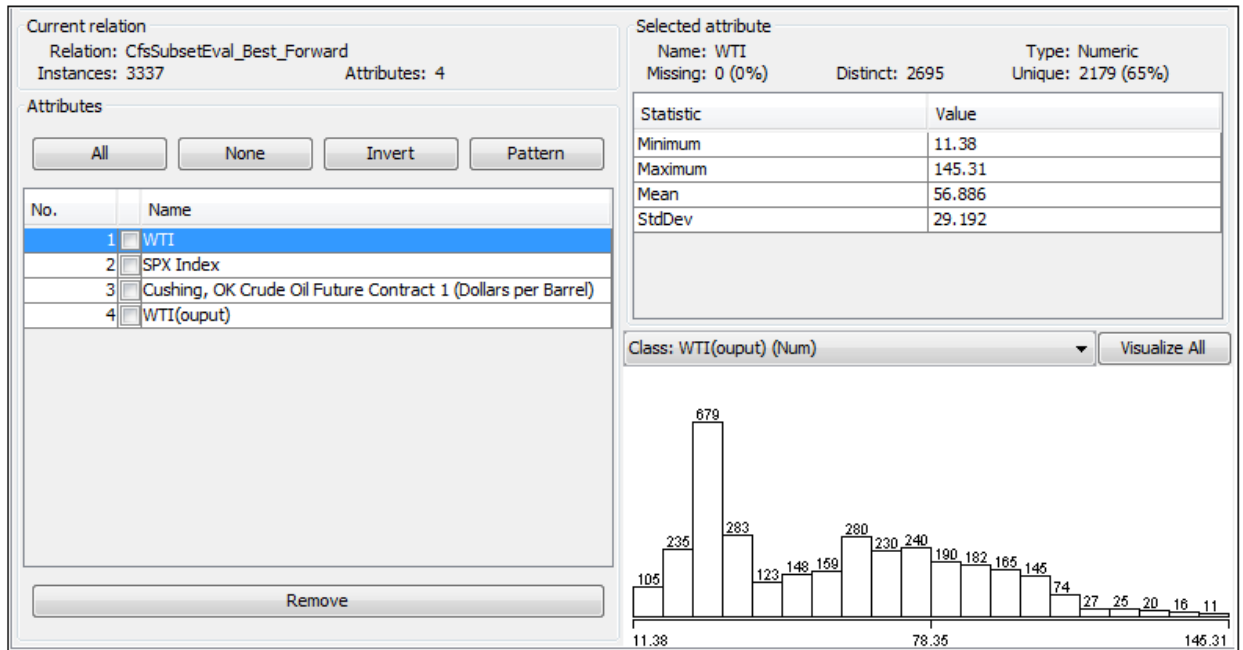


Figure A.1 Represent SBDS<sub>1</sub> using WEKA

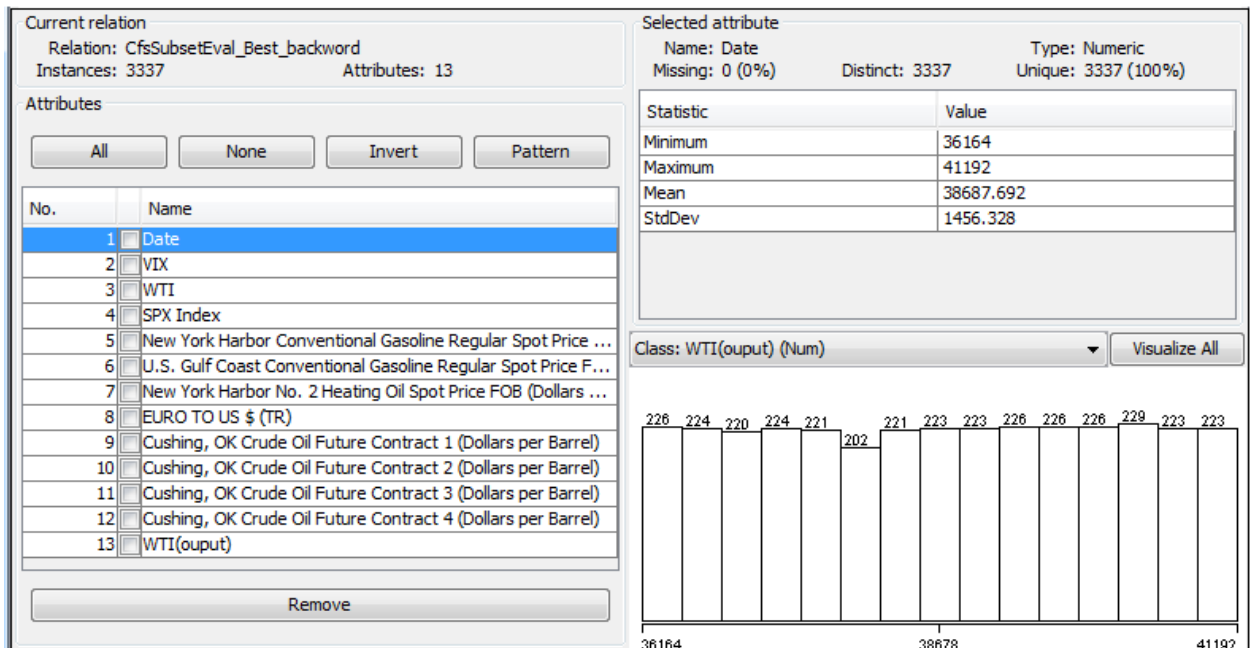


Figure A.2 Represent SBDS<sub>2</sub> using WEKA

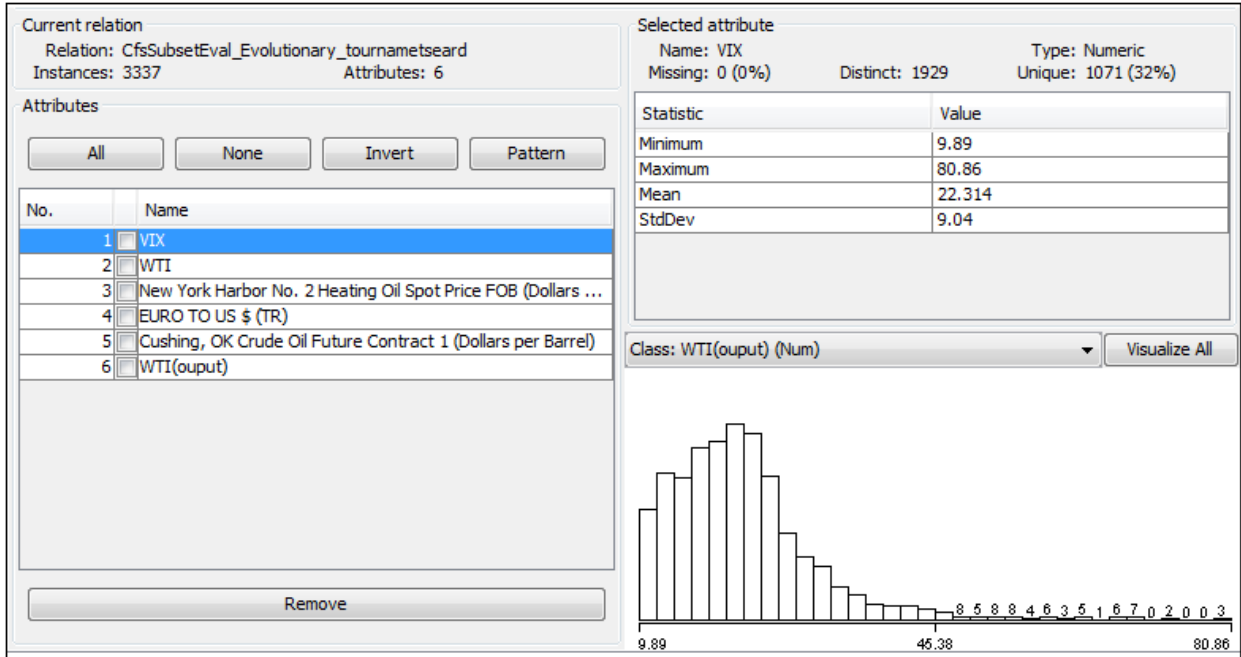


Figure A.3 Represent SBDS<sub>3</sub> using WEKA

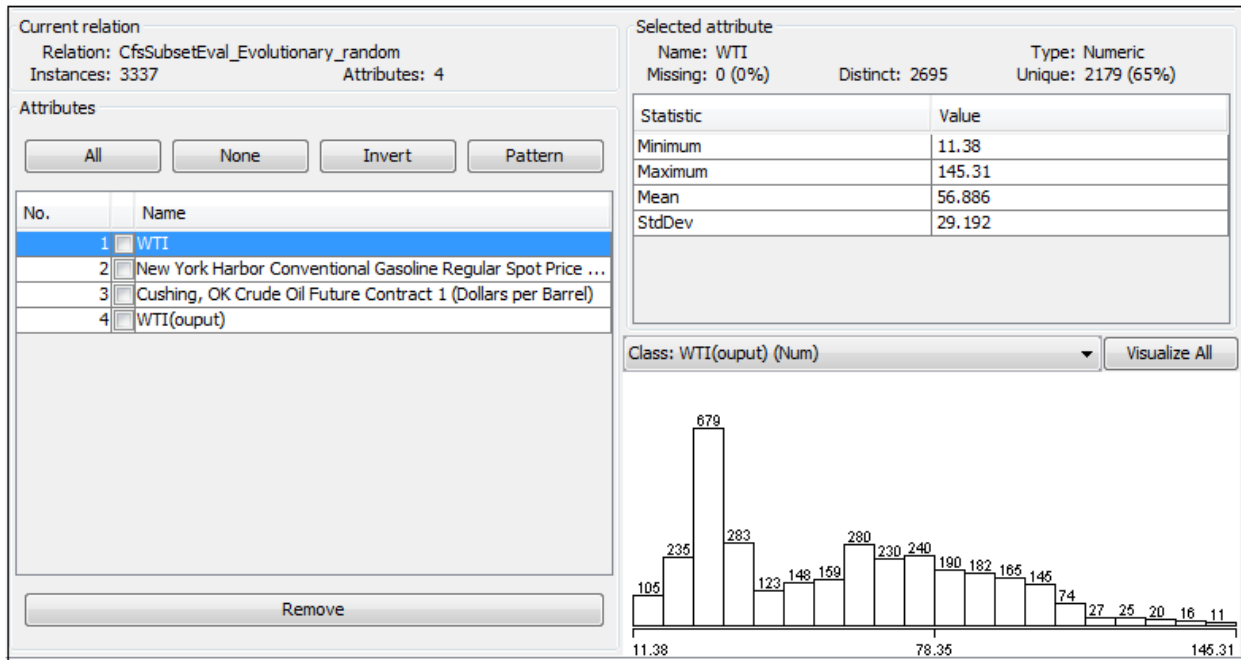


Figure A.4 Represent SBDS<sub>4</sub> using WEKA

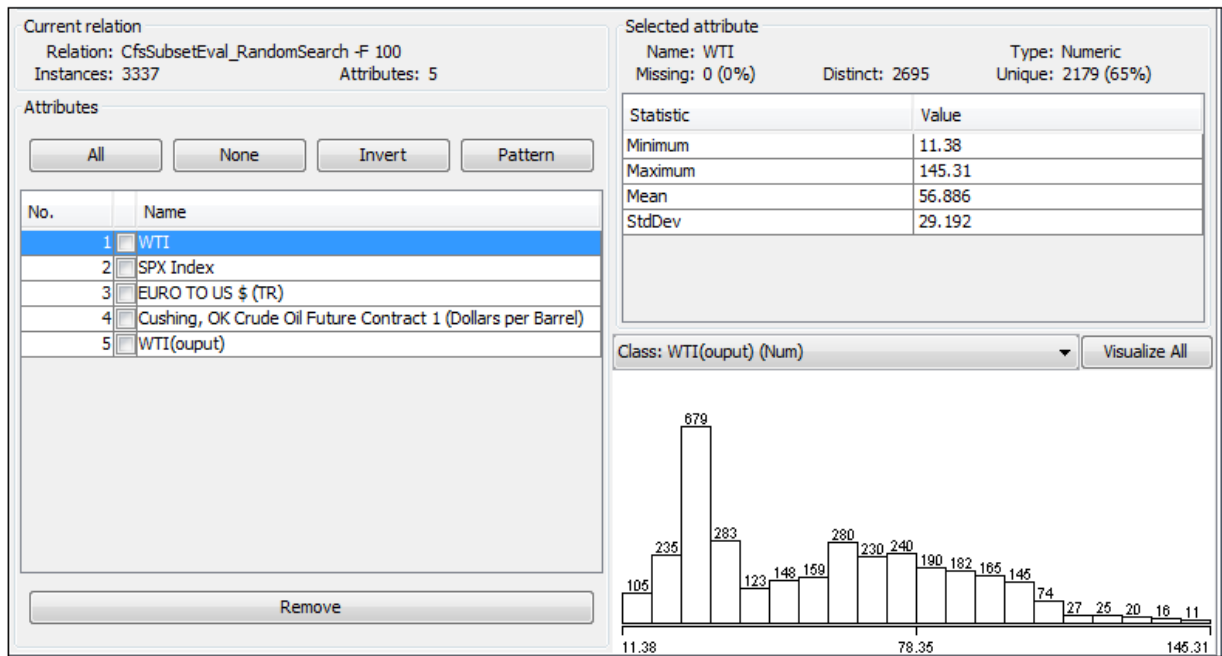


Figure A.5 Represent SBDS<sub>5</sub> using WEKA

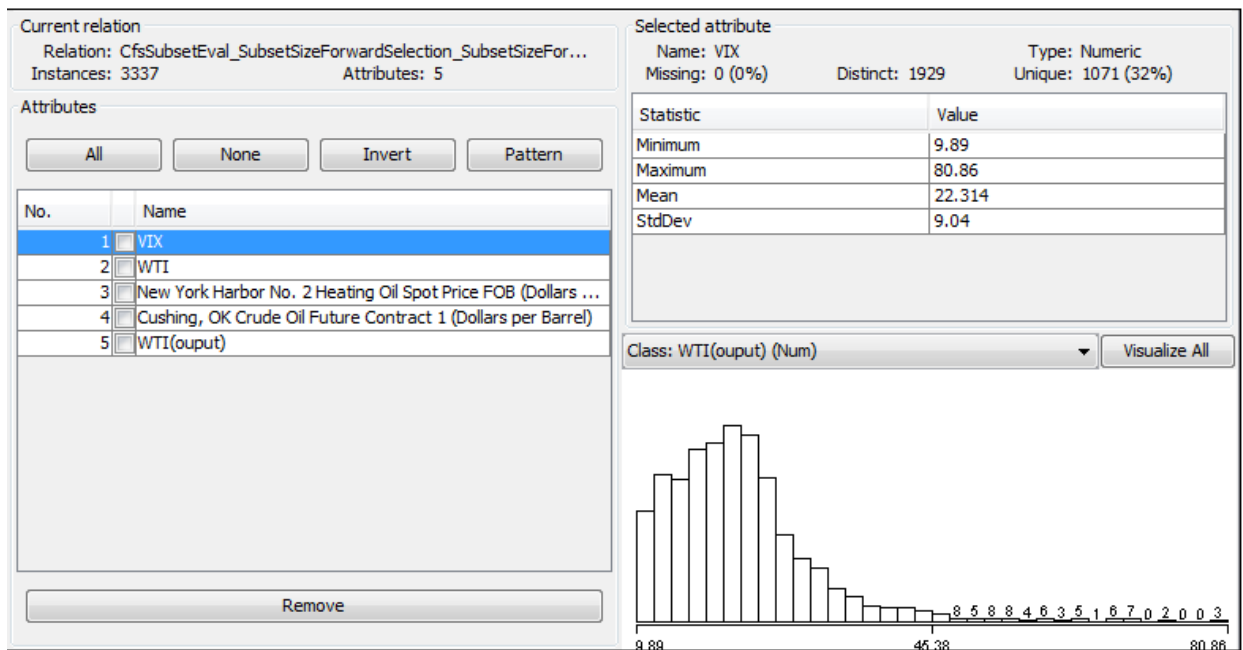


Figure A.6 Represent SBDS<sub>6</sub> using WEKA

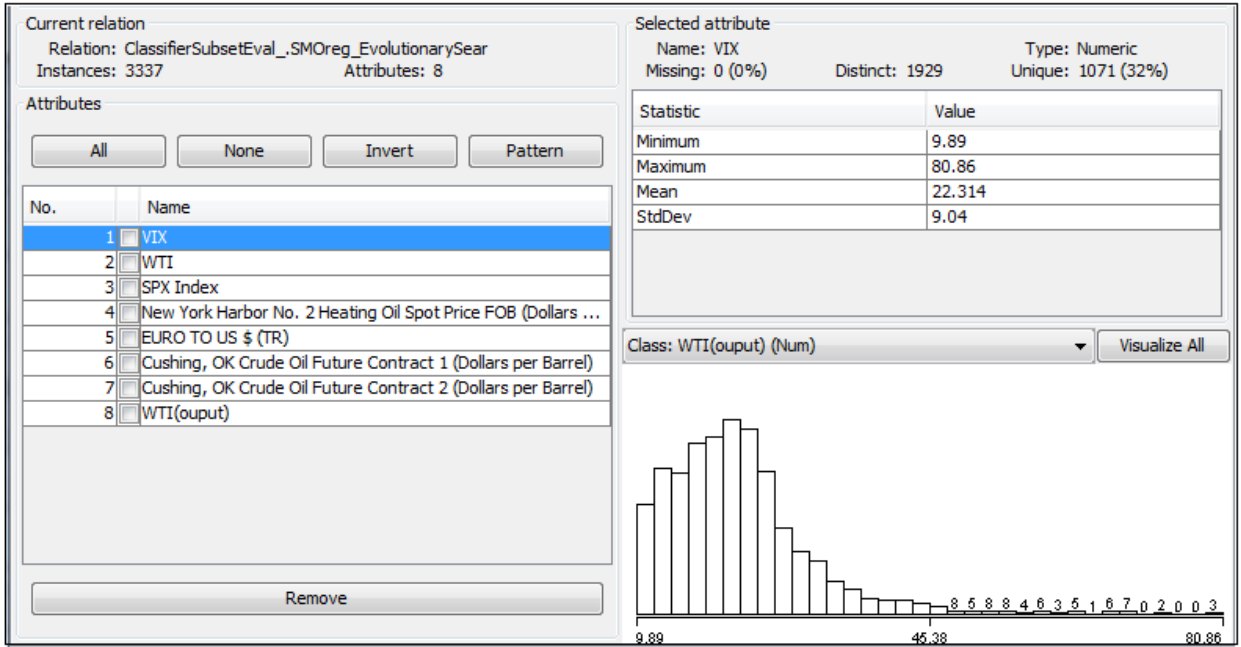


Figure A.7 Represent SBDS<sub>7</sub> using WEKA

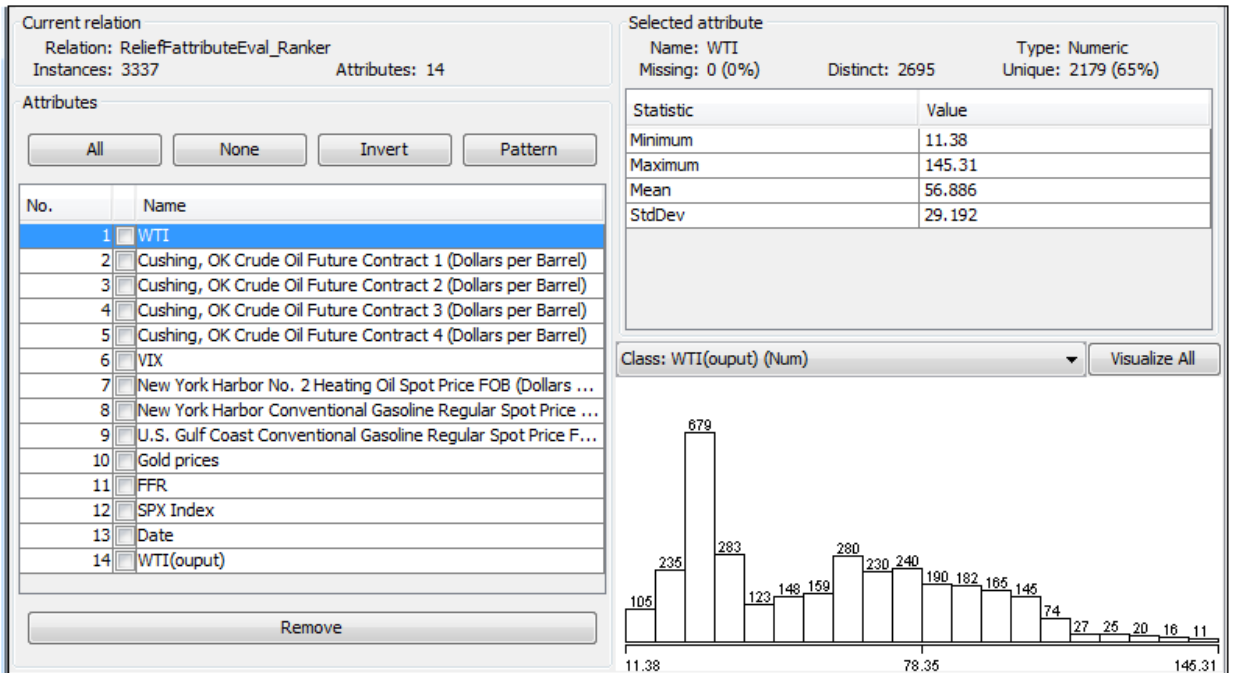
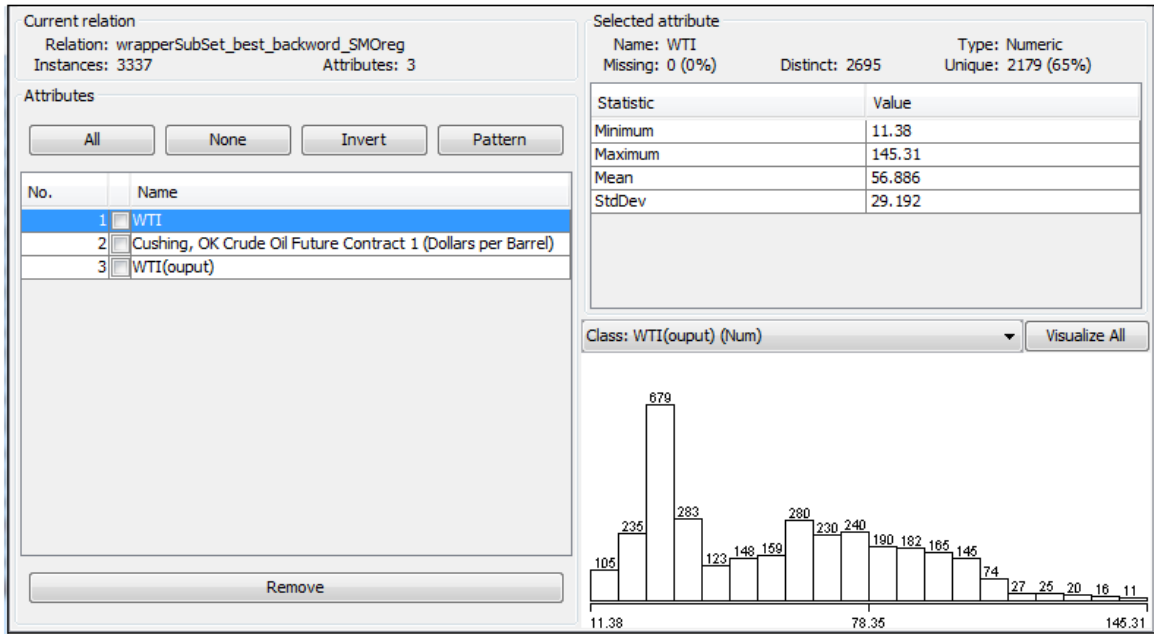


Figure A.8 Represent SBDS<sub>8</sub> using WEKA



**Figure A.9** Represent SBDS<sub>9</sub> using WEKA



**Figure A.10** Represent SBDS<sub>8</sub> using WEKA

## Appendix B: Samples of computations made for performance evaluation

This section lists the samples of computations made for performance evaluation using three type of neural networks approach model with one sub-dataset. The calculations include the computation of MAE, MSE, and RMSE.

**Table B.1** Feed forward performance for SBDS<sub>2</sub>

Training algorithm	Subdataset <sub>2</sub>					
	Neurons	40	45	50	55	60
; feedforward (newff) ; 90%- 10%;						
(Levenberg-Marquardt)	MAE	1.43100E-03	1.13120E-02	4.50600E-03	9.17000E-04	3.29500E-03
	MSE	2.00000E-04	1.50000E-03	6.00000E-04	1.00000E-04	4.00000E-04
	RMSE	1.41421E-02	3.87298E-02	2.44949E-02	1.00000E-02	2.00000E-02
; feedforward (newff) ; 90%- 10%;						
(Bayesian Regularization)	MAE	4.23300E-03	3.66300E-03	3.33700E-03	3.12300E-03	2.83700E-03
	MSE	5.68100E-01	4.91600E-01	4.47900E-01	4.19100E-01	3.80700E-01
	RMSE	7.53724E-01	7.01142E-01	6.69253E-01	6.47379E-01	6.17009E-01
; feedforward (newff) ; 90%- 10%;						
( BFGS quasi-Newton )	MAE	1.37000E-02	9.90000E-03	2.30000E-02	3.43000E-02	2.47000E-02
	MSE	9.00000E-04	2.10000E-03	1.50000E-03	1.10000E-03	1.90000E-03
	RMSE	3.00000E-02	4.58258E-02	3.87298E-02	3.31662E-02	4.35890E-02
; feedforward (newff) ; 80%- 20%;						
(Levenberg-Marquardt)	MAE	4.48000E-04	8.20900E-03	5.21000E-04	2.24500E-03	1.24010E-02
	MSE	0.00000E+00	8.00000E-04	1.00000E-04	2.00000E-04	1.20000E-03
	RMSE	0.00000E+00	2.82843E-02	1.00000E-02	1.41421E-02	3.46410E-02
; feedforward (newff) ; 80%- 20%;						
(Bayesian Regularization)	MAE	3.50800E-03	3.07100E-03	2.86100E-03	2.53200E-03	2.18100E-03
	MSE	3.50300E-04	3.05700E-04	2.84200E-04	2.52300E-04	2.17100E-04
	RMSE	1.87163E-02	1.74843E-02	1.68582E-02	1.58840E-02	1.47343E-02
; feedforward (newff) ; 80%- 20%;						
( BFGS quasi-Newton )	MAE	4.00000E-04	6.00000E-04	1.00000E-03	5.00000E-04	1.40000E-03
	MSE	8.60000E-03	1.47000E-02	8.20000E-03	1.83000E-02	1.18000E-02
	RMSE	9.27362E-02	1.21244E-01	9.05539E-02	1.35277E-01	1.08628E-01
; feedforward (newff) ; 70%- 30%;						
(Levenberg-Marquardt)	Neurons	40	45	50	55	60
	MAE	2.74700E-03	1.92690E-02	2.37750E-02	2.85300E-03	1.40840E-02
	MSE	2.00000E-04	1.10000E-03	1.40000E-03	2.00000E-04	8.00000E-04

	<b>RMSE</b>	1.41421E-02	3.31662E-02	3.74166E-02	1.41421E-02	2.82843E-02
; feedforward (newff) ; 70%- 30%;						
<b>(Bayesian Regularization)</b>	<b>Neurons</b>	<b>40</b>	<b>45</b>	<b>50</b>	<b>55</b>	<b>60</b>
	<b>MAE</b>	6.68300E-03	1.02200E-03	5.56600E-03	5.15100E-03	4.83500E-03
	<b>MSE</b>	3.80600E-04	5.82000E-05	3.16900E-04	2.93300E-04	2.75300E-04
	<b>RMSE</b>	1.95090E-02	7.62889E-03	1.78017E-02	1.71260E-02	1.65922E-02
; feedforward (newff) ; 70%- 30%;						
<b>( BFGS quasi-Newton )</b>	<b>Neurons</b>	<b>40</b>	<b>45</b>	<b>50</b>	<b>55</b>	<b>60</b>
	<b>MAE</b>	3.40000E-03	2.00000E-03	1.90000E-03	3.70000E-03	1.40000E-03
; feedforward (newff) ; 60%- 40%;						
<b>(Levenberg-Marquardt)</b>	<b>Neurons</b>	<b>40</b>	<b>45</b>	<b>50</b>	<b>55</b>	<b>60</b>
	<b>MAE</b>	5.33000E-03	1.91690E-01	1.19450E-02	3.15700E-03	6.81400E-03
	<b>MSE</b>	2.75100E-04	9.53000E-04	5.93900E-04	1.56900E-04	3.38800E-04
	<b>RMSE</b>	1.65861E-02	3.08707E-02	2.43701E-02	1.25260E-02	1.84065E-02
; feedforward (newff) ; 60%- 40%;						
<b>(Bayesian Regularization)</b>	<b>Neurons</b>	<b>40</b>	<b>45</b>	<b>50</b>	<b>55</b>	<b>60</b>
	<b>MAE</b>	8.40000E-04	1.20930E-02	9.92900E-03	8.67900E-03	7.69000E-04
	<b>MSE</b>	4.18000E-05	6.01200E-04	4.93700E-04	4.31500E-04	3.82000E-05
	<b>RMSE</b>	6.46529E-03	2.45194E-02	2.22194E-02	2.07726E-02	6.18061E-03
; feedforward (newff) ; 60%- 40%;						
<b>( BFGS quasi-Newton )</b>	<b>Neurons</b>	<b>40</b>	<b>45</b>	<b>50</b>	<b>55</b>	<b>60</b>
	<b>MAE</b>	1.20000E-03	1.60000E-03	2.50000E-03	1.00000E-03	1.30000E-03

**Table B.2** Radial basis function performance for SBDS<sub>1</sub>

Subdataset <sub>1</sub>					
Neurons	40	45	50	55	60
; Radial basis ; 90%- 10%;					
<b>MAE</b>	1.03146E-04	9.34926E-05	1.00340E-04	5.70070E-05	5.08729E-05
<b>MSE</b>	1.38420E-05	1.26140E-05	1.34620E-05	7.65030E-06	6.82710E-06
<b>RMSE</b>	3.72048E-03	3.55162E-03	3.66906E-03	2.76592E-03	2.61287E-03
; Radial basis ; 80%- 20%;					
<b>MAE</b>	1.53000E-04	1.48000E-04	1.09000E-04	2.58000E-04	2.40010E-05
<b>MSE</b>	1.47180E-05	1.43120E-05	1.05300E-05	1.51780E-05	2.31470E-06
<b>RMSE</b>	3.83640E-03	3.78312E-03	3.24500E-03	3.89590E-03	1.52141E-03
; Radial basis ; 70%- 30%;					
<b>MAE</b>	4.84160E-05	5.01626E-05	5.02828E-05	5.02436E-05	4.98570E-05
<b>MSE</b>	2.81400E-06	2.85650E-06	2.86330E-06	2.86116E-06	2.83900E-05
<b>RMSE</b>	1.67750E-03	1.69012E-03	1.69213E-03	1.69150E-03	5.32823E-03
; Radial basis ; 60%- 40%;					
<b>MAE</b>	3.88153E-05	3.81595E-05	3.81548E-05	3.81543E-05	3.81548E-05
<b>MSE</b>	1.93000E-06	1.89700E-06	1.89700E-06	1.89700E-06	1.89700E-06
<b>RMSE</b>	1.38924E-03	1.37732E-03	1.37732E-03	1.37732E-03	1.37732E-03



**Table B.3** Recurrent performance for SBDS<sub>4</sub>

Recurrent subdataset <sub>4</sub>		
; 90%- 10%;		
(Levenberg-Marquardt)	<b>Hidden layer sizes (default = 10</b>	
	MAE	7.68000E-04
	MSE	1.03020E-04
	RMSE	1.01499E-02
; 90%- 10%;		
(Bayesian regularization)	<b>Hidden layer sizes (default = 10</b>	
	MAE	4.80285E-05
	MSE	6.44540E-06
	RMSE	2.53878E-03
; 90%- 10%;		
( BFGS quasi-Newton )	<b>Hidden layer sizes (default = 10</b>	
	MAE	1.58640E-02
	MSE	2.10000E-03
	RMSE	4.58258E-02
; 80%- 20%;		
(Levenberg-Marquardt)	<b>Hidden layer sizes (default = 10</b>	
	MAE	5.20102E-05
	MSE	5.01600E-06
	RMSE	2.23964E-03
; 80%- 20%;		
(Bayesian Regularization)	<b>Hidden layer sizes (default = 10</b>	
	MAE	3.94799E-05
	MSE	3.80750E-06
	RMSE	1.95128E-03
; 80%- 20%;		
( BFGS quasi-Newton )	<b>Hidden layer sizes (default = 10</b>	
	MAE	7.16800E-03
	MSE	6.91340E-04
	RMSE	2.62933E-02
; 70%- 30%;		
(Levenberg-Marquardt)	<b>Hidden layer sizes (default = 10</b>	
	MAE	4.73000E-04
	MSE	2.69600E-05
	RMSE	5.19230E-03
; 70%- 30%;		

( BFGS quasi-Newton )	<b>Hidden layer sizes (default = 10</b>	
	MAE	2.17658E-04
	MSE	1.23940E-05
	RMSE	3.52051E-03
; 70%- 30%;		
(Bayesian Regularization)	<b>Hidden layer sizes (default = 10</b>	
	MAE	2.00530E-02
	MSE	1.10000E-03
	RMSE	3.31662E-02
; 60%- 40%;		
(Levenberg-Marquardt)	<b>Hidden layer sizes (default = 10</b>	
	MAE	2.08400E-03
	MSE	1.03630E-04
	RMSE	1.01799E-02
; 60%- 40%;		
(Bayesian Regularization)	<b>Hidden layer sizes (default = 10</b>	
	MAE	1.81824E-04
	MSE	9.03990E-06
	RMSE	3.00664E-03
; 60%- 40%;		
( BFGS quasi-Newton )	<b>hidden layer sizes (default = 10</b>	
	MAE	6.17200E-03
	MSE	3.06840E-04
	RMSE	1.75168E-02

**Table B.4** ANFIS and RBF performance for all subdatasets

ANFIS	Subdataset <sub>1</sub>	90-10	1.23522E-05	Subdataset <sub>4</sub>	90-10	3.44406E-04	Subdataset <sub>10</sub>	90-10	7.95274E-06		
		80-20	1.56636E-05		80-20	3.69172E-04		80-20	4.70906E-07		
		70-30	5.81167E-05		70-30	8.19114E-04		70-30	1.77291E-06		
		60-40	4.84699E-05		60-40	1.90370E-04		60-40	2.47023E-02		
	Subdataset <sub>2</sub>	90-10	8.94070E-06	Subdataset <sub>6</sub>	90-10	9.96863E-04					
		80-20	1.32849E-05		80-20	1.50347E-03					
		70-30	2.87358E-05		70-30	7.59600E-03					
		60-40	2.90189E-01		60-40	1.23274E-02					
	Subdataset <sub>3</sub>	90-10	3.75200E-02	Subdataset <sub>9</sub>	90-10	2.39222E-06					
		80-20	2.89157E-02		80-20	6.78925E-06					
		70-30	2.12495E-01		70-30	9.50917E-06					
		60-40	1.12921E+00		60-40	1.90200E-05					
	Radial basis	Subdataset <sub>1</sub>	90-10	5.08729E-05	Subdataset <sub>4</sub>	90-10		2.08946E-04	Subdataset <sub>10</sub>	90-10	2.34000E-04
			80-20	2.40010E-05		80-20		6.08910E-05		80-20	3.98158E-05
			70-30	4.84160E-05		70-30		1.10437E-04		70-30	5.34009E-05
			60-40	3.81543E-05		60-40		3.80000E-04		60-40	1.16000E-04
Subdataset <sub>2</sub>		90-10	1.33255E-03	Subdataset <sub>6</sub>	90-10	1.60000E-04					
		80-20	1.68000E-03		80-20	1.04000E-04					
		70-30	2.85700E-03		70-30	2.34880E-04					
		60-40	4.30200E-03		60-40	8.74530E-05					
Subdataset <sub>3</sub>		90-10	2.75000E-04	Subdataset <sub>9</sub>	90-10	9.04440E-02					
		80-20	2.24000E-04		80-20	2.20646E-05					
		70-30	5.11000E-04		70-30	4.47060E-05					
		60-40	1.51400E-03		60-40	3.35330E-05					

**Table B.5** RCN and FFN performance for all subdatasets

Recurrent	Subdataset <sub>1</sub>	90-10	1.14102E-03	Subdataset <sub>4</sub>	90-10	4.80285E-05	Subdataset <sub>10</sub>	90-10	2.48000E-04				
		80-20	5.79000E-04		80-20	3.94799E-05		80-20	1.47000E-04				
		70-30	6.98300E-03		70-30	2.17658E-04		70-30	1.01200E-03				
		60-40	7.69800E-03		60-40	1.81824E-04		60-40	4.30824E-04				
	Subdataset <sub>2</sub>	90-10	3.77000E-04	Subdataset <sub>6</sub>	90-10	3.94114E-05							
		80-20	1.74000E-04		80-20	1.12000E-04							
		70-30	1.36700E-03		70-30	4.41680E-04							
		60-40	3.32000E-04		60-40	3.62000E-04							
	Subdataset <sub>3</sub>	90-10	5.22000E-04	Subdataset <sub>9</sub>	90-10	1.83283E-04							
		80-20	1.05000E-04		80-20	1.72000E-04							
		70-30	2.82000E-04		70-30	1.82716E-03							
		60-40	1.68000E-04		60-40	9.69000E-04							
Feed-forward	Subdataset <sub>1</sub>	90-10	3.84294E-05	Subdataset <sub>4</sub>	90-10	5.74155E-05					Subdataset <sub>10</sub>	90-10	1.90200E-03
		80-20	1.35000E-04		80-20	6.45000E-05						80-20	1.97000E-04
		70-30	3.55000E-04		70-30	3.04484E-04						70-30	2.25700E-03
		60-40	4.38000E-04		60-40	1.94000E-04						60-40	4.50000E-04
	Subdataset <sub>2</sub>	90-10	9.00000E-04	Subdataset <sub>6</sub>	90-10	9.65000E-04							
		80-20	4.00000E-04		80-20	6.80000E-05							
		70-30	1.02200E-03		70-30	3.73327E-04							
		60-40	7.69000E-04		60-40	2.10900E-03							
	Subdataset <sub>3</sub>	90-10	1.26594E-04	Subdataset <sub>9</sub>	90-10	6.23000E-05							
		80-20	1.24897E-04		80-20	6.04640E-05							
		70-30	3.10000E-03		70-30	3.49000E-04							
		60-40	1.80000E-03		60-40	1.62000E-04							

## Appendix C: Samples of of FFN and ANFIS code

This section lists the samples of FFN and ANFIS code:

### Source code(1)

```
clc
clear, close all
%*****
****
%70-30 training & testing feedforward network
%*****
****
Data_Inputs=          xlsread('CfsSubsetEvalBestbackword70-
30.xlsx','train');
Testing_Data      =    xlsread('CfsSubsetEvalBestbackword70-
30.xlsx','test');

% create network %
Shuffling_Inputs = Data_Inputs(randperm(2336),1:13);
Training_Set     =  Shuffling_Inputs(1:2336,1:12);% specify
training set
Target_Set      = Shuffling_Inputs(1:2336,13); % specify target
set
Testing_Set     = Testing_Data(1:1001,1:12); % specify Testing
set
Testing_Target_Set = Testing_Data(1:1001, 13); % specify
Testing set, Target
[pn,ps] = mapstd(Training_Set');
[tn,ts] = mapstd(Target_Set');

[testn, tests] = mapstd(Testing_Set');
[targettestn,targettests] = mapstd(Testing_Target_Set');

% neurones = [10:10:213];
neurones = [40:5:60];
% neurones = [10:5:70];
for j = 1:length(neurones)
    neurones(j)
% train the network%
```

```

% for i=1:10
rand('seed',2051974);
%rng(2051974);

net = newff(pn,tn,[neurones(j)]); % this command for one
layer
%net = newff(pn,tn,[neurones(j) neurones(j)]); % this com-
mand for two layer
%net = newff(pn,tn,[neurones(j) neurones(j) neurones(j)]);
% this command for three layer
net.divideFcn = '';
net.trainFcn = 'trainlm';

% net.trainparam.min_grad = 0.00000001;
net.trainParam.epochs =1000;
% net.trainParam.lr = 0.05;
% net.trainParam.mu = 0.001;
% net.trainParam.goal =1e-5;
net.performFcn = 'mae';
net = train(net,pn,tn);
%save(net,'net');
%*****
% test the network %
%*****
y = sim(net,pn);
error =(y-tn);
% performance_train(i) = mae(error);
performance_train = mae(error);
ytest = sim(net,testn);
x1_again(:,j) = mapstd('reverse',ytest',targetttests')
errortest =(ytest-targetttestn);
%performance_test(i) = mae(errortest);
performance_test(j) = mae(errortest);
end
% final_result = mean(performance_test)

xlswrite('lubnamsetrainlm.xlsx',x1_again,1)
end

```

## **Source code(2):**

```
clc
clear, close all
%*****
*****
%90-10% training & testing anfis network
%*****
*****
Data_Inputs    =    xlsread('wrapper_best_backword_SMOreg90-
10','train');
Testing_Data   =    xlsread('wrapper_best_backword_SMOreg90-
10','test');

% create network %
Shuffling_Inputs = Data_Inputs(randperm(3003),1:3);
Training_Set     =    Shuffling_Inputs(1:3003,1:2);%    specify
training set
Training_Target_Set = Shuffling_Inputs(1:3003,3); % specify
target set
Testing_Set      =    Testing_Data(1:334,1:2); % specify Testing
set
Testing_Target_Set = Testing_Data(1:334,3); % specify Test-
ing set, Target
% [pn,ps] = mapstd(Training_Set');
% [tn,ts] = mapstd(Target_Set');
% [testn, tests] = mapstd(Testing_Set');
% [targettestn,targettests] = mapstd(Testing_Target_Set');

%FIS structure
trnData = [Training_Set    Training_Target_Set];
testData = [Testing_Set    Testing_Target_Set];

%%Initialize the fuzzy system with command genfis1. Use 5
trapazoid membership functions.
NumMf = [2 2 2]
mfType =('trapmf')
outmfype =('linear')
in_fis = genfis1(trnData,NumMf,mfType,outmfype);

%The initial membership functions produced by genfis1 are
plotted
```

```

figure(1)
subplot(2,2,1)
plotmf(in_fis,'input', 1)
xlabel('input1');ylabel('output');title('initial membership
functions');
subplot(2,2,2)
plotmf(in_fis, 'input', 2)
xlabel('input2');ylabel('output');title('initial membership
functions');

% Apply anfis-command to find the best FIS system - max
number of iterations = 100
epoch = 100;
out_fis = anfis(trnData,in_fis,epoch);
figure (2);
subplot(2,2,1)
plotmf(out_fis,'input', 1);
subplot(2,2,2);
plotmf(out_fis,'input',2);

% Evaluate the output of FIS system using train data
Tr_predict= evalfis(Training_Set,out_fis)
showrule(out_fis)
ruleview(out_fis)
Error_Tr=(Training_Target_Set-Tr_predict);
SquareErrorTrain=Error_Tr.^2;
RMSETr= sqrt(mean(SquareErrorTrain))

%Evaluate the output of FIS system using test data
out_fis2=anfis( testData,in_fis,epoch);
Test_predict=evalfis(Testing_Set,out_fis2);
errorTest= (Testing_Target_Set-Test_predict);
SquareErrorTest=errorTest.^2;
RMSETest= sqrt(mean(SquareErrorTest))
MAE_Test=abs(errorTest);
Final_MAE=mean(MAE_Test)

```

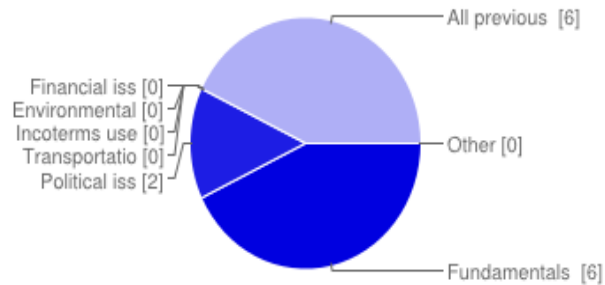


## Appendix D: Questionnaire item and document analysis

Summary of questionnaire for information enterprise architecture for crude oil pricing and prediction

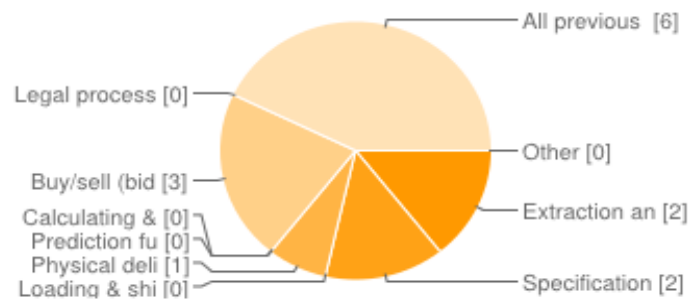
### 1. What is the important information, entities required for prediction/pricing of crude oil?

Fundamentals of supply and demand	6	43%
Political issues (such as Gulf War, Iraq War, Venezuela Unrest,..)	2	14%
Financial issues and uncertainty Financial (such as Global Financial Crises, Asia Financial Crises,...)	0	0%
Environmental issues	0	0%
Incoterms used	0	0%
Transportation and Storage	0	0%
All previous entities	6	43%
Other	0	0%



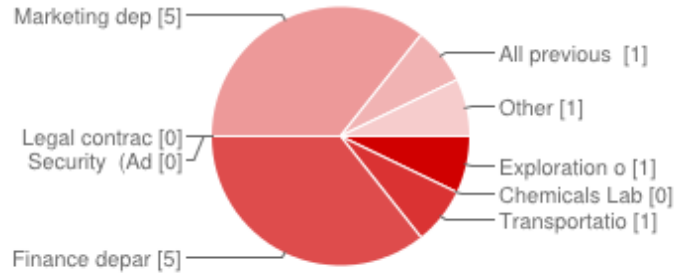
**2. What are the processes that have contribution and an impact on crude oil pricing/prediction?**

Extraction and search	2	14%
Specification & classification crude oil	2	14%
Loading & shipping Cargoes	0	0%
Physical delivery (spot & long term)	1	7%
Calculating & Assessment prices	0	0%
Prediction future prices	0	0%
Buy/sell (bid or offer) and withdrawn	3	21%
Legal process	0	0%
All previous process	6	43%
Other	0	0%



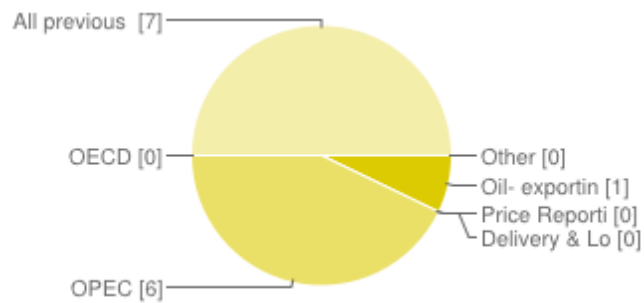
**3. Who are the (people/organizations/departments) that participate in the crude oil price prediction/pricing system?**

Exploration of Crude oil department (engineers, geologists,.....)	1	7%
Chemicals Laboratories (engineers, laboratory analysts,.....)	0	0%
Transportation and Shipping (traders, workers,.....)	1	7%
Finance department (investors, traders, brokers , buyers , consumers, suppliers, speculators,.....)	5	36%
Legal contracts (legal,.....)	0	0%
Security (Administrators,.....)	0	0%
Marketing department (Analysts,.....)	5	36%
All previous participants	1	7%
Other	1	7%



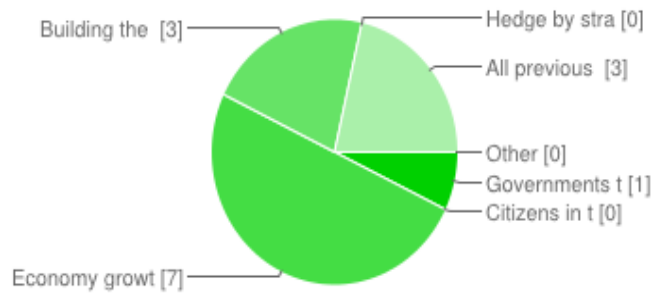
**4. What are the individual locations and places that have a relationship with crude oil prediction/pricing system?**

Oil- exporting & importing countries	1	7%
Price Reporting Agencies & International exchanges	0	0%
Delivery & Loading ports	0	0%
OPEC	6	43%
OECD	0	0%
All previous participants	7	50%
Other	0	0%



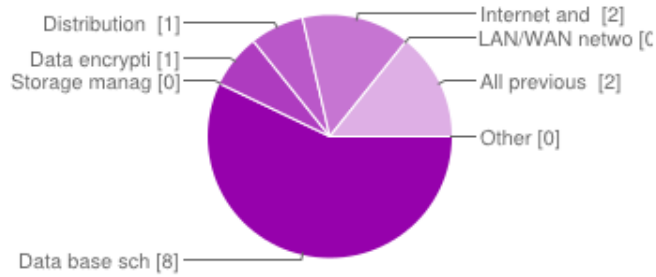
**5. Why are the prediction/pricing crude oil price important to the government, citizen and economy etc. (objectives)?**

Governments to secure the supply of oil to their nations.	1	7%
Citizens in transportation–manufacturing-agriculture-heating –lightening	0	0%
Economy growth depends on energy	7	50%
Building the government’s budget	3	21%
Hedge by strategic oil stocks	0	0%
All previous reasons	3	21%
Other	0	0%



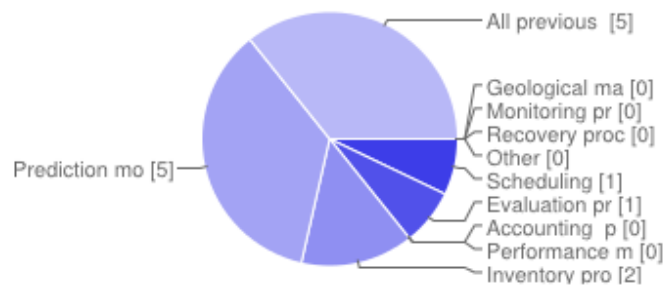
**6. What are the technology data entities that support new crude oil prediction/pricing system?**

Data base schema	8	57%
Storage management	0	0%
Data encryption	1	7%
Distribution system	1	7%
Internet and web application	2	14%
LAN/WAN networks	0	0%
All previous technology	2	14%
Other	0	0%



**7. What are the processes that you expect to need for computerization of the crude oil prediction/pricing system?**

Geological mapping	0	0%
Monitoring process	0	0%
Recovery process	0	0%
Scheduling	1	7%
Evaluation process	1	7%
Accounting process	0	0%
Performance management	0	0%
Inventory process	2	14%
Prediction model	5	36%
All previous process	5	36%
Other	0	0%



**8. How the current prediction is done within your industry and your research?**

- By geological mapping and evaluation process
- Using prediction model

- GIS mapping of potential extraction for long term
- Using predictive analytics
- We look at supply and demand fundamentals, stocks, geo-political factors as well at the marginal cost of production
- Statistical analysis of historical data
- From world bank price data
- Computational intelligence algorithm
- The Price of oil will increase by monitoring and adding a fixed percentage + or - 5%
- Based on my contribution, in an article, aimed to predict oil prices, I used a combination of neural network and wavelet decomposition
- The prediction of crude oil prices done in our Ministry through marketing Dept which, in turn, collect the data needed from many international publications sources like, Platts, Rim and so forth.
- Via the manage-float system whereby the price of fuel will be determined based on automatic -pricing mechanism whereby the ministry of Domestic Trade, Co-operatives and Consumerism will announce the oil price
- Based on worldwide price

**9. How often oil experts update the computerized prediction models?**

- Using all previous technology
- 1-2 per year
- Every day
- By mentioning their logical opinions regarding the price in consideration of all other parameters
- To update the prediction models they tend to include some determinants that may influence the price prediction within adopting new forecasting approaches and demonstrating their effectiveness compared to the classical ones.
- Quarterly, or when large discovery event
- No idea
- By prediction model
- As soon as new data becomes available.
- I do not know.
- After each tender
- On ad-hoc basis when the data is available
- No response
- Frequently

**10. How they take into account of the influence of other inputs (example: politics, war etc.)?**

- Using prediction model
- Using expert system
- For qualitative variables, they are included as dummy variables in the model.
- Any relevant input is considered but numerically.
- Not considered
- All previous risk in monitoring process
- They takes all that by looking for the transportation,insurance ,risk management .....etc
- Politics and Government
- Geopolitical factors play a role as long as supply is tight but demand can also be impacted by political turmoil
- Its added within a fixed error of assumption .. 5% + or - ..
- Oil prices are not determined entirely by supply, demand and market sentiment toward the physical product. Rather, supply, demand and sentiment toward oil futures contracts, which are traded heavily by speculators, play a dominant role in price determination. Cyclical trends in the commodities market may also play a role.
- No response
- I do not know
- Risk analysis

**11. Do you support that the crude oil prediction /pricing is becoming a sophisticated integrated system? If yes why?**

- No
- Yes because his impact is very important in political and Financial issues
- It is becoming complicated but still possible if one is aware of the fundamentals
- The crude oil market is a strongly fluctuating and non-linear market and the fundamental mechanism governing the complex dynamic is not understood. The oil market is the most volatile of all the markets and shows strong evidence of chaos. Therefore, oil price prediction is very important topic, albeit an extremely hard one due to its intrinsic difficulty and practical applications.
- Yes because the oil prediction might impact on the high dependence of most modern industrial transport, agricultural, and industrial systems on the low cost and high availability of oil.

- Yes, because so many factors are integrated to affect crude oil price eg politics, demand, supply, and war
- Yes because many important factors are necessary to be included in computerized model
- Yes. Because since it affects the life of a nation, it becomes very essential.
- Yes ,because the media for analyzing the data is dramatically improved and also there are so many analysts for the market specially for oil market
- No response
- I do not know.
- Not possible - Too many chaotic participants
- Yes .. by using mathematical models and updated data can increase the accuracy and corrected any errors of calculation ..
- Still in the research

**12. How the user can interact with the computerized model of the crude oil prediction/pricing system ? What are the minimal requirements? (If possible give brief description for system interface)**

- Using conventional interface design and following the computing principle of Input - Processing - Output (IPO).
- Short term energy outlook
- It depend on how easy and friendly is the model for the user .. if the critical calculation is done by the model in the back end and the input are clear , real, available and producing accurate results it will guarantee the sustainable use of the system ..
- Use existing, canned system such as Salesforce or Oracle
- It depends on how complex the model is. At my company, we still do crude price forecasts by personal judgment but refined product prices with the help of simple excel models
- Through system interface and the minimum requirement is to be computer literate
- Using monitoring process
- Single companies find many difficulties to do that because their results will no be consider globally so they most of the time depend on international publication witch prepare by professionals
- No response
- The best is to have a very user-friendly interface, which can help user to analyse the oil price via different type of graph. It is possible to have a simulator that can



estimate the oil price in the future depending on the input variable from the users.

- The best way to interact is to conduct a wide range of simulations dealing with training and testing phases and make the necessary change. I mean to give the required modification in order to make the model flexible with reference to the data complexity. It is also recommended to consider several software in doing simulations. Furthermore, it may be possible to manipulate the model internal structure and adapt it to the prediction requirements.
- System interface must interact with all previous entities including data base and other factors
- I do not know
- Still in the research

**13. Please determine the sequence (order) and timelines (timing) of the following activities and process in crude oil prediction /pricing system ;(timeline e.g daily ,weekly, immediately,.....)**

- Identification Extraction Distribution of crude Refining Distribution of refined
- Oil extraction (daily) Crude Specification (weekly) Classification of crude oil (weekly) Calculating price (daily) Cargo loading (monthly) Physical delivery (Spot price) (daily) Physical delivery (long term) yearly Assess prices by international publication (spot price) (daily) Buy/sell (bid or offer)(daily) Prediction of future prices (daily)
- Oil extraction - daily crude specification - weekly classification of crude oil - weekly calculating price - daily cargo loading - immediately physical delivery - daily physical delivery (long term) - weekly assess by international publication - daily buy/sell - immediately prediction of future price – daily
- Oil extraction - Once a year. Crude specification - once in three years. Classification of crude oil - once in three years. calculating price - monthly basis. Cargo loading - monthly. physical delivery -monthly. Assess prices by international publication - daily. buy/sell - daily. Prediction of future prices – monthly
- Oil extraction, .daily Crude Specification, daily Classification of crude oil , daily Calculating price, daily Cargo loading, upon tender period 3 month before delivery Physical delivery (Spot price), on tender close Physical delivery (long term), 3 months Assess prices by international publication (spot price) on tender preparation ,Buy/sell (bid or offer), quarterly Prediction of future prices,• on budget preparation twice year Other process with their order and timing
- Oil extraction - daily Crude Specification - daily Classification of crude oil - daily Calculating price - daily Cargo loading - weekly Physical delivery (Spot price)

- weekly Physical delivery (long term) - monthly Assess prices by international publication (spot price) - Monthly Buy/sell (bid or offer) - Daily Prediction of future prices - as needed.
- Oil extraction,prediction of future prices,assess prices by international publication, delivery
- No response
- Oil extraction, calculating price, cargo loading, etc
- Oil extraction, crude specification, physical delivery (spot and long term) buy/sell (bid or offer),assess prices and prediction
- Oil extraction (daily), Crude Specification (weekly), Classification of crude oil(weekly) , Calculating price(daily), Cargo loading(weekly), Physical delivery (Spot price) (weekly), Physical delivery (long term)(weekly), Assess prices by international publication (spot price)(daily) ,Buy/sell (bid or offer)(daily),Prediction of future prices(weekly)
- All the timing methods are required but important one the daily prediction because it is applied in most of the international publication
- Crude specification: daily, Classification of crude oil: immediately, Calculating price: monthly, Cargo loading: monthly, Physical delivery (spot price):monthly, Physical delivery (long term): quarterly, Prediction of future prices: quarterly.
- Daily

**14. As per their experience how many times the price changes more than 10% over a period of (3-5 business days)?**

- Zero
- 3
- 2
- 1
- 0
- Monthly
- Two times
- No response
- 3-5 times or on daily basis
- Once in two years
- One time
- No response
- 2 times
- Not sure

## List of survey participants

Name	Email	Institution	Country
Rich Burkhard	Ric ard.burkhard@sjsu.edu	San Jose State University	USA
Chiroma Haruna	freedonchi@yahoo.com	University of Malaya, Kuala Lumpur,	Malaysia
Haitham Babikir Yousif	hbabikir@gmail.com	Sudapet .. IT consultant	Sudan
Elfatih M Nour	elfatih201@hotmail.com	Manager	Sudan
Abdelrazig Abdelrahim	madomrahim@gmail.com	Sudapet	Sudan
Ehsan Ul-Haq	eulhaq@kbcac.com	KBC Energy Economics	United Kingdom
Fatai Adesina Anifowose	anifowo@kfupm.edu.sa	King Fahd University of Petroleum and Minerals	Saudi Arabia
Fathia kraiem	ktaiemfathia@yahoo.fr	Geologist	Tunisia
KraiemFahim		Entreprise Tnisienne d'Activites Pétrolières	Tunisia
		Petronas	Malaysia
Uwais Al-Qarni		Bonded	Malaysia
Tanjung		PET	Malaysia
Nadia	shaifulmilln@gmail.com	Company Cap Ayam	Malaysia
Rania Jammazi	jamrania2@yahoo.fr	faculty of economic and management Sciences and Management of Sousse	Tunisia