THE DRIVERS OF OIL PRICES

A DATA MINING APPROACH

By

NEHA SEHGAL

Under the Guidance of

Guide

Dr. KRISHAN KUMAR PANDEY

Assistant Dean - Research & Associate Professor
College of Management and Economic Studies

Submitted

In partial fulfillment of the requirement of the

DEGREE OF DOCTOR OF PHILOSOPHY

TO

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

DEHRADUN
MARCH, 2014

To My Family

Acknowledgement

This thesis would not have been written without the endless support, insightful supervision and consistent encouragement of my supervisor, Associate Professor Krishan K. Pandey. His insatiable appetite to gain knowledge, enthusiasm and patience have been invaluable to my training as a researcher. I am deeply indebted to him for leading me into the interesting field of energy and data-mining. Being a statistician, it was hard to make up my mind to switch from fundamental statistics to the world of data-mining during initial years of my graduation. My supervisors belief's in me was the energy drink to fall in love with the task of exploring and playing with hidden patterns in data.

I would like to thank other faculty and staff members of College of Management & Economic Studies department for their valuable feedback on this work and for providing a friendly and enjoyable environment during my time in the university.

I would also like to thank my sister, Mansi Kalra for having endless discussions regarding my work and convincing me to use LaTeX for writing thesis. I would like to thank my mother-in law, Aruna Sehgal for her affection and help in managing my life in busy times. Without her, it would have been difficult to focus on my research. My warmest thanks to my brother, Rahul Mehra for managing my food, stays and travels to university. I would like to thank my parents, Som Narian Mehra & Narender Kaur Mehra for their love and for nurturing me all these years. Their dreams and belief in me was the driving force for my efforts. Finally, I am faithfully indebted to my partner, Sidharth Sehgal for his endless support and love. I would like to thank him for the special efforts he puts in every time to make me feel lively during good and bad times.

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

(Neha Sehgal/March, 2014)

Thesis Completion Certificate

This is to certify that the thesis on **"The Drivers of Oil Prices - A Data Mining Approach "** by Neha Sehgal in Partial completion of the requirements for the award of the Degree of Doctor of Philosophy is an original work carried out by her under our joint supervision and guidance.

It is certified that the work has not been submitted anywhere else for the award of any other diploma or degree of this or any other University.

Internal Guide

(Dr. Krishan Kumar Pandey)

External Guide

(Dr. Neeraj Pandey)

Contents

Executive Summary

According to Energy Information Administration, oil will remain the world's primary fuel and is projected to remain the energy source with largest share for many years to come. Crude oil is traded in global market and fluctuations in oil prices have become a prime feature of any economy. The "price of oil "is a critical factor that has substantial impact on world economics, be it part of OPEC or Non-OPEC countries. This single international price is a key component that dominates investment decisions and plays a significant role to find ways to overcome prolonged instability within economy, to form better economic policies and to overcome sudden change in supply-demand framework. Oil prices have been steadily rising for several years and in July 2008 stood at a record high of $145 per barrel. Later, it declined due to global economic crisis at the end of 2008, and then recovered to $75 per barrel by 2010. Recently, oil prices have set records by surpassing $100 per barrel for the first time in the year (in money-of-the-day terms). This rise or decline in oil prices stimulates for studying in detail the factors behind movements in the price of oil.

Forecasting of crude oil prices has never been an easy task, though it is important for so many economic policies. Many institutions – including central banks and international organization–are currently using NYMEX oil futures as key indicator for the market's expectation regarding spot price development, but there are high number of external factors which are complex, noisy, and uncertain that drive crude oil prices. These include the behaviour of future markets, an assessment of the degree to which speculation are driving current price development, supply–demand framework, expected future reserves and impact of geopolitics on behaviour of stock market. Any rise or decline in these factors creates imbalance in the market which critically impact oil market participants and makes markets unpredictable. Thus, it becomes crucial to develop predictive models using various influential factors that drive crude oil prices to understand the

complex and dynamic nature of oil prices. There is no solitary indicator that act as stand-alone instrument for forecasting spot price movement but there are high number of external factors, which are complex, noisy and uncertain that drives oil prices.

There is colossal collection of data for factors, ranging from demand-supply, inventories, reserves to varied market, but an important task is to discover knowledge by identifying useful patterns in data. The process of mining information from data includes an important step of selecting an appropriate set of input variables. The output is significantly dependent on how much information is contained in the set of input variables for the study. The empirical literature is very far from any consensus about selecting the appropriate features/ indicators that can explain the true characteristics of oil market. In economics literature, there are many studies that had examined the relationship between oil prices and macroeconomic variables but there seem be to no consensus on the extent to which these macroeconomic variables are related to oil prices. Further, all existing methods of predicting oil prices have accounted for non-linearity, non-stationary and time-varying structure of the oil prices but seldom have focused on selecting significant features with high prediction power. Most of the researchers have considered predictor variables for oil price prediction based on judgemental criterion or trial and error method. To address this issue, primarily, there is an essential need to identify influential and informative features that can explain the characteristics of oil market.

A central problem in data mining is to identify a representative set of features to construct a prediction or classification model for a particular task. It is principal task to identify key factors driving oil prices through feature selection algorithm before proceeding for model building and evaluation. Feature selection play an important role in data mining to extract relevant and non-redundant features. An appropriate set of features can help in high prediction performance and thus, due care should be taken to select a set of relevant and non-redundant features. Most of the feature selection methods are based on the assumptions of conditional independence or need the number of features to be extracted. But still couldn't provide the minimal set of features that are most relevant and non-redundant for the study. The basic assumption of conditional independence of feature selection methods degrades the performance of model if features are strongly

inter-connected. Most of the real world problems contain features that are strongly inter-related to each other. Due to above mentioned research gaps, there is lack of robust feature selection method to select relevant and non-redundant factors for oil price forecasting which can incorporate complexities of crude oil prices. Thus, the objective of the study can be stated as "Mining the key factors driving oil prices using robust feature selection method for achieving high prediction performance."

To overcome this research gap, this thesis propose two new MI^3 and I^2MI^2 algorithms as feature selection methods to assess non-linear dependencies between oil prices and input variables. This thesis addresses the problem of feature selection for data mining through MI^3 and I^2MI^2 algorithms. These algorithms are built on the pillars of information-theoretic approaches. This thesis focus on identifying the relevant and non-redundant features for different time horizons, use artificial intelligent models as forecasting engines and achieve high prediction performance. This thesis focus on deriving the explanatory power of selected features based on proposed methodologies and their contribution as drivers of oil prices over decades.

MI^3 and I^2MI^3 algorithm are evaluated for two groups of dataset - A & B. Both algorithms are compared with competitive feature selection methods such as Modified Relief, Correlation based Feature Selection and Modified Relief + Mutual Information. The features selected from the proposed algorithms for both groups are fed to three forecasting engines: Multilayered Perceptron Neural Network, General Regression Neural Network and Cascaded Neural Network for comparison. In most analysis, prediction accuracy using the reduced features from MI^3 and I^2MI^2 algorithm is more accurate in comparison to other competing feature selection methods.

Experiments on both groups shows that I^2MI^3 algorithm quickly identifies most relevant and non-redundant features. Without perturbing on number of features to be extracted, on natural domain, MI^3 and I^2MI^2 algorithms eliminates more than $\frac{1}{2}$ and $\frac{1}{4}$ of the features. The features selected by I^2MI^2 algorithm is the minimal set of most relevant and non-redundant features for deciding direction of future oil prices. I^2MI^2 algorithm can enhance the performance of data mining problems, while at the same time can achieve significant reduction in number of features used in the study. The proposed algorithm operates on original feature set and doesn't incur

the high computational cost associated with repeatedly invoking the learning algorithm.

The out-of-sample forecasts using I^2MI^2 algorithm with General Regression Neural Network performed superior for both groups in comparison to EIA's STEO econometric model forecasts. Further, the proposed I^2MI^2 algorithm together with General Regression Neural Network is used to extract information regarding the explanatory power of factors and their contribution in influencing oil prices before and after 2008 financial crisis. The proposed algorithm is superior in extracting the influence of emerging economies in driving oil prices. The results shows the shift in influence of OECD consumption to Non-OECD consumption as a key indicator driving oil prices. China consumption and its reserves have emerged as influential factors driving oil prices post 2008 financial crisis with drastic increase in their percentage contribution. OPEC Supply is dominating the fluctuations in oil prices due to sudden change in production targets or policies. With recent increase in Non-OPEC production, the influence of older giants (OPEC) as the most influential factor driving oil prices is diminishing. the results highlighted that NYMEX future price is not a stand-alone instrument for predicting spot oil prices but there are high number of external factors that are required to be identified. Since oil is traded in global market and most of the trade has operated and continue to operate in dollars, U.S Dollar Index remains an influential factor driving oil prices. The results highlighted that the overall mechanism of oil market broke due to 2008 financial crisis. Speculation and reserves played an important role in driving oil prices before crisis while CPI and EPPI have largest contribution as key drivers of oil prices after crisis. The importance of reserves before the crisis was repercussion of cuts in OPEC production targets or changes in OPEC policies.

List of Figures

List of Tables

List of Symbols

R^2	Coefficient of Determination
A	Annual
AC	Analog Complexity
ACF	Auto-Correlation Function
ACIX	Autoregressive Conditional Interval Model with Exogenous Explanatory Interval Variable
AE	Absolute Error
AI	Artificial Intelligent
ALNN	Adaptive Linear Neural Network
AMIN	AI framework of Amin-Naseri et al.
ANN	Artificial Neural Network
APARCH	Asymmetric Power ARCH
AR	Annualised Return
ARIMA	Autoregressive Integrated Moving Average
BFGS	Broyden−Fletcher−Goldfarb−Shanno−Quasi Newton
BiP Sig	Bipolar Sigmoid
BLR	Bias Learning Rule
BNN	Boltzmann Neural Network
BP	Back-Propagation
BPNN	Back-Propagation Neural Network

BR	Bayesian Regulation
Br	Brent Crude Oil Market
BVaR	Bayesian Vector Auto-Regression
CA	Correlation Analysis
Ca-Var	Conditionally Autoregressive VaR
CC	Cluster Classifier
CrI	Crisis Index
D	Daily
DA	Day Ahead
Db	Daubechies
DirS	Direct Strategy
DNN	Decomposition based Neural Networks
DS	Directional Statistics
DT	Delta Test
Du	Dubai Oil Market
ECM	Error Correction Model
EGARCH	Exponential GARCH
EM	Expectation Maximization
EMD	Empirical Mode Decomposition
ENN	Elman Neural Network
Eqn.	Equation
FBS	Forward Backward Selection
Fig	Figure
FIGARCH	Fractionally Integrated GARCH
FIML	Full Information Maximum Likelihood

FLNN	Functional Link Neural Network
FM	Fuzzy Model
FNN	Fuzzy Neural Network
FP	NYMEX Future Prices
GA	Genetic Algorithm
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GB	Geometric Brownian Process
GD	Gradient Descent
GDX	Gradient Descent BEP
GPMGA	Generalized Pattern Matching Genetic Algorithm
GRNN	General Regression Neural Network
GSM	Grey System Model
GT	Gamma Test
HaT	Harr a Trous
HM	Hidden Markov Model
HQIC	Hannan-Quinn Info Criterion
HR	Hit Rate
HTS	Hyperbolic Tangent Sigmoid
HWBT	Hull White with Binomial Tree
IBL	Instance Based Learning
IGARCH	Integrated GARCH
IGP	Inverse Gaussian Process
JC	Judgemental Criterion
KAB	Genetic Programming framework of Kaboudan
L-RIM	Linear Relative Inventory Model

LD	Log-Differenced
Lgs	Logistic
LM	Levenberg-Marquardt Algorithm
LS	Logarithmic Sigmoid
LSE	Least Square Error
M	Monthly
MA	Month Ahead
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MFA	Manual Feature Extraction
MLP	Multi-layered Feed Forward Neural Network
MoGNN	Mixture of Gaussian NN
MRP	Mean Reverting Process
MSE	Mean Squared Error
NL-RIM	Non-linear Relative Inventory Model
NMSE	Normalised Mean Squared Error
NN	Neural Networks
NORM	Normalization
NRW	Naïve Random Walk
NSR	Noise-to-Signal Ratio
OLS	Ordinary Least Square
OU	Ornstein-Uhlenbeck Model
PACF	Partial Autocorrelation Function
PARCH	Power ARCH
PCP	Percentage of Correct Predictions

PGRP	Persian Gulf Region Prices
PMI	Partial Mutual Information
PR	Prediction Rate
PRMS	Pattern Modelling in Recognition System Approach
RBF	Radial Basis Function
RecS	Recursive Strategy
RM	Regression Model
RMA	Relative Change of Moving Average
RMS	Regime Markov Switching Stochastic Volatility Model
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RS	Regime Switching
RT	Return Transformation
RW	Random Walk
S-SVM	Standard SVM
SA	Step Ahead
Sig	Sigmoid
SM	Stochastic Model
SMAPE	Symmetric MAPE
SMP	Smoothing Procedure
SNR	Signal-to-Noise Ratio
SoMLP	Self-organizing MLP
SP	Spot Prices
SR	Scaling Range
SSE	Sum of Square Error

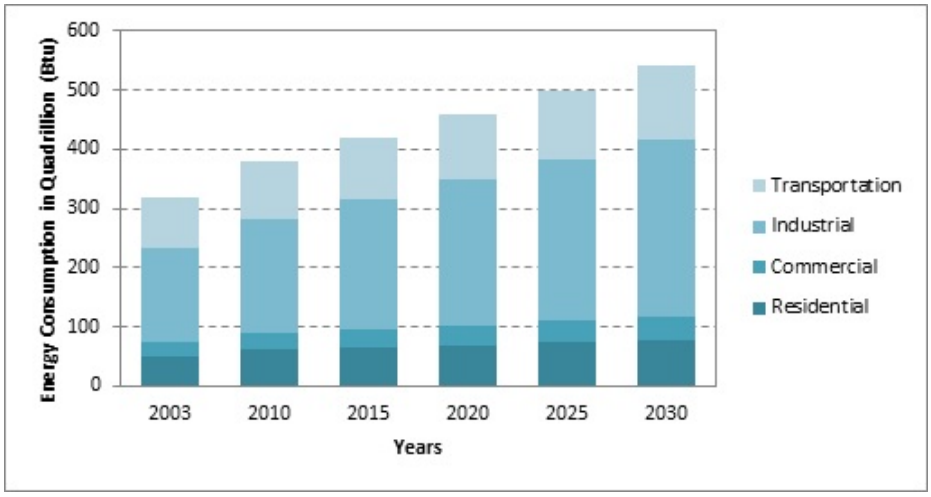
STEO	EIA's Short-Term Energy Outlook Econometric Model
SVM	Support Vector Machine
SVR	Support Vector Regression
TE	Trial and Error Method
TGARCH	Threshold GARCH
TM	Text Mining
TPA	Time Period Ahead
TSig	Tangent Sigmoid
TSK	Takagi-Sugano-Kang
VaR	Value-at-Risk Model
VECM	Vector Error Correction Model
W	Weekly
WA	Week Ahead
WANG	AI framework of Wang et al.
WCI	Without Crisis Index
WDE	Wavelet Decomposition Ensemble
WNN	Wavelet Neural Network
WSP	Without Smoothing Procedure
WT	Wavelet Transform
WTI	West Texas Intermediate Crude Oil Market

Chapter 1

Introduction

1.1 Background

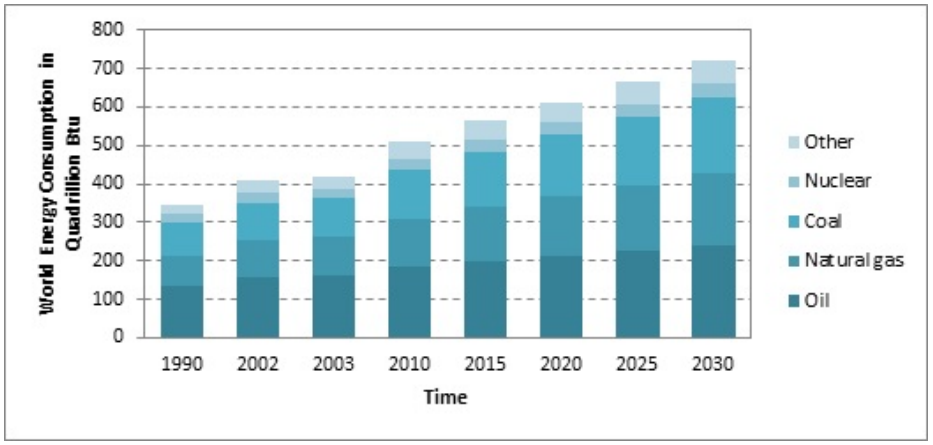
According to IEA [1], the global energy demand is expected to grow by one-third from 2011 to 2035 with emerging economies like China, India and the Middle-East accounting for 60% of the increase. On the basis of global scenario, strong economic growth is expected to continue increasing the energy uses [1]. A nation's industries, offices, homes and vehicles are altogether power-driven by energy, which energies the economic growth of the world [2]. Fig 1.1 shows the world energy consumption by different end use sectors. It shows that energy demand is concentrated in transport sector.



Source: IEA World Energy Outlook, 2012

Figure 1.1: World energy consumption by end user sectors (2003–2035)

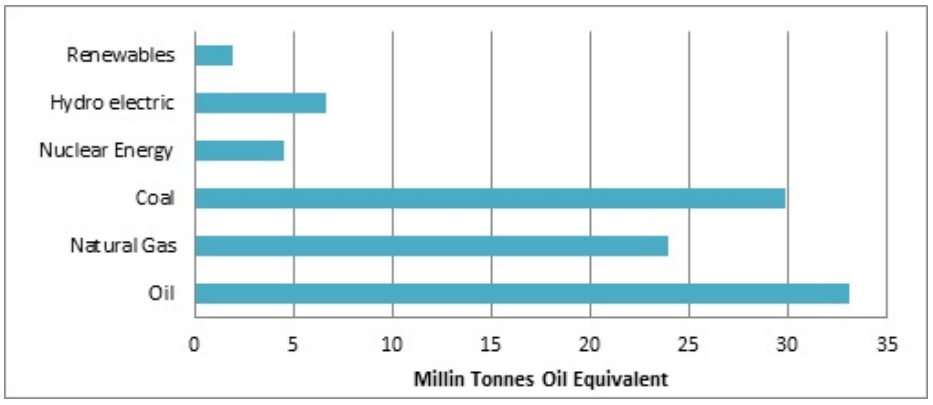
Time series analysis shows that energy and economic growth co-integrate and energy use Granger causes economic growth in long run [3] [4]. The



Source: IEA World Energy Outlook, 2006

Figure 1.2: World supply of primary energy by fuel type (1990–2030)

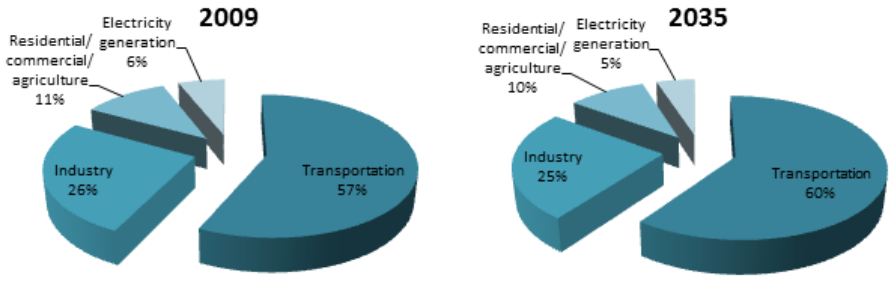
world supply of primary energy by fuel type is shown in Fig 1.2. Fossil fuels currently account for 87% of primary energy demand and are projected to still make up to 82% of the global demand by 2035 [1] . It is evident from Fig 1.2 that oil will remain the energy source with the largest share for most of the projected periods and will continue to play a foremost part in satisfying world energy needs [1]. Oil plays an inevitable role in the current economic scenario and is thus termed as black gold [1]. According to BP [5], oil remains the world's primary fuel, accounting to 33.1% of global energy consumptions as shown in Fig 1.3.



Source: Statistical Review of World Energy, 2013

Figure 1.3: Percentage share of primary energy world consumption by fuel type in 2012

Further, the percentage share of world oil demand by sectors in 2009 to 2035 is shown in Fig 1.4. The main sector that affects oil demand is the transportation sector. In 2009, transportation accounted for 57% of global oil use followed by 26% of industrial usage The residential and agriculture



Source: OECD/ IEA Energy Balance of OECD/ Non-OECD countries, 2011

Figure 1.4: Percentage share of world oil demand by sector (2009–2035)

sector together with demand from commercial sector contribute to 11% share in world oil demand. As a result of economic growth, number of cars is expected to double by 2035 and road freight would be responsible for almost 40% of the increase in global oil demand [1]. Therefore, the projected economic growth by 2035 inclines toward investments in the transportation and energy E&P infrastructure simultaneously [6].

1.2 Prominence of Oil Prices

On a day to day basis, millions of tonnes of oil is being moved around the world and it then refined to make diesel, gasoline, lubricants and other petroleum products. Though petrol and diesel are used to run engines, other fuels are used in numerous industries to produce commodities like metal, plastic and furniture. The key businesses of any country are dependent on oil prices e.g. power sector, mining sector, auto-mobile industry, airline industry, consumer goods industry, transport industry, chemical and pharmaceutical industry, food industry, textile industry and aerospace industry.

Petroleum based fuels currently meet the majority of human energy requirements [7]. Hence, economy of any country can't run without oil and any fluctuation in crude oil prices affects its economy directly. When crude oil prices are high, it leads to slower economic growth [4]. Since crude oil prices are directly linked with the economy, therefore high oil prices can lead to high inflation and thereby retarding economic growth. Any fall in crude oil price leads to lower inflation which increases economic growth. Since inflation decreases, interest rates reduces while consumer and business spendings increases. Any fluctuation in crude oil prices has distinct

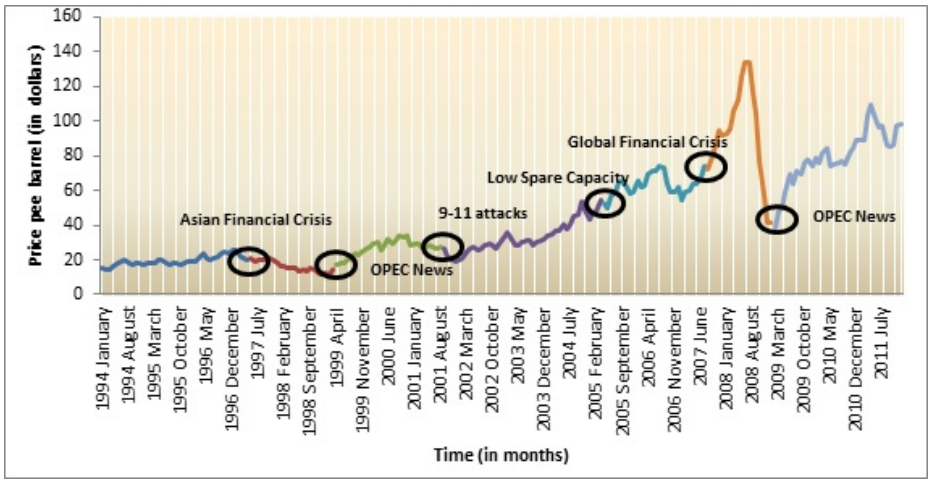
effect on stock market as well [8] [9]. This fact is difficult to conceptualize but same has been observed throughout the world.

This single international price is a key component that will continue to dominate investment picture for years to come on. It acts as a key variable in evaluation of economic development, energy policy decisions and stock markets [5]. A prior knowledge of oil price fluctuations helps oil producers to take decision regarding increase or decrease in production levels. Oil prices help strategically in macroeconomic projections and macroeconomic risk analysis for central and private banks. They are helpful in predicting recession in business cycles [10]. They are helpful in planning regulatory policies regarding taxes & standards. Businesses dependent on oil will be benefited as firms will be in position to take measures to control manufacturing and sales of their products in line with expected trend of forecasted oil prices. Accurate forecasting helps Non-OPEC countries to take effective measures so that their growth remains robust and thus benefiting their consumers. Further, economic policies can be formulated in a way to overcome recession and unemployment. Oil price fluctuations are of deep economic risk, both to producers as well as to consumers. Fluctuating prices and regional supply disruptions lead to considerable uncertainty to the near-term outlook.

1.3 Oil Price Scenario

Crude oil is traded in global market and thus oil price volatility has become a prime feature of global oil market. According to EIA, geopolitical and economics events have strong impact on crude oil markets over last 40 years. Oil prices are affected by geopolitical and economic events that have potential to disrupt the flow of oil and petroleum products to markets. These events cause disruption in the supply–demand framework, leading to an increase in volatility of oil prices. Crude oil prices have react to variety of events ranging from happening of cuts in OPEC production target, financial crisis, terrorist attacks to political disruption in oil exporting countries. Fig 1.5 befittingly demonstrates the effects of the various geopolitical and economic events on the crude oil prices.

Oil prices rose to $30 per barrel by end of 1996 but due to Asian financial crisis in 1997 quarter 1, prices declined drastically to $16 per barrel



Source: U.S. Energy Information Administration, Thomson Reuters

Figure 1.5: Crude oil prices and Geopolitical & Economic events (1994-2012)

by end of 1998. As a consequence of cuts in OPEC production targets by 1.7 mmbpd, oil prices increased to more than $35 per barrel in late 2000. The impact of such extreme events is of prime importance as they impact the oil market. These price fluctuations were reproduce due to divergent factors such as Iran-Iraq political conflicts, OPEC supply disruptions, 9/11 attacks or global financial crisis. In quarter 3, 2001 when oil prices were around $34 per barrel, terrorist activity (9/11 attack) led to increase in volatility of oil prices, soaring oil prices above $54 per barrel in late 2004. Further, oil prices steadily rose for several years and reached a record high of $145 per barrel in July 2008 due to low spare capacity. Afterwards, global financial crisis in 2008 caused oil prices to plunge to around $43 per barrel by end of 2008. In quarter 1 2009, OPEC cut production targets by 4.2 mmbpd and thus oil prices rose from $43 per barrel to $91 per barrel by end of 2011. Basis above details, the question that arises is whether volatility in oil prices is due to variation in availability or are there any other political or economic indicators to blame. Therefore, it is essential to analyse the key indicators driving oil prices.

1.4 Rationale & Motivation

The “price of oil ”is a critical factor that has substantial impact on world economics, be it part of OPEC or Non-OPEC countries. Oil prices have been steadily rising for several years and in July 2008 stood at a record high of $145 per barrel. Later, it declined due to global economic crisis at

the end of 2008, and then recovered to $75 per barrel by 2010. Recently, oil prices have set records by surpassing $100 per barrel for the first time in the year (in money-of-the-day terms). This rise or decline in oil prices stimulates for studying in detail the factors behind movements in the price of oil.

Understanding complex oil price movements and indicators driving them was the impetus for Energy Information Administration (EIA) to launch a monthly report assessing the physical market, financial and trading factors influencing oil prices. Looking ahead, there are several issues that can inform us about the direction of oil prices. These include the behaviour of future markets, an assessment of the degree to which speculation are driving current price development, supply–demand framework, expected future reserves and impact of geopolitics on behaviour of stock market. Any rise or decline in these factors creates imbalance in the market which critically impact oil market participants and makes markets unpredictable.

Forecasting of crude oil prices has never been an easy task, though it is important for so many economic policies. Many institutions – including central banks and international organization–are currently using NYMEX oil futures as key indicator for the market's expectation regarding spot price development, but there are high number of external factors which are complex, noisy, and uncertain that drive crude oil prices [11]. Thus, it becomes crucial to develop predictive models using various influential factors that drive crude oil prices to understand the complex and dynamic nature of oil prices. As discussed, there is no solitary indicator (lags, future prices or macroeconomic variables) which can provide a complete picture of how prices can be determined but definitely, there are several factors that can inform us about the direction of future price path. The key factors driving oil prices can give us a snapshot of some fluctuations in oil prices and by modelling these key snapshots can give a clear picture on direction of oil prices. Taking into account that complex and chaotic tendency of crude oil prices are due to significant variation in key external factors, a combination of these key factors can help to overcome some of the drawbacks of previous econometric models used for forecasting of oil prices.

There is colossal collection of data for factors, ranging from demand-supply, inventories, reserves to varied market, but an important task is to discover knowledge by identifying useful patterns in data. The process of mining

information from data includes an important step of selecting an appropriate set of input variables. In most of the studies, the choice of input variables for oil price forecasting is carried on judgemental criterion or trial and error procedures. Little attention is paid on selecting influential factors and more is on assessing new techniques for oil price forecasting. The central problem of identifying a representative set of features to construct a prediction model for oil prices is a major issue of concern. To address this issue, primarily, there is an essential need to identify most relevant and non-redundant features that can explain the characteristics of oil market.

1.5 Outline of the Study

It is apparent that the central problem is to primarily identify a representative set of input variables for data mining the characteristics of oil prices. However, a comprehensive robust feature selection method is needed that can account for non-linear, chaotic and complex relationship between oil prices and candidate features. Existing methods of predicting oil prices have accounted for complexities in oil prices but seldom focus on identifying key factors driving them (chapter 2 covers this in detail). What constitute the basic elements of oil price prediction model, at a fundamental level, are required to be known to draw up the entire architecture of this intervention.

The main objectives of this thesis are to review the methodological approaches adopted for forecasting oil prices, identify key indicators driving oil prices by studying the association and dependency structure between oil prices and factors driving them, select relevant and non-redundant drivers of oil prices by building robust feature selection algorithms based on pillars of information-theoretic approaches, use artificial intelligent models as forecasting engines, evaluate the results of proposed methodology with existing pool of methods and empirically test the proposed methodology for forecasting oil prices in different time-horizons.

To achieve the above mentioned objectives, this thesis does an exhaustive literature review of oil price forecasting methodologies used by researchers, discuss about the factors used by them to develop oil price forecasting models and identify the basic building blocks (variables) that are needed to provide the details about direction of oil prices. In this thesis, two robust feature selection methods are proposed basis the objectives of the

study. The study identifies the minimal representative set of factors driving oil prices in different time-horizons and use them as inputs to neural networks. The proposed framework is used for one-month and twelve-month ahead predictions of oil prices. The results are presented in chapter 4 & 5 of this document.

1.6 Business Problem

The accuracy of predicted crude oil prices is the biggest hurdle faced by the global oil markets. Accurate forecasting of crude oil will help both oil producers and consumers to a large extent. Oil producing countries shall be benefited as they shall be in a position to increase or decrease their productions basis the predicted prices. Consumers shall be benefited as their economy is highly dependent on oil. Accurate forecasting will help Non-OPEC countries to take effective measures so that their growth remains robust. The businesses which are dependent on oil will be benefited as they shall be in a position to take measures in advance to control their manufacturing, sales, and inventories basis expected trends of forecasted oil prices. With accurate projection of oil prices, economic policies would be formulated to overcome recession, high inflation & unemployment.

Using future prices only as a stand-alone instrument to forecast spot price development doesn't seem to be recommendable. For designing better structural forecasting models, it is important to identify the key factors driving crude oil prices during the time frame of happening of such extreme events. In this context, a feature selection algorithm (based on dependency and association of key factors with oil prices) is required to identify the minimal set of key factors for achieving high prediction performance.

Thus, the business problem of this thesis is to identify key drivers of oil prices for better investment decisions, find ways to overcome prolonged instability within economy and form better economic policies to overcome sudden change in demand-supply framework so as to reduce public deficits.

1.7 Contribution of the Study

This research leads to general contributions to the field of data mining and applied energy. The contributions of this study are as follows:

- A new three stage I^2MI^2 algorithm for feature selection method, that performs very competitively when compared with several state-of-the-art feature selection methods. The study present both theoretical and empirical contributions. (Data Mining contribution)

- A new two stage MI^3 algorithm for oil price prediction which simultaneously improves the performances of oil price predictions using significant input variables. (Data Mining contribution)

- The new proposed algorithms provide 100% non-redundant and relevant features than previous feature selection methods for applications in various disciplines. The explanatory power of key indicators influencing oil price market before and after financial crisis is presented. (Data Mining and Applied energy contribution).

- A new ensemble learning algorithm ($I^2MI^2 + GRNN$) for prediction of oil prices with extensive empirical evaluation with EIA's STEO econometric model (Data Mining Contribution).

- A framework which can be used for predicting future value of oil prices depending upon movements in significant inputs driving them (Applied Energy Contribution)

- The novel I^2MI^2 algorithm , which can be seen as a realization and an application of the proposed framework. Our experiments on real world problem show that the proposed algorithm performs better compared to other competitive ensemble models. (Data Mining Contribution)

1.8 Organization of the Thesis

This thesis is organised as follows:

The study consists of six chapters. The first chapter is the introduction to the topic that includes the world energy consumption to sustain its current economic growth, the contribution of fossil fuels in primary energy supply and the share of oil that continue to play a foremost role in satisfying world energy needs. This chapter also discusses about the significance of oil prices and how the fluctuations in oil prices cause disruption in economy. It highlights the world oil price scenario as a consequence of happening of

various geopolitical and economic events. The first chapter explains the rationale and motivation to conduct this research work together with the contribution of the study in various disciplines.

The second chapter reviews the literatures that study the methodological approaches used by researchers for oil price forecasting, emphasizes on artificial intelligent methods which are now being used extensively and attempts to provide in-depth review on essential parameters for the study. The variables used in artificial intelligent models as inputs are also presented. This chapter also highlights the research gaps as a result of extensive literature review. It explains the rationale of the study followed by the statement of the research problem, objectives of the study, research questions and scope of the study.

The third chapter discusses in detail the data mining process followed in the study and basic concepts of information-theoretic approaches used for building two proposed feature selection algorithms. The neural networks are used as forecasting engines for the study and the key issues related to them are addressed in this chapter. It explains the data sample and preparation done for analysis of primary data.

The fourth chapter introduces a new MI^3 feature selection algorithm with both theoretical and empirical analysis to identify key drivers of oil prices. An alternative ensemble learning algorithm called I^2MI^2 for finding the minimal set of non-redundant and relevant features as input for designing oil price prediction model is explained. The models design and analysis are explained step-by step in this chapter. Both algorithms are evaluated and confirms the superiority of proposed algorithms in comparison to other known feature selection methods. The formulated research model is empirically tested and consequent results are reported.

The fifth chapter identifies the shift in influential power of factors driving oil prices in rising and downturn period. The contribution of factors to oil price volatility before and after 2008 financial crisis is discussed in detail.

In the end, the sixth chapter concludes the thesis with additional remarks and addresses some potentially important research directions for future work. Bibliography is given at the end as reference.

1.9 Concluding Remarks

Oil plays an inevitable role in the current scenario and is termed as black gold. Oil will remain the world's primary fuel and is projected to remain the energy source with largest share for many years to come. This single international price acts as a key variable in evaluation of economic development, energy policy decisions and stock markets. Oil prices have risen or fallen due to happening of various geopolitical and economic events. These fluctuations in oil prices raise an important question whether this rise or fall in oil price is due to squeeze in availability or are there any other political or economic indicators to blame. Researchers have forecast oil prices using future prices as key indicator but future prices as sole indicator is not sufficient. There are high number of external factors that drive oil prices. The key factors driving oil prices can provide a clear picture about direction of oil prices.

This thesis attempts to fill the gaps owing to nascences literature on factors that may drive the direction of oil prices, by identifying a set of variables that form the core component of oil price prediction model. The selected features are used to develop & empirically test a model for oil price prediction. It is expected that this thesis shall add to the existing body of knowledge as literature on the oil price forecasting using influential factors is at nascent stage. In next chapter, extensive review is presented to understand the current pool of literature and to identify research gaps that exist.

Chapter 2

Literature Review

2.1 Overview

This chapter reviews whole spectrum of methodological approaches adopted for forecasting crude oil prices by using various influential factors that drive oil prices. The summary of variety of approaches employing time series models, financial models and structural models is presented. Once the literature is studied to develop an overall understanding of the current scenario, it becomes imperative to build on the foundation of existing body of knowledge by identifying the research gap that exist. Thus, these research gaps become the starting point of the research work. The objectives of the literature review are:

- To study the overall oil price forecasting methodologies used by researchers.
- To discuss the types of methods used in Time Series models framework for oil price forecasting.
- To discuss the types of methods used in Fundamental models framework for oil price forecasting.
- To briefly discuss about the factors used by researchers in developing stochastic or regression models for oil price forecasting.
- To identify and elaborate on the list of artificial intelligent models used for forecasting of oil prices to create the master list of input variables (identified through literature review).
- To identify the research gaps in the existing body of knowledge that then becomes the basis for conducting this research work.

- To set the research objectives for the study, followed by research questions and then scope of the study.

2.2 Econometric Models

As discussed in chapter 1, forecasting of crude oil prices is an important task for better investment management, macroeconomic policies and risk management. It is important to analyses the probabilistic assumption of oil prices in terms of normality, leptokurticy, linearity and serial correlation [12]. To forecast crude oil prices, a variety of approaches have been proposed by numerous authors employing time series [13] [14] [15] [16] [17] [18], financial models [19] [20] and structural models [21] [22] [23] [24] [25] [26].

2.2.1 Time Series Model

Time series analysis is a method of forecasting that focuses on the historical behaviour of dependant variable. Oil prices are assumed to be normally distributed in many studies but their departure from normal distribution was disregarded due to misinterpretation of Central Limit Theorem [12] [27]. Crude oil prices are found to be non Gaussian. Forecasting crude oil prices through fundamental method is a complex task due to uncertainty, noisiness and non-stationary inbuilt in indicators that drive them. Therefore, time series models provide an alternative to analyse and predict future movements based on past behaviour of oil prices [11]. The price-forecasting models based on time-series approach have been further classified into three subsets as shown in Fig 2.1. The summary of time series forecasting models

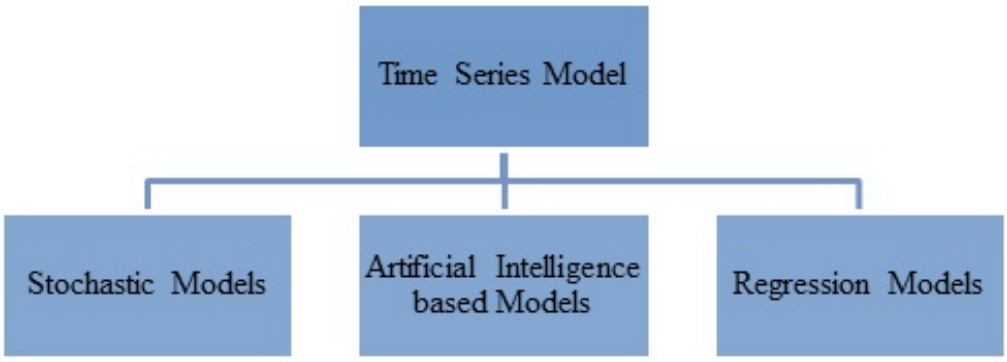


Figure 2.1: Classification of time series oil price-forecasting models

based on type of model being employed and methods used by researchers

have been presented in Table 2.1. Stochastic models are inspired by financial literature and are widely applied in forecasting of oil prices. There are several stochastic models which have been employed for modelling and forecasting of crude oil prices such as Random Walk [17] [15] [28] [29] [14], Mean Reverting Processes [30], Brownian Motion Processes [31], Ornstein-Uhlenbeck Processes [31], Inverse Gaussian Process [27] and Jump Diffusion Processes [32].

Regression type models are based on the relationship between oil price and number of exogenous variables that are known or can be estimated. The most common approaches employing regression models are ARIMA models [33] [20] and GARCH-family models [33] [16] [14] [34] [35] [36]. Arouri et al. [13] employs GARCH models to forecast conditional volatility of spot and future oil prices with structural breaks for better forecasting performance. Huang et al. [37] and Hou [34] presented superior performance of non-parametric GARCH models relative to parametric GARCH models (in-sample and out-of-sample volatility forecasts). Researchers concluded that non-linear dynamical approach is more appropriate for characterizing and predicting crude oil prices than linear approach [38] [39]. The parameters of forecasting models for crude oil prices have been estimated by either Least Square Method [40] [41] [42] [43] [44] [30] [20], Full Information Likelihood Method [24], Kalman Filter [30] [31] or under Bayesian Framework [45]. However, these numerous estimation algorithms have failed to achieve high prediction accuracy. Stochastic models involving certain characteristics of oil prices and regression models have been kept outside the scope of this review. A review of these econometric time series models for oil price forecasting has been presented by Frey at al. [67].

Table 2.1 provide summary of time series models for crude oil price forecasting. In recent times, artificial intelligent models are extensively being used to capture unknown or too complex structure in the time series. Researchers have used artificial intelligent model based approach for oil price forecasting in more than 50% of the studies listed in Table 2.1. Out of thirty six studies listed in Table 2.1, twenty eight studies have considered WTI crude oil spot prices as dependent variable in their studies. Section 2.3 covers various artificial intelligent models ranging from single models (e.g. neural networks, support vector regression, wavelets) to more complex hybrid versions.

Table 2.1: Summary of time series models for crude oil price forecasting

Author	**Oil Market**	**Model Type**	**Methods**
[17]	WTI	SM, RM	RW, VaR, ECM
[27]	WTI	SM	IGP
[46]	WTI	RM	ARIMA
[30]	WTI	RM, SM	OLS , MRP
[47]	Br	AI	WNN
[31]	FP	SM	GB, OU
[8]	WTI, Br	RM	GARCH
[33]	FP	RM, AI	ARIMA, GARCH, ANN
[48]	FP	AI	GSM
[18]	WTI	AI	FM
[49]	WTI, Br	AI	SVM
[50]	WTI	AI	SVM
[51]	WTI, Br	AI	ANN
[52]	WTI	AI	WNN
[15]	WTI, Br	AI, RM, SM	WDE, RW, ARMA
[53]	WTI	AI	ANN, WDE
[13]	WTI, FP	RM	GARCH
[54]	FP	RM	RS-EGARCH
[55]	WTI, Br	AI	GA + ANN
[56]	WTI	AI	FNN
[57]	WTI	AI	RMS
[34]	WTI	RM	GARCH
[58]	FP	RM	GARCH
[59]	WTI	AI	WDE
[36]	WTI, Br	RM	GARCH
[60]	WTI	RM	RS-GARCH
[29]	WTI	SM, RM	VECM, NRW, ARIMA
[61]	WTI	RM	CA-VaR
[62]	FP	AI	GA + ANN
[63]	WTI	AI	SVR
[64]	WTI	AI	SVR
[65]	WTI, Br, Du	AI	WNN
[66]	WTI, Br	AI	FNN
[35]	WTI, Br, Du	RM	CGARCH, FIGARCH, IGARCH
[14]	WTI	SM, RM	RW, HM, ARIMA, GARCH, EGARCH, TGARCH, PARCH, CGARCH
[16]	SP	RM	GARCH, EGARCH, APARCH, FIGARCH

2.2.2 Fundamental Models

Fundamental models predict oil prices based on their relationship with economic, financial, social and political indicators that drive them. This study assumes financial and structural models as part of fundamental models. Crude oil prices have been influenced by large number of factors which are complex, noisy, and uncertain [11]. There is no single indicator which can provide a comprehensive portrait of how prices can be determined. Each indicator can give us a snapshot of present condition and modelling of these significant snapshots together provides a clear picture of direction of oil prices. Similar to time series models, fundamental models can also be classified into two major classes: Regression models and Artificial Intelligent based models as shown in Table 2.2. This table enlists the analytical methods used by authors for forecasting crude oil prices.

Globalization hypothesis holds that oil prices (WTI-Brent, WTI-Dubai, WTI-Maya and Dubai-Maya) move together and exhibit greater conditional dependency [68]; therefore, most of the study listed in Table 2.2 considers WTI spot crude oil prices as benchmark price. It is evident from Table 2.2 that around 50% of the studies have incorporated artificial intelligent models for forecasting oil prices. As evident from Table 2.2, researchers have preferred ordinary least square methods for parameter estimation in a regression model. The different input variables, along with the class they belong to, used by different researchers are presented in the next section.

Factors driving Oil Prices

Oil prices had shown upward trend in 1996 but prices declined drastically by end of 1998. As a consequence of cuts in OPEC production targets, oil prices increased again in late 2000. The impact of such extreme events is of prime importance as they effect the direction of oil prices and thereby the objective of increasing the predication accuracy of crude oil prices. It is important to identify the key indicators driving crude oil prices (during the time-frame of happening of such extreme events) for designing better structural forecasting models. In 2001, 9/11 attack led to increase in volatility of oil prices that soared oil prices till 2004. Oil prices have been steadily rising for several years and in July 2008 stood at a record high of $145 per barrel due to low spare capacity. Later, it declined due to global economic crisis at the end of 2008 and then recovered to around $75/bbl

Table 2.2: Summary of fundamental models for crude oil price forecasting

Author	**Oil Market**	**Model Type**	**Methods**
[69]	WTI	AI	ANN
[70]	WTI	AI	WNN
[41]	WTI	RM	OLS
[42]	WTI	RM	OLS
[43]	WTI	RM	OLS
[44]	WTI	RM	OLS, ECM
[71]	WTI	RM	OLS
[25]	WTI	AI	WNN
[40]	WTI	RM	OLS
[24]	WTI	RM	OLS, FIML, ECM
[20]	WTI	RM	OLS, ARMA
[72]	Br	AI	GRNN
[26]	WTI	AI	ANN, TM
[73]	WTI	AI	FNN
[22]	PGRP	AI	ANN
[74]	WTI	RM	VECM
[75]	WTI	RM	ACIX
[76]	WTI	AI	WDE
[77]	FP	AI	WDE
[78]	FP	RM	GARCH
[79]	FP	RM	GARCH
[39]	Br, WTI, Du	RM	ESTAR
[80]	WTI	RM	CGARCH
[45]	WTI	RM	BVaR
[81]	FP	RM	ECM
[82]	SP	AI	ANN
[83]	WTI	AI	SVR
[84]	NF	AI	SVR
[85]	SP	AI	GA
[52]	WTI; Br	AI	WNN
[86]	WTI	AI	WNN

by 2010. In 2013, oil prices have set record by surpassing $100/bbl for the first time in that year (in money-of-the-day terms). This rise or decline in oil prices stimulates for studying in detail the factors behind movements in oil prices. There are large number of factors, which are complex, noisy, and uncertain influencing crude oil prices [11]. Understanding complex oil price movements and indicators driving them was the impetus for Energy Information Administration (EIA) to launch a monthly report to assess the financial, trading and physical market factors that influence oil prices.

Beside geopolitical and economic events, any fluctuation in demand or supply side also creates imbalance in the market that critically impact oil market participants and makes markets unpredictable. The demand-supply framework plays a crucial role to an extent that it determines the directions of crude oil prices but has not been sole indicator that drives oil prices. Researchers have considered enormous factors such as GDP, inventories, emerging economies and stock market fluctuations to study their influence on oil prices (Table 2.3 covers this in detail). There is no solitary indicator (lags, future prices or macroeconomic variables) that can provide a complete picture on how prices have been determined, however there are few key indicators that have governed and ruled crude oil prices. The key indicators can give us snapshots of fluctuations in oil prices and modelling of these significant snapshots only (as input variables) can give a clear picture on directions of oil prices. The fluctuations in these factors cause complex, volatile, non-linear and chaotic tendency of crude oil prices, therefore, it is important to find the key strategic indicators that are ruling crude oil prices from decades. Thus, it has become crucial to develop predictive models using various influential factors that drive crude oil prices to understand the complex and dynamic nature of oil prices.

This review has identified petroleum inventory level as a virtuous market indicator of change in crude oil price. Inventory levels have been a measure of balance or imbalance between production and demand [70]. Ye et al. proposed a linear forecasting model using relative inventory level as input for oil prices [42] but later improved Linear-RIL model to Non-Linear-RIL model due to dynamic relationship between them [43]. Pang et al. [70] proposed to consider both crude oil inventory level and petroleum product inventory level as input factors for better forecasting performance. Weiqi et al. [23] constructed a structural econometric model using relative inven-

tory and OPEC production as explanatory variables for short-run oil price forecasts. Other than inventories level, variables like production, net imports and forward prices were taken as independent variables to estimate spot prices by Considine and Heo [87]. The evidence for the non-linear relationship between GDP growth and oil prices has been examined by Hamilton [40] and Kim [88].

Zamani [44] examined a short term quarterly econometric forecasting model using OECD stocks, Non-OECD demand and OPEC supply to forecast WTI crude oil prices. Ye et al. [41] considered the possible substitution for large proportion of world demand and inventory (in form of OECD demand) as input variable compared to U.S demand alone. Dées et al. [71] assessed a structural econometric model to show the immediate impact of OPEC quota decisions and capacity utilization on oil prices. According to BP [5], emerging economies (especially Asia) accounted for major net growth in energy consumption whereas OECD demand remains falling for a third time in last four years. Ratti and Vespignani [89] has highlighted that intense increase in China and India liquidity led to increase in real oil price and production. Li and Xiaowen Lin [90] has provided evidence indicating demand from China and India as leading driver in the world oil pricing system since 2003. Zhang and Wang [91] indicated a greater contribution (95.71%) of crude oil future markets in price discovery function of spot price. Alvarez-Ramirez et al. [21] used fourier analysis to examine the strong relationship between U.S macro economy and crude oil prices. Dées et al. [24] has examined the dynamic relationship between oil prices and OPEC capacity utilization. Murat and Tokat [74] examined the forecasting power of crack spread futures under vector error correction framework to predict spot oil markets. Yang et al. [75] studied the influence of European debt crisis and financial crisis on crude oil prices with the proposed Autoregressive Conditional Interval Model with Exogenous Explanatory Interval Variable (ACIX). He et al. [77] used error correction models to examine the influence of Kilian economic index (as global activity indicator) on crude oil prices.

Bu [78] examined the relationship between speculative traders' position and crude oil future prices using estimates of GARCH model. Basher and Haug [92] proposed a structural vector autocorrelation model to investigate the dynamic relationship between emerging markets, stock prices

and oil prices. Chai et al. [45] developed a oil price VaR model based on path-analysis using core influential factors. Zhang et al. [76] proposed an empirical mode decomposition-based event analysis method to estimate the impact of extreme events on crude oil prices. Fong and See [60] suggested regime switching models framework to study factors driving crude oil prices volatility. Ignoring the clear evidence of presence of non-linearity or structural breaks in the oil price series can lead to false imprint on predictability and persistence. Oil prices are shown more sensitive to any change in oil supply by Chai et al. [45] but weekly dependent on exchange rates by Reboredo [68]. The study of factors driving oil prices that are considered in various stochastic or regression models of oil prices forecast is a major area of research and has been kept outside the scope of this review.

Despite above mentioned attempts, oil price prediction has remained a difficult problem due to its complex non-linear and time-varying nature. In addition, recent studies lay emphasis on developing structural econometric models for forecasting crude oil prices without focusing on finding the key drivers of oil prices. Most recently, a category of artificial intelligent models have emerged and are being attempted to predict oil prices. AI based framework for oil price prediction are discussed in detail in section 2.3. The factors influencing oil prices may be classified within the categories: **C1 -** supply, **C2 -** demand, **C3 -** inventories, **C4 -** price, **C5 -** reserves, **C6 -** economy, **C7 -** world events, and **C8 -** properties. There are as many as fifty one variables used in different AI related studies for crude oil price forecasting. A detailed list of factors driving oil prices considered in artificial intelligent models, along with the class to which they belong are presented in Table 2.3. In order to build an effective model, careful attention should be paid on selecting informative and influential inputs which cause changes in prices [69]. However, until recently, the input variables of oil price forecast have been selected on judgemental criteria or trial and error procedures [93] [49] [83] [63] [84] [64]. This review discovers that most of the studies were concentrated on non-linearity, non-stationary and time varying properties of oil prices but seldom focused on feature selection method for selecting significant inputs to improve forecasting accuracy. The study identifies that historical oil prices (either daily, weekly or yearly) are the most popular input variables used by researchers. Abdullah [26] has used 22 input variables from the categories as mentioned in Table 2.3 to achieve high prediction accuracy but the variables were selected on judge-

Table 2.3: Summary of factors influencing oil prices

Class	**Input Variable**	**Key**
Supply	OPEC Production	V1
	Non-OPEC Production	V2
	World Production	V3
	Oil Supply	V4
Demand	OECD Consumption	V5
	China Consumption	V6
	India Consumption	V7
	Seasonal Demand	V8
	Global Demand	V9
	Non-OECD Consumption	V10
Inventories	U.S. Refinery Capacity	V11
	OPEC Total Liquid Capacity	V12
	U.S. Gasoline Ending Stocks	V13
	OECD Stocks	V14
	U.S. Ending Stocks	V15
	U.S. Petroleum Imports from OPEC	V16
	U.S. Petroleum Imports from Non-OPEC	V17
	U.S. Crude Oil Imports from OPEC	V18
	U.S. Crude Oil Import from Non-OPEC	V19
	OECD Industrial Inventory Level	V20
	Crude Oil Distillation Capacity	V21
Price	Historical Prices	V22
	Heating Oil Spot Price	V23
	Gasoline Oil Spot Price	V24
	Natural Gas Spot Price	V25
	Propane Spot Price	V26
	NYMEX Crude Oil Futures	V27
	NYMEX Heating Oil Futures	V28
Reserves	OPEC Reserves	V29
	OECD Reserves	V30
	No. of Well Drilled	V31
Economy	GDP Growth Rate	V32
	U.S. Dollar Nominal Effective Exchange Rate	V33
	Foreign Exchange of GBP/USD	V34
	Foreign Exchange of YEN/USD	V35
	Foreign Exchange of Euro/USD	V36
	U.S. Inflation Rate	V37
	U.S. Consumer Rrice Index	V38
	Population of Developed Countries	V39
	Population of Less Developed Countries	V40
	S&P500	V41
	Gold Prices	V42
	Producer Price Index	V43
World Events	World Event Impact Factor	V44
	OPEC Quota Tighten (April 99)	V45
	Sept 11, 2001 Attack	V46
Properties	API Density	V47
	Sulphur Content	V48
	Country	V49
	Time	V50
	No. of Weeks	V51

mental criteria. To handle optimal long-term oil price forecasting, Azadeh et al. [73] developed a flexible algorithm based on artificial neural networks and fuzzy regression by using oil supply, crude oil distillation capacity, oil consumption of Non-OECD, USA refinery capacity and surplus capacity as economic indicators. The study concluded ANN models outperform FR models in terms of mean absolute percentage error (MAPE). Section 2.3 discusses in detail the characteristics of different types of artificial intelligent models that used factors listed in Table 2.3 as input variables for oil price forecasting.

2.3 Review of Methodology

For forecasting of oil prices, artificial intelligent methods are being extensively used as an alternate approach to conventional techniques. There has been a whole spectrum of artificial intelligent techniques to overcome the difficulties of complexity and irregularity in oil price series. The potential of AI as a design tool for oil price forecasting has been reviewed in this study. The following price forecasting techniques have been covered: (i) artificial neural network, (ii) support vector machine, (iii) wavelet, (iv) genetic algorithm, and (v) hybrid systems. In order to investigate the state of artificial intelligent models for oil price forecasting, thirty five research papers (published during 2001 to 2013) had been reviewed in form of table (for ease of comparison) based on the following parameters: (a) input variables, (b) input variables selection method, (c) data characteristics (d) forecasting accuracy and (e) model architecture.

This review reveals procedure of AI methods used in complex oil price related studies. The review further extended above overview into discussions regarding specific shortcomings that are associated with feature selection for designing input vector, and then concluded with future insight on improving the current state-of-the-art technology. In this section, the studies are segregated on the basis of type of models, starting from single models (such as neural networks, support vector regression, wavelets, genetically evolved models) to more complex and hybrid models. Recently, hybrid models are been extensively used for building oil price forecasting models as they overcome the limitations of single models and provide better forecasting accuracy.

2.3.1 Neural Network based Models

In the past , neural networks were being used extensively for oil price forecasting. Neural Networks can model richer dynamics and can approximate any continuous function of inputs [69]. In this category, seven researchers have forecasted oil prices using artificial neural networks as single model. There are many research studies that suggest integration of neural networks with other traditional methods (such as support vector regression, genetic algorithm, or wavelets) by mean of hybrid approach for improving the prediction performance. These studies are discussed in detail in section 2.3.5. In Table 2.4, information regarding the data, time scale, input variables for the study, method of input variable selection and preprocessing techniques employed are discussed. Researcher have utilized neural networks for all major oil markets. It is evident from the Table 2.4 that neural networks can handle large number of input variables. Abdullah and Zeng [26] integrated

Table 2.4: Data characteristics, preprocessing technique, input variables & its selection method

Paper	**Oil Market**	**Time Scale**	**Input Variable**	**Variable Selection**	**Preprocessing Technique**
[69]	WTI	D	V22, V27	JC	RMA, NORM
[33]	NF	D	V22	JC	-
[72]	Br	M	V11, V32, V33, V12, V1, V13	TE	CrI is formed
[22]	PGRP	M	V47, V48, V49, V50	SA	SR
[93]	SP	-	V22	JC	SR
[26]	WTI	M	V1, V2, V5, V6, V7, V14, V15, V16, V17, V18, V19, V29, V30, V31, V34, V35, V36, V32, V37, V38, V39, V40	MFA	LD
[82]	SP	D	V22, V23, V24, V25, V26	CA	CC

22 quantitative input variables (sub-factors of demand, supply, economy, inventory and population) together with the qualitative data (collected from experts' view and news) using neural networks to predict oil prices for long and short term time period. The authors have utilized manual feature extraction method for finding significant input variables for the study. Most of the studies have pre-processed the raw price data either by scaling range, normalization or cluster classification as shown in Table 2.4.

This review identified that most of studies have selected input variables based on judgemental criteria or trail and error basis. In Table 2.5, forecasting performance of various neural networks models has been compared. Neural networks showed superior results compared to benchmark TEI@I

Table 2.5: Forecasting performance comparison of Neural Network models

Paper	Training Data	Testing Data	Forecast Horizon	Comparison with other models	Level of accuracy
[69]	90%	10%	3 DA	-	RMSE: 0.53- 0.78 ; HR: 53 - 79
[33]	86%	14%	1 TPA	ARIMA, GARCH	MSE: 8.14; RMSE: 2.85; MAE: 2.04
[72]	86%	14%	1 MA	WCI - GRNN	MSE: -1.48 - 9.84
[22]	70%	30%	-	-	MSE: 7.24 - 8.82
[93]	-	-	5 DA	-	NMSE: -0.35; DS: 61; SNR: 25.37; AR: 92
[26]	80%	20%	-	TEI@I, EMD-FNN-ALNN	RMSE: 2.26;; NMSE: 0.009; DS: 94
[82]	80%	20%	1 MA	RM	MSE: -2.15 - 4.73

methodology and ARIMA-GARCH models as shown in Table 2.5. The models are validated using Root Mean Square Error (RMSE), Hit Rate (HR), Mean Square Error (MSE), Mean Absolute Error (MAE), Normalized Mean Square Error (NMSE), Annualized Return (AR) and Directional Statistics (DS). Malliaris and Malliaris [82] studied five inter-related energy products for forecasting one-month ahead prices using neural networks. The results thus obtained through neural network consistently led to a MSE less than half than that of the regression predictions. Malliaris and Malliaris [82] used correlation analysis to find significant input variables for their study. The model architecture of seven studies considered under this category is shown in Table 2.6. Mahdi et al. [93] examined three different neural network models: Multi-layered Perceptron (MLP), Functional Link Neural Network (FLNN) and Self-Organized MLP (SoMLP) for the oil price series. Mahdi et al. [93] compared the prediction capability of SoMLP with MLP and FLNN for ten different data sets including oil prices. The experimental results thus demonstrated that all neural networks performed better by using stationary data and failed to generate profits while using non-stationary data.

Haider et al. [69] presented a short term forecasting model to understand oil price dynamics based on multi-layer feed forward neural network. Several transformation methods were tested with original data and results showed that relative change of simple moving average is the best method amongst other methods. Moshiri and Foroutan [33] examined chaos in daily crude oil future prices using BDS and Lyapunov test. The results indicated that future prices follow complex non-linear dynamic process and

showed superiority of ANN model as compared to ARIMA and GARCH models. Movagharnejad et al. [22] designed a neural network to predict the prices of seven different crude oils in Persian Gulf region, provided that the benchmark light oil of Saudi Arabia is known or predicted by another hybrid forecasting method.

Further, this review identifies that number of hidden neuron varies across studies. Haider et al. [69] has fixed the number of neurons in hidden layer ranging from 1 - 10 while few authors [33] [22] [82] have fixed a constant value based on judgemental criteria for number of neurons in hidden layer. There is no rule of thumb applied by researchers for finding the optimal number of neuron to be set for hidden layer. Alizadeh and Mafinezhad [72] proposed a general regression neural network forecasting model for Brent crude oil price with particular attention on finding number of features as input data to achieve best performance. Most of the studies have used

Table 2.6: Neural Networks model's architecture

Paper	**NN type**	**Learning Algorithm**	**Hidden neurons**	**Activation Function**
[69]	MLP, RNN	LM	1-10	Sig
[33]	MLP	GD	5	HTS, Id
[72]	GRNN	-	TE	-
[22]	MLP	BP	15	LSig
[93]	MLP, FLNN, SoMLP	BP, IA	-	-
[26]	MLP	BP	TE	Sig
[82]	-	-	20	-

sigmoid function as preferable activation function as shown in Table 2.6. It is evident from the Table 2.6 that the multi-layered perceptron neural network with back propagation as the learning algorithm is the most popular among researchers for price forecasting.

2.3.2 Support Vector Regression Models

Oil prices are complex series with mixture of linear and non-linear characteristics underlying data generating processes of different nature. He at al. [64] introduced morphological component analysis to explore the complex nature underlying oil prices . There are few authors who have tested for non-linearity [63] [64] [50] and normality assumptions [63] [64] of oil price series. Support vector regression has an advantage of reducing the problem

of over-fitting or local minima. Khashman [83] experimental results proved SVM could be used with a high degree of precision in predicting oil prices. Bao et al. [49] presented a comparative study of recursive and direct strategies of multi-step ahead prediction for both WTI and Brent crude oil spot prices with support vector regression. As Compared to results obtained through benchmark ARMA and Random Walk models for crude oil price prediction, He et al. [63] confirmed the superiority of the proposed slantlet denoising algorithm based on SVR model. Zhu [84] formulated a two-stage structure for modelling oil future prices by partitioning the whole input data space into mutually exclusive regions by K-mean clustering algorithm and then corresponding SVM models. Table 2.7 showed that each study have preprocessed raw price data either by scaling range, return transformation or by cluster classifier. Most of the authors have used WTI as

Table 2.7: Data characteristics, preprocessing technique, input variables & its selection method

Paper	**Oil Market**	**Time Scale**	**Input Variable**	**Input Selection**	**Preprocessing**
[49]	WTI; Br	W	V22	JC	SR
[83]	WTI	W	V22, V27, V8, V51, V44, V9	JC	SR
[63]	WTI	D	V22	JC	RT
[84]	NF	D	V22, V33, V28, V41	JC	CC
[64]	WTI	D	V22	JC	WT; RT
[50]	WTI	M	V22	JC	RT

benchmark oil price data in their studies. An important research gap in selection of input variables through judgemental criteria or by literature review is highlighted under this category. It is evident from Table 2.8 that most of the authors have compared SVM with linear models. Xie et al. [50] has shown SVM model performed better than back-propagation neural networks. The model proposed by Zhu [84] has shown better performance in terms of MSE, MAE and MAPE compared to standard SVM model. Radial basis kernel function is the most popular choice among researchers for a price forecast problem as seen from Table 2.9. The values for epsilon, cost and gamma vary across different studies. He at al. used gradient search method to set appropriate model parameters [63] [64]. Xie et al. [50] and He et al. [64] have used directional statistics as performance criterion to compare their respective models with traditional stochastic or regression models.

Table 2.8: Forecasting performance comparison of Support Vector Regression models

Paper	Training Data	Testing Data	Forecast Horizon	Comparison with other models	Level of accuracy
[49]	71%	29%	4, 8, 12 WA	DirS, RecS	RMSE: 5.05 - 52.11; MAPE: 10.23
[83]	50%	50%	1 WA	-	PR: 81.27%
[63]	60%	40%	-	ARMA, RW	MSE: 4.65
[84]	95%	5%	-	S-SVM	MSE: 1.37; MAE: 0.95; MAPE: 1.15
[64]	60%	40%	-	RW, ARMA	MSE: 8.74; DS: 53.07%
[50]	88%	12%	1 MA	ARIMA, BPNN	RMSE: 2.19; DS: 70.83%

Table 2.9: Support Vector Regression model architecture

Paper	Kernel Function	Epsilon	Cost	Gamma	Model Parameters
[49]	RB	0.01	-	-	-
[83]	RB	-	2965820	0.001953	$2^{-15} - 2^{15}$
[63]	RB	6.64×10^{-16}	24.25	27.86	GSM
[84]	RB	-	-	-	[0,1]
[64]	RB	7.81×10^{-3}	$0.5 - 8$	3.90×10^{-3}	GSM
[50]	RB	-	-	-	-

2.3.3 Genetically Evolved Models

Kaboudan [85] performed short term monthly forecasting of crude oil price using genetic programming (GP) and neural networks. The study presented that GP can produce impressive one-month ahead forecast compared to that by Random Walk and ANN. This GP based oil price forecasting framework by Kaboudan [85] is considered as benchmark for comparison by Amin-Naseri [62]. Xiao et al. [94] combined transfer learning techniques

Table 2.10: Data characteristics, preprocessing technique, input variables & its selection method

Paper	Oil Market	Time Scale	Input Variable	Input Selection	Preprocessing
[85]	SP	M	V22, V3, V5, V15	TE	-
[94]	WTI; Br	D	-	TE	-

with analog complexing and genetic algorithm for crude oil price forecasting. However, there is no information as to how the input variables are selected in studies mentioned in Table 2.10. According to Table 2.11, Xiao et al. [94] showed that genetically evolved models performed better in comparison to neural networks and ARIMA family models based on MSE,

Table 2.11: Forecasting performance comparison of Genetically evolved models

Paper	Forecast Horizon	Comparison with other models	Level of accuracy
[85]	1-12 MA	ANN; RW	MSE: 0.24 − 1.85;
[94]	1 TPA	ARIMA; ANN; GPMGA; AC	RMSE: 1.0691; DS: 79.02

Table 2.12: Genetically evolved model's architecture

Paper	Cross-over probability	Mutation probability	Fitness function
[85]	0.02	0.06	–
[94]	0.9	0.05	–

RMSE and directional statistics. There is lack of information with respect to preprocessing technique and fitness function used by authors as shown in Table 2.12. The new combined models based on genetic algorithm with neural networks and that with SVM are discussed in section 2.3.5.

2.3.4 Wavelet-based Models

A wavelet-based prediction model is proposed to provide forecast over 1-4 months' horizon and to compare with future oil price data by Yousuf [95]. He et al. [15] introduced the wavelet decomposed ensemble model to analyse dynamic changing nature of underlying oil market structure. This study found that hybrid version comprising of wavelet with neural networks is more appealing to researchers as compared to single wavelet based model.

2.3.5 Hybrid Models

Genetic Algorithm and Neural Network

Amin-Naseri [62] proposed a hybrid artificial intelligence model combining local approximation techniques with genetically evolved neural network. The author used Hannan-Quinn info criterion (HQIC) as fitness function and set number of hidden neurons in range from 1-30. The performance of the model was evaluated with three competing frameworks (STEO, KAB and WANG) for oil price forecasting. The proposed model has performed well in terms on MSE, RMSE and directional statistics, and has found to be

effectively mapping the non-linearity and non-normality present in crude oil price data. The proposed model has been considered as benchmark model for comparison by Alexandridis and Livanis [86]. Fan et al. [55] presented Generalized Pattern Matching based on Genetic Algorithm (GPMGA) to predict future prices. GPMGA overcomes some limitations of Elman Networks and Pattern Modelling in Recognition System (PRMS) approach for multi-step prediction of oil prices. As evident from Table 2.13, authors

Table 2.13: Data characteristics, preprocessing technique, input variables & its selection method

Paper	**Oil Market**	**Time Scale**	**Input Variable**	**Input Selection**	**Preprocessing**
[62]	FP	M	V22	PACF	CC
[55]	WTI; Br	D	V22	GT; ACF	Standardization

have preprocessed raw price data to achieve high level of accuracy. Authors have preferred autocorrelation function and partial autocorrelation function to determine the optimal number of lags for model identification and estimation.

Table 2.14: Forecasting performance comparison of Genetic Algorithm and Neural Network models

Paper	**Training Data**	**Testing Data**	**Forecast Horizon**	**Comparison with other models**	**Level of accuracy**
[62]	95%	5%	-	STEO; KAB; WANG	MSE: 0.90 − 9.10; RMSE: 0.95 − 3.02; DS: 71 − 81%
[55]	99%	1%	1 MA	PRMS; ENN	RMSE: 1.57 −2.43

Table 2.15: Genetic Algorithm and Neural Networks model's architecture

Paper	**Model Type**	**Learning Algorithm**	**Hidden Neurons**	**Activation Function**	**Fitness Function**	**Cross-Over Probability**	**Mutation Probability**
[62]	MLP	LM; GD	1 − 30	LSig	HQIC	0.9	0.01
[55]	RNN	BP	TE	TSig	-	0.9	0.09

The review states that genetically evolved neural networks are superior to other competitive models as mentioned in Table 2.14. The studies under this category have utilize sigmoid function as activation function. The number of neurons in hidden layers varies as seen from Table 2.15.

Wavelet and Neural Network

Mingming et al. [52] proposed a multiple wavelet recurrent neural network based hybrid method for international crude oil prices. This model utilized wavelet analysis to capture multi scale data characteristics, while designing an appropriate recurrent neural network to predict oil prices at different time scales, followed by a standard back-propagation neural network to combine these independent forecasts. He et al. [65] proposed an ensemble approach incorporating wavelet and feed-forward neural network for estimating VaR in crude oil market to further improve modelling accuracy and reliability of three oil markets: WTI, Brent and Dubai. Jinliang et al. [25] decomposed crude oil price time series into several trend and random component. For higher prediction accuracy, the trend component of oil prices is predicted with Boltzmann neural network and the random component is predicted with Gaussian kernel density function as shown in Table 2.18. Jammazi and Aloui [53] examined a short term forecasting of monthly WTI prices with different input-hidden nodes combinations and three types of activation functions. The results highlighted combination of Harr A Trous wavelet function with back propagation neural network as a promising forecasting tool. Pang et al. [70] proposed to predict monthly oil prices using OECD inventory level as independent variable, and used wavelet theory based feed forward neural network to model the non-linear relationship between oil prices and inventory. The proposed model achieved lower RMSE, MAPE and MAE in comparison to both linear and non-linear relative inventory models.

Qunli et al. [47] decomposed the original price sequence successfully using discrete wavelet transform as input layer of radial basis function neural network. Alexandridis and Livanis [86]used wavelet neural network to forecast monthly WTI crude oil spot prices using price lags, world crude oil production and the producer price index for petroleum as explanatory variables. Out of seven articles listed in Table 2.16, only one author has emphasized on selecting input variables based on correlation analysis. Correlation analysis is a measure of linear relationship between variables but macroeconomic variables exhibit non-linear relationship with oil prices. Therefore, there is a requirement to develop a method for identifying significant input indicators based on non-linear relationship that exists between variables. It can be observed that Daubechies has been adopted by most

Table 2.16: Data characteristics, preprocessing technique, input variables & its selection method

Paper	Oil Market	Time Scale	Input Variable	Input Selection	Preprocessing	Wavelet Function
[52]	WTI, Br	A	V22, V42	JC	WT	Db
[65]	WTI, Br, Du	W	V22	JC	RT, WT	HaT; Db; Coiflet
[25]	WTI	M	V22, V42	JC	WT	Db
[53]	WTI	M	V22	JC	WT	HaT
[70]	WTI	M	V22, V20, V45, V46	JC	WT	Morlet
[47]	Br	M	V22	JC	WT, SR	Db
[86]	WTI	M	V22, V43, V3	CA	WT	-

Table 2.17: Forecasting performance comparison of Wavelet and Neural Network models

Paper	Training Data	Testing Data	Forecast Horizon	Comparison with other models	Level of accuracy
[52]	70%	30%	4-8; 8-16; 16-32 YA	-	MSE: 3.88 - 4.06
[65]	36%	24%	-	ARMA-GARCH	MSE: 0.0059 - 0.0131
[25]	-	-	-	-	-
[53]	80%	20%	19 MA	MLP	MSE: 3.89; HR: 73%; R^2: 0.997
[70]	82%	18%	1 MA, 2MA, 3MA	L-RIM, NL-RIM	RMSE:1.486; MAE:1.073; MAPE:2.263
[47]	68%	32%	-	-	SSE: 6.16×10^{-5}
[86]	56%	50%	1MA, 3MA, 6MA	WANG, AMIN, STEO	MSE: 2.05; MAE: 1.02; Max AE: 7.36

Table 2.18: Wavelet and Neural networks model's architecture

Paper	Model Type	Learning Algorithm	Hidden Neuron	Activation Function
[52]	RNN; MLP	BP	TE	LSig
[65]	MLP	LM	6	LSig
[25]	BNN	-	-	-
[53]	MLP	BP	TE	BiP Sig
[70]	MLP	GD	8	-
[47]	RBF	-	4	-
[86]	WNN	-	1	-

researchers as a wavelet function. The prediction accuracy of combined models is evaluated with linear ARMA family, non-linear neural networks, STEO, WANG and AMIN models.

Fuzzy Neural Network

Panella et al. [66] favoured the quality of forecasting accuracy based on neurofuzzy approach (adaptive neuro-fuzzy inference system) in comparison to other linear and neural network models. Ghaffari and Zare [56] presented a method based on soft computing approaches to forecast WTI crude oil spot prices for one-month ahead forecast horizon. Azadeh et al. [73] used oil supply, surplus capacity, Non-OECD consumption, U.S. refinery capacity and crude oil distillation capacity as input variable for designing a flexible ANN-FR algorithm to model noisy and complex oil prices. Further, the ANOVA and Duncan multiple range test are used to test the significance of the forecast obtained from ANN and FR models. Table 2.19 indicates that the input variables has been selected based on judgemental criterion and no prior selection method has been applied. Ghaffari [56] explored the pos-

Table 2.19: Data characteristics, preprocessing technique, input variables & its selection method

Paper	Oil Market	Time Scale	Input Variable	Input Selection	Preprocessing
[66]	WTI; Br	D	V22	JC	LT; CC
[56]	WTI	D	V22	JC	SMP
[73]	WTI	A	V11, V4, V21, V10	JC	-

sibility of smoothing procedure as preprocessing tool to explore the pattern of oil prices while Panella et al [66] incorporated both log transformation and cluster classification. As clear from Table 2.20, the prediction accuracy of proposed neuro-fuzzy model by Panella et al. [66] has been compared with linear and non-linear models on the basis of noise-to-signal ratios.

Ghaffari [56] has compared the results of their respective models with and without smoothing procedure and observed that prediction quality in terms on percentage of correct predictions has been improved by smoothing oil price data. It is evident from Table 2.21 that researchers preferred to develop fuzzy model by adopting Takagi-Sugano-Kang as fuzzy inference system and Gaussian as membership function. The multi-layered perceptron

neural network is the most popular among researchers for oil price forecasting under this category of neuro-fuzzy approach. Azadeh [73] model architecture includes MLP along with five variant of learning algorithm to improve the forecasting performance and achieved MAPE as low as 0.035.

Table 2.20: Forecasting performance comparison of Fuzzy Neural Network models

Paper	Training Data	Testing Data	Forecast Horizon	Comparison with other models	Level of accuracy
[66]	67%	33%	1 SA	LSE; RBF; MoGNN	NSR: -46 − -24
[56]	80%	20%	30 DA	WSP	PCP: 68.18 − 70.09
[73]	80%	20%	-	ANN, FR	MAPE: 0.035

Table 2.21: Fuzzy Neural Networks model's architecture

Paper	Model Type	Learning Algorithm	Hidden Neurons	Fuzzy Inference System	Membership Function
[66]	MLP	LSE + BP	TE	TSK	Gaussian
[56]	MLP	LSE + GD	TE	TSK	Gaussian
[73]	MLP	BFGS; BR; BLR; GDX; LM	TE	-	Gaussian

Decomposition based Neural Network

Yu et al. [51] proposed a "decomposition-and-ensemble"strategy using EMD-based NN ensemble learning model to predict oil prices. Empirical mode decomposition is proposed to decompose oil price data into eleven intrinsic mode functions. Xiong et al. [46] evaluated the performance of EMD-based feed-forward neural network framework incorporating slope-based method for oil price forecasting with three leading strategies : direct, iterative and multiple-input multiple-output (MIMO).

Xiong et al. [46] used Symmetric MAPE (SMAPE) as a forecasting performance criterion to evaluate the performance of EMD based neural network. In Table 2.22, partial mutual information is used to examine the relationship between historic prices and oil prices together with forward backward

Table 2.22: Data characteristics, preprocessing technique, input variables & its selection method

Paper	Oil Market	Time Scale	Input Variable	Input Selection	Preprocessing
[51]	WTI; Br	D	V22	JC	–
[46]	WTI	W	V22	PMI; DT; FBS	SR

selection and delta test for model identification and estimation. Multi-layered perceptron neural network is the most widely used neural network architecture by researchers for hybrid models as evident from Table 2.24.

Table 2.23: Forecasting performance comparison of Decomposition based Neural Network models

Paper	Training Set	Testing Set	Forecast Horizon	Comparison with other models	Level of accuracy
[51]	72%	28%	1 DA	ARIMA; MLP	RMSE: 0.273 −0.225; DS: 86.99−87.81
[46]	67%	33%	4 WA	MLP	SMAPE: 2.28−8.15; MAE: 0.81−1.28; DS:54−87

Table 2.24: Decomposition based Neural Networks model's architecture

Paper	Model Type	Learning Algorithm	Hidden Neurons	Activation Function	Decomposition Method	No. of IMF's
[51]	MLP; ALNN	BP	–	Lgs; Lin	EMD	11
[46]	MLP	LM	15	–	EMD	–

Support Vector and Genetic Algorithm

Guo [96] improved traditional SVR forecast precision by using genetic algorithm optimized parameter of SVR in accordance with the training data. The model is found to be effective in mapping the complexities of oil price series. Gabralla [97] investigated performance of two different algorithms for feature selection together with several machine learning methods (IBL, KStar and SMOreg) for oil price prediction.

2.4 Inference drawn from the Literature Review

- There is no solitary indicator driving crude oil prices. The output is based on how much information is contained in the set of input variables selected for the study.
- In most of the studies, the design of input vector for oil price forecasting model is carried out on judgemental criteria or trial and error procedures. Little attention is paid on selecting influential factors and more on assessing new techniques for oil price forecasting.
- The effect of input variables is considered to constantly driving oil prices in different studies. There is shift in the influence of input variables subject to happening of any geopolitical and economic events in a given time-period but there is no literature available that highlights this point. Predicted oil prices are dependent on short term macroeconomic indicators whose effects are subject to structural changes.
- Artificial neural network is most widely method used by researchers for price forecasting. Recently, researchers suggests integration of neural networks with traditional methods as support vector regression, genetic algorithms or wavelets by mean of hybrid approach to overcome limitations of single models.

2.5 Research Gap

- As per researcher knowledge, there is no literature which has done a comprehensive literature review of artificial intelligent based oil price forecasting models.
- Few studies have been carried out using artificial intelligent models to forecast complex oil price series as compared to its application in diverse fields.
- There are many studies that had examined the relationship between oil prices and macroeconomic variables but there seem no consensus on the extent to which these macroeconomic variables drive oil prices.
- Existing methods of predicting oil prices have accounted for non-linearity, non-stationary and time-varying structure of the oil prices

but have seldom focus on selecting significant features with high predicting power. The empirical literature is very far from any consensus about selecting the appropriate features/ indicators that explains the characteristics of oil market.

- There is a lack of robust feature selection method for designing the input vector of oil price forecasting model that depends on the association and dependency structure of oil prices and exogenous variables.

- Till date, there is no literature which has paid attention on finding significant indicators driving oil prices constraint to structural change using any competent feature selection method. There is a need to select appropriate features/ indicators that explains the characteristics of oil markets subject to major structural changes so as to improve forecasting accuracy.

2.6 Objective of the Study

The following are the objectives of the research work:

To study the association and dependency structure between oil prices and strategic indicators that drive them.

- Data Mining the drivers of oil prices by developing a two-stage competent feature selection method and using artificial intelligent models as forecasting engines.

To study the non-linear dependence between oil prices and strategic indicators by employing interaction information and mutual information as measure of redundancy(or synergy) and relevance.

- To develop a three-stage competent feature selection method for finding the minimal set of key factors that drive oil prices.

To develop and empirically test artificial intelligent based forecasting models for short term period based on significant factors driving oil prices.

- To find significant factors driving oil prices constraint to happening to geopolitical and economic events.

2.7 Research Questions

Based on the research objectives mentioned above, the following research questions have been identified that needs to be answered through this research work.

Central Research Question (RQ): What are the factors that can contribute to change in direction of oil prices?

- **Additional RQ1**
  Is the proposed two-stage feature selection method competent to identify drivers of oil prices together with artificial intelligent models as forecasting engines?
- **Additional RQ2**
  Is the proposed three-stage feature selection method competent to identify drivers of oil prices together with artificial intelligent models as forecasting engines?
- **Additional RQ3**
  What are the key indicator driving crude oil prices in pre and post-2008 financial crisis scenario to develop artificial intelligent models for short-term forecasting?

2.8 Scope of the Study

The scope of the study is restricted to build an artificial intelligent model using proposed feature selection methods for oil price forecasting. The scope of the study is restricted to broad spectrum of input factors (covering supply, demand, inventory, reserves, economy, weather, exchange market, speculation and stock market). The study focussed on developing competent feature selection for identifying key drivers of oil prices from any set of input variables. Due to lack of availability of data in a desired way, the scope of the study confined to input variables as listed in Chapter 3.

2.9 Concluding Remark

Off late, artificial intelligent models are being extensively used to capture unknown or too complex structures in time series. This chapter focussed

on artificial intelligent based oil price forecasting models and attempted to provide in-depth review based on the following parameters: (i) type of model, (ii) input variables, (ii) input variable selection method, (iv) data characteristics, (v) forecasting performance, and (vi) model architecture. It enlisted the numerous key indicators used as input variables in artificial intelligent based oil price forecasting models and attempted to highlight a serious issue of selecting input variables based on judgemental or trial and error basis.

This chapter concludes that there is no single indicators driving oil prices and there is need to identify the relevant input variables for oil price predictions. Multi-layered perceptron neural network is most widely used by researchers for price forecasting. Recently, researchers suggests integration of neural networks with traditional methods as support vector regression, genetic algorithms or wavelets by mean of hybrid approach to overcome limitations of single models. It also highlighted that effect of factors is constraint to happening of geopolitical and economic events. The research gap highlighted to develop a robust feature selection method that can account for non-linearity and time-varying structure of oil prices.

Chapter 3

Methodology - Concepts & Definitions

3.1 Overview

This chapter provides a comprehensive understanding of the procedure of data mining process followed in the study. This chapter discusses the basic concepts of information-theoretic approaches that include mutual information, conditional mutual information and interaction information. It is further followed by description of artificial intelligent methods used as forecasting engines for the study. It provides an overview of two proposed algorithms: MI^3 and I^2MI^2 for feature selection to achieve high prediction performance for oil prices. This chapter further describes the characteristics of datasets used for the study.

3.2 Data Mining Process

We are in an age that is often referred as the information age. In this age, basic understanding of how to plan, execute, refine and build model to discover knowledge is an important part of any business project. There is colossal collection of data, ranging from simple measurements to more complex information such as wind speed, text documents and gene regulatory patterns. Data mining is a process of discovering knowledge by identifying novel, non-trivial and useful patterns or information in existing data. Data mining is an extension of traditional statistical data analysis as it includes analytical techniques drawn from a range of disciplines, but not limited to, artificial intelligence, econometrics, numerical analysis and information

technology. The process of data mining comprises of few steps starting from defining a problem, raw data collection to extracting meaningful new information and patterns for knowledge enhancement. Data mining can offer a broad spectrum of industries such as:

- Data mining can help government organizations to focus on proposing new wind farm ventures based on wind speed, wind density and other significant variables database.
- Data mining can help in determining the target customers who are likely to respond to recently launched product.
- Data mining can help in discovering and exploring regularities in gene regulator systems.
- Data mining can determine market characteristics based on historical database as well as to predict stock performance.

The iterative data mining process is illustrated in the following steps:

- **Defining Business Problem**
  The first step is to define the problem, determine the project goals and requirements from a business perspective, identify key studies and learn about the current solutions to the problem. This step involves formulating the gained knowledge into a data mining problem definition. To achieve the business objectives as stated in chapter 1, the data mining goal for the study is "Mining the key factors driving oil prices using robust feature selection algorithms for achieving high prediction performance".
- **Data Gathering**
  This step starts with initial screening of data required for the project. In this step, sample data is collected and background knowledge is used as guide to decide the data subset of interest for the project. An extensive literature review is performed to get familiar with data, identify data quality, discover first insight into data and form subset of input indicators. Data is collected from EIA, CFTC, Bloomberg and World Bank databases and briefly described in section 3.7.
- **Data Preparation**
  This phase covers all activities covering data cleaning, data integration, data transformation and data sampling to construct the final

dataset for modelling purpose. In this step, data is checked for completeness, redundancy, missing values, plausibility of attributes, etc. Data cleaning removes noise and irrelevant data from the collection. This cleaned data may be further processed by feature selection algorithms by deriving new attributes using discretization. Under data integration, data is collected from multiple heterogeneous sources and integrated in a combined source. Data transformation is a phase in which the selected data is transformed into an appropriate form for mining procedure. Data sampling consist of dividing data into training, testing and validation subsets for model building.

In this research, the selected data is checked for missing values and outliers as part of data cleaning process. Linear transformation is applied to data for synchronising diverse measurement scales. An important part of data mining process is feature selection. Feature selection helps to decide the set of relevant and non-redundant features for the study. An appropriate set of features can help in high prediction performance and thus, due care should be taken to select an appropriate set of relevant and non-redundant features. It is evident from Chapter 2 that most of the studies have selected input variables based on judgemental criterion or trial and error method. This study proposes a new two-stage MI^3 algorithm and its extended version I^2MI^2 algorithm for selecting the set of input variables that perform best for forecasting oil prices.

- **Model Building & Evaluation**
  It is the most crucial step in which various modelling techniques for classification and regression are selected and their parameters are optimized to extract patterns. The discovered knowledge is extracted and evaluated by understanding the results. The most important goal of model building is to provide a stable model for prediction that holds true when applied to unseen data. This study utilizes various neural network models (multi-layered perceptron, cascaded neural network and general regression neural network) as forecasting engines. The models are evaluated for one and twelve-month ahead forecasts with EIA's STEO econometric model forecasts.

- **Knowledge Deployment**
  The goal of this step is to organize and present the knowledge gained

in report format or the way client can use. This phase requires simple reports to be generated or implementing a repeatable data mining process. The conclusions based on the analysis performed in this research are shown in chapter 6.

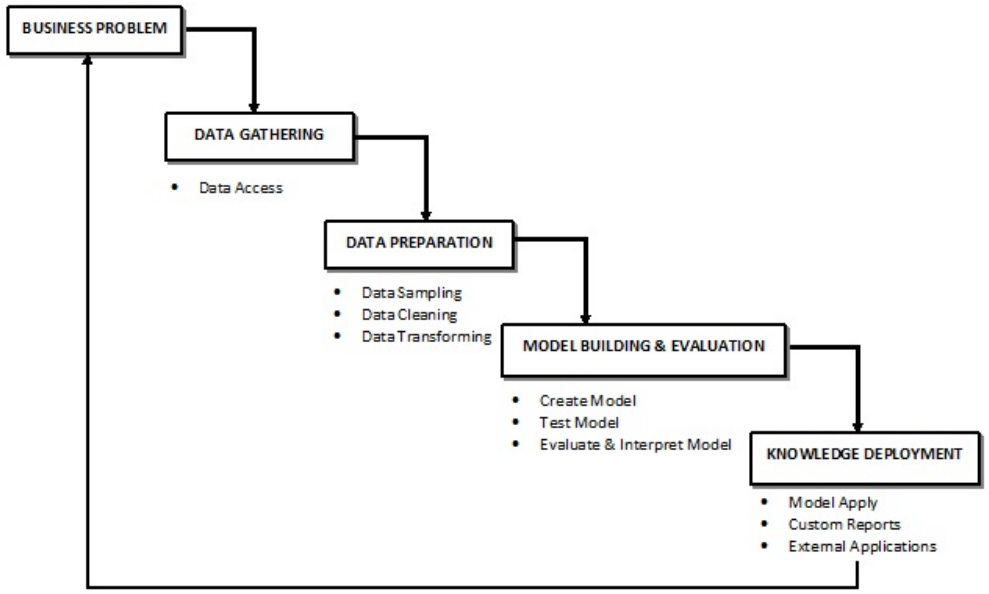


Figure 3.1: Flowchart of data mining process

3.3 Basic Terminology

3.3.1 Mutual Information

Information theory provides quantifiable tools to measure the amount of information within a random variable or between n-random variables. A fundamental concept in information theory is the entropy $H(X)$ [98] of random variable (WTI prices), that provides self-information about a random variable. It quantifies the amount of uncertainty of a random variable (WTI prices).

Definition 1: The entropy of a continuous variable X with probability distribution $p(X)$ is defined as

$$H(X) = -\int p(X)\, \log p(X)\, dX \tag{3.1}$$

where the unit of measurement is the bit, when logarithm to the base 2 are used. If X is a discrete random variable, the entropy is defined as follows:

$$H(X) = -\sum_{x \in X} p(x) log\, p(x) \tag{3.2}$$

The joint entropy $H(X, Y)$ and conditional entropy $H(Y \mid X)$ are extensions that measure the uncertainty in the joint distribution of a pair of random variables and the uncertainty in the conditional distribution of a pair of random variables.

Definition 2: The joint entropy $H(X, Y)$ of two random variables with probability distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) log\, p(x, y) \tag{3.3}$$

Definition 3: The conditional entropy H(Y|X) is defined as

$$H(Y|X) = -\sum_{x \in X} p(x) H(Y|X = x) \tag{3.4}$$

The Mutual Information (MI) measures how much on average one random variable tells us about another random variable. It is a measure of how much entropy of Y (WTI prices) is reduced if one is aware of X where X can be OPEC supply, Reserve-Production ratio or China consumption.

Definition 4: The mutual information is defined as:

$$I(X, Y) = -\sum_{x \in X, y \in Y} p(x, y) log\, \frac{p(x, y)}{p(x)p(y)} \tag{3.5}$$

Alternatively, mutual information can be expressed in terms of entropy and conditional entropy as follows:

$$I(X, Y) = H(X) - H(X|Y) \tag{3.6}$$

or

$$I(X, Y) = H(X) - H(X|Y) - H(Y|X) \tag{3.7}$$

A high value of mutual information indicates close relationship between variables Y and X. However, a higher value of mutual information between

input variables indicates redundancy.

3.3.2 Interaction Information

The interaction information [99] or co-information [100] is the generalization of mutual information that measures the amount of information (redundancy or synergy) bound up in a set of n variables. For the three variables case $\{X, Y, Z\}$, it measures whether the dependency between X and Y is entirely due to the influence of common cause Z or not.

Definition 5: The interaction information $I(X, Y, Z)$ is defined as:

$$I(X, Y, Z) = I(X, Y|Z) - I(X, Y) \tag{3.8}$$

where $I(X, Y)$ is the mutual information between X and Y, and $I(X, Y \mid Z)$ is the conditional mutual information between X and Y given Z. Unlike mutual information, interaction information can be either positive or negative. If the interaction information is negative, it indicates Z constrains the information shared between X and Y; emphasizing the set of redundant variables. Further, if the interaction information is positive, it indicates Z boosts the correlation between X and Y; emphasizing the synergy in set of variables.

3.4 MI^3 Algorithm for Feature Selection

Based on the information-theoretic approaches discussed in previous section, an original two-stage non-linear feature selection method called MI^3 Algorithm composed of interaction information and mutual information is proposed for finding relevant drivers of oil prices. The proposed algorithm consists of two stages. In the first stage, mutual information based irrelevance filter is proposed to select the most relevant features from the set of candidate inputs. The selected features from stage one are filtered based on a threshold value. The filtered features are sent to stage two.

In the second stage, interaction information based redundancy filter is proposed to remove the redundant features from selected relevant candidates. The set of features having negative interaction information are used to filter redundant features. The selected features filtered from stage one & two

is a set of relevant and non-redundant features. These selected features are used as set of input variables to build neural networks for predicting oil prices. The ensemble model can provide an insight into the explanatory power of selected relevant and non-redundant features and their contribution in deriving the direction of oil prices. The structure of the proposed MI^3 Algorithm is shown in Fig 3.2.

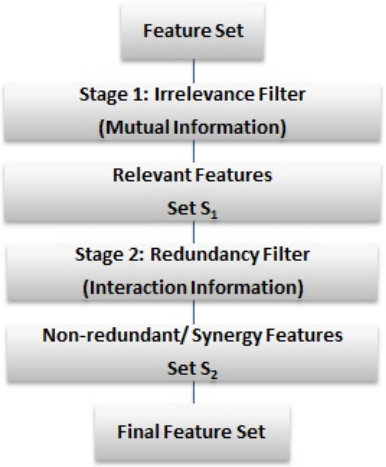


Figure 3.2: Flowchart of the proposed MI^3 algorithm

3.5 I^2MI^2 Algorithm for Feature Selection

The novel three stage feature selection method called I^2MI^2 algorithm is an extended version of MI^3 Algorithm build on pillars of interaction information and mutual information. It is used for selecting relevant and non-redundant features that drive oil price. The proposed algorithm consists of three stages. In the first stage, mutual information is computed between target variable and candidate inputs. The variables are ranked based on normalized mutual information value and the irrelevant features are filtered out based on a threshold value. The selected variables are added to set S_1.

In stage two, three-variable interaction information is computed to removes the redundant features among the selected features from set S_1. The set of selected features having negative interaction information are used to filter out the redundant features. The study incorporates the concept of interaction information so as to filter redundant input variables instead of correlation analysis or partial correlation analysis. Interaction information

is favoured over correlation analysis as it measures non-linear dependency. This stage provides list of features that are relevant and non-redundant in nature.

Further, in the third stage, mutual information is computed between the selected features from stage two and ranked according to normalized mutual information value. Depending on a threshold value, redundant features in stage three are filtered according to relevance rank in stage one. The selected features are used to build neural networks for oil price prediction. R-codes are written for each stage of the proposed algorithms together with manual computation required for stage two. The study further used DTREG software for modelling purposes. The structure of the proposed I^2MI^2 algorithm is shown in Fig 3.3.

Figure 3.3: Flowchart of the proposed I^2MI^2 algorithm

3.6 Forecasting Engines

3.6.1 Neural Networks For Oil Price Modelling

Neural Networks, also called artificial neural networks, are data driven non-parametric methods for classification and prediction that do not require many constricting assumptions related to data generation. The neural network is based on the biological activity in the brain, having trivial interconnected units called neurons. Neural networks are data motivated procedures that learn from the sample data [101] and hence doesn't require any underlying probabilistic assumption of prices in terms of normality, leptokurticy, non-linearity and autocorrelation [12]. Artificial neural networks perform well with non-stationary data also [102]. Neural Networks models perform well even when the market fluctuated heavily [70] and can approximate any continuous function [69]. Among many non-linear models, neural network has been one of the popular methods in financial studies for forecasting applications. The main characteristics of neural networks can be highlighted within the environment of non-linear time series models. Lets consider the linear $AR(1)$ time series model:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \vartheta_t \tag{3.9}$$

where Y_t is a stochastic process, α_0 are parameter and ϑ_t is white noise. Neural networks are extensions of linear autoregressive models that allow the relationship between Y_t and Y_{t-1} to be non-linear by augmenting Eqn. 3.10 by a transfer function (logistic, hyperbolic-tangent, linear, log-sigmoid etc.):

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \phi_1 F_t + \nu_t \tag{3.10}$$

where ν_t is an error term and F_t is given by:

$$F_t = \frac{1}{1 + \exp^{-(\alpha_0 + \alpha_1 Y_{t-1})}} \tag{3.11}$$

The neural network parameters $\{\alpha_0, \alpha_1, \phi_1, \beta_0, \beta_1\}$ can be estimated by choosing the parameters β_0 and β_1 randomly for first step, thereby, regressing Y_t on $\{1, Y_{t-1}, F_t\}$ and estimating the remaining parameters $\{\alpha_0, \alpha_1, \phi_1\}$. Neural networks consist of three layers of architecture: input, hidden and output layer. Input layer consists of all input variables, output layer consists of target variables and hidden layer refers to transfer function (i.e.

logistic function as mentioned in Eqn. 3.11) which forms the non-linear relationship between target and input variables. Neural networks methodology is examined under the following categories:

- **Data Preprocessing**: It helps to minimize the effect of magnitude among the inputs by normalization of input variables on a specified scale (e.g. between 0 and 1 or -1 and 1). It further includes data cleaning to remove noise, resolves missing values and emphasizes correlation in the input variables.

- **Network Structure**: The structure of neural network depends on number of neurons in input, hidden and output layer and the type of transfer function used to control the output of neurons.

- **Avoiding Over fitting**: The problem of over fitting can be resolved by using a network size (i.e. number of neurons in hidden layer) just large enough to provide a suitable fit.

- **Network Training**: The data is trained to estimate the weights that lead to best predictive results.

Neural networks can model richer dynamics and can accurately approximate any continuous function of inputs [69]. Moshiri and Foroutan [33] highlighted the existence of chaos in movement of oil prices and tested for chaos using neural networks. Stengos [103] showed that non-parametric models are better and provide superior forecast as compares to parametric models, when the economic series shows existence of chaos. Qunli et al. [47] showed superior performance of radial basis function neural network model for complex brent spot price data.

This study incorporated three forecasting engines: Multi-layered Perceptron, Cascaded neural networks and General Regression neural networks for forecasting oil prices based on features selected through proposed I^2MI^2 and MI^3 algorithm. Multi-layered perceptron neural network consist of input layer, hidden layer and output layer. Initially, data is divided into three sets: training, validation and testing. The goal of training process is to find the set of neuron's weight that results in lowest RMSE, MAE or MAPE between target prices and predicted prices. To find the optimal number of neurons in hidden layers, process of cross-validation can be incorporated. Cascaded neural networks begin with a minimal network

consisting of input layer and output layer with no hidden layer [104]. It trains and adds neurons in hidden layer one by one, creating a cascaded structure. It is suitable for problems where few of input factors are required to be predicted first [105]. AlFuhaid et al. [106] and Kouhi & Keynia [107] showed that forecasting approach of cascaded neural network is more effective for short term load forecasting as compared to that of artificial neural networks. Cascaded neural networks overcome the shortcoming of back-propagation neural networks and provide faster convergence speed and low error rate [108]. The V-fold cross validation is performed to build a CNN for model testing and validation. The third forecasting engine: General Regression neural networks [109] are much faster to train and are often more accurate than multi-layered perceptron [110]. General regression neural network has proved to be effective for complicated non-linear financial problems [111]. Zhang et al. [112] used general regression neural network to model change in length of day and showed the effectiveness and feasibility of GRNN over MLP for complex problems.

Artificial neural networks have been shown to exhibit superior prediction performance compare to traditional approaches but still they have been labelled as "black box"because in early 90s they were not able to provide an insight on the explanatory power of input variables. In recent years, various data mining software's (SAS, Statistica, DTREG and WEKA) provide the details regarding explanatory power of predictor variable in predicting oil prices and have overcome this limitation of neural networks.

In this thesis, there are three neural network architectures that have been used as forecasting engines. The issues involved in designing and training each one of them for the current research are discussed below. The measures taken to overcome issues involved in designing and training multi-layered perceptron neural networks are as follows:

- **Data Preprocessing**
  It helps to minimize the effect of magnitude among the inputs by normalization input variables on a specified scale (e.g. between 0 and 1 or -1 and 1). It further includes data cleaning to remove noise, resolve missing values and emphasize correlation in the input variables. The input layer for multi-layered perceptron neural network is standardized so that range of each variable is from -1 to 1.

- **Deciding number of neurons in each hidden layer**
  It is one of the most important characteristic of MLP network. There is no rule of thumb for fixing the number of neuron in hidden layer. The network may over fit the data if there are too many neurons being used. When over fitting occurs, the network begins to model random noise in the data that leads to extremely well model fit to training data, but poor to new data. The problem of over fitting has been taken care by using a network size (i.e. number of neurons in hidden layer) just large enough to provide a suitable fit. To find the optimal number of neurons in hidden layer, the numbers of neurons ranging from 1 (minimum) to 20 (maximum) are tested. Numerous models were built by using varying number of neurons and cross-validation is used to measure the quality.

- **Convergence to optimal solution**
  Given a set of randomly selected starting weight values, there is need to optimize these values. An appropriate procedure to adjust the weights towards convergence is required. The scaled conjugate gradient algorithm is used in the study to overcome the instability by combining the model-trust region approach from the Levenberg-Marquardt algorithm with the conjugate gradient approach. This algorithm developed by Martin Fodslette Moller converge twice as fast as traditional conjugate gradient and up to 20 times as fast as back propagation using gradient descent.

The measures taken to overcome issues involved in designing and training general regression neural networks are as follows:

- **Data Preprocessing**
  The ranges of values in input layer are standardize by subtracting the median and dividing by the interquartile range.

- **Optimal Sigma Value**
  The primary work of GRNN is to optimize sigma value to control the spread of radial basis function. The study uses conjugate gradient algorithm to compute the optimal sigma value. One of the limitations of GRNN is that there is one neuron for each training row which leads to large number of neurons in the network. To overcome this limitation, unnecessary neurons are removed using cross validation.

The measure to overcome issues involved in designing and training cascaded neural networks are as follows:

- **Over fitting protection control**
  The process of cross validation and pruning model to optimize size is performed to avoid over fitting.

There are several issues which are common for all above mentioned neural networks. They are discussed as follows:

- **Division of data**
  Cross validation is a technique that is frequently used in ANN modelling and has a significant impact on the way the available data is divided. Cross validation is used to compare the generalization ability of different models. For Group A, nearly 17 year data (January 1994-December 2011) consisting of 216 monthly data points is used for training and validation, whereas 12 month ahead data (January 2012-December 2012) is used for testing purpose. For Group B, nearly 17 year data (January 1995-November 2012) consisting of 215 monthly data points is used for training and validation, whereas 12 month ahead data (December 2012-November 2013) is used for testing purpose.

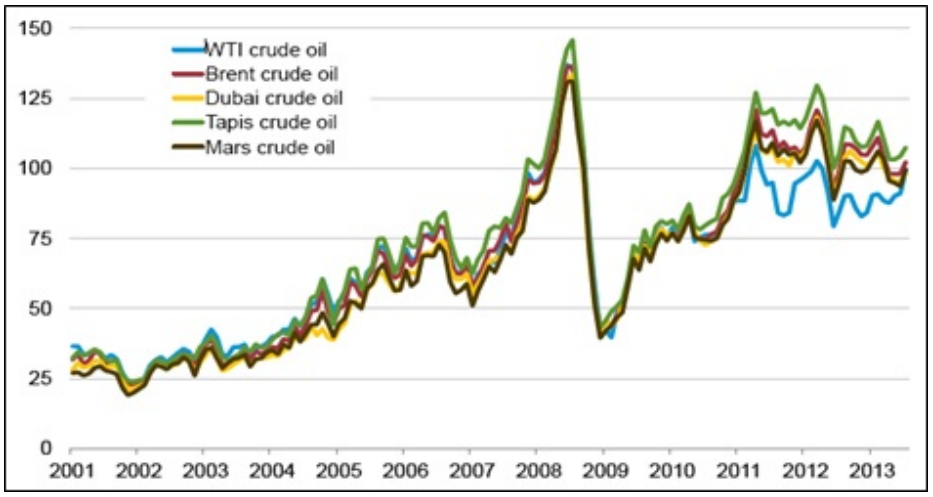
- **Determination of model Inputs**
  The selection of appropriate inputs is an important task. However, it is evident from chapter-2 that little attention is given to this task. Most of the studies have selected input variables based on judgemental criterion or trial and error basis. To focus on this important issue, this thesis propose two new feature selection methods: MI^3 and I^2MI^2 built on pillars of mutual information and interaction information concepts. The methods are used to select the set of relevant and non-redundant features for achieving high oil price prediction performance.

- **Transfer Function / Kernel Function**
  Sigmoid function is used as transfer function for multi-layered perceptron neural networks. Radial basis function is used as kernel function in general regression neural networks to compute weights (influence) of each input whereas Sigmoid and Gaussian function is used as hidden layer kernel functions in cascaded neural networks to optimize weights.

- **Validating the neural network to test for over-fitting**
  V-fold cross-validation is performed for model testing and validation.

3.7 Data Sample and Preparation

According to Energy Information Administration (EIA), world oil prices (i.e. WTI, Brent, Mars, Tapis and Dubai) move together due to arbitrage as shown in Fig 3.4. Though there are obstinate differences between types of oil based on sulphur content, temperature and pressure qualities but prices of oil produced globally tends to move together. Globalization hypothesis



Source: Bloomberg, Thomson Reuters

Figure 3.4: World oil prices move together according to globalization hypothesis

of oil prices moving together holds independently of whether the market is crashing or booming [68]. In this study, WTI crude oil spot price is chosen as target variable to select significant drivers as it is considered as a benchmark oil in global terms [113] [114] [115]. There are two sets of input variables used in this thesis. The description of two groups is given below.

3.7.1 Group A

The factors driving oil prices are classified into eight classes: Supply (1–2), Demand (3–7), Reserves (8–13), Inventory (14–17), Future Market (18), Exchange Market (19–21), Economy (22–26) and Weather (27–28) as shown in Table 3.1. The features are selected on basis of extensive literature review as discussed in Chapter 2. According to data availability, 28 representative index were chosen within eight classes. The data for supply, demand,

reserve, inventory, future prices, weather and economy is collected from Energy Information Administration (EIA) and data for China and India GDP is collected from World Bank Database. The data for exchange rates is collected from Blomberg database. There were only 17 index available on monthly basis, 2 on quarterly basis and 9 on yearly basis. The quarterly and yearly data are converted to monthly data using interpolation.

Table 3.1: Description of input variables under Group-A

S.No.	**Feature Name**	**Feature Code**
1	Non-OPEC Production	Non-OPEC-P
2	OPEC Supply	OPEC-S
3	OECD Consumption	OECD-C
4	Non-OECD Consumption	Non-OECD-C
5	India Consumption	IC
6	China Consumption	CC
7	Primary Energy Consumption	PEC
8	OPEC Spare Capacity	OSC
9	OECD Reserves	OECD-R
10	OPEC reserves	OPEC-R
11	China Reserves	CR
12	Reserves-Production Ratio	RP
13	Strategic Petroleum Reserves	SPR
14	OECD Petroleum Stocks	OPS
15	U.S Petroleum Imports from OPEC	I-OPEC
16	U.S Petroleum Imports from Non-OPEC	I-Non-OPEC
17	U.S Refiner Capacity	RC
18	NYMEX future prices	FP
19	GBP/USD	GU
20	JPY/USD	JU
21	EUR/USD	EU
22	China GDP	C-GDP
23	India GDP	I-GDP
24	U.S Inflation	Inf
25	Geopolitics and Economic events	GE
26	U.S. GDP	GDP
27	U.S. Heating Degree Days	HD
28	U.S. Cooling Degree Days	CD

The study uses nearly 17 year (January 1994-December 2011) monthly data as the training and validation sample, whereas the 12 month ahead data (January 2012-December 2012) is used as testing sample. The data for numerous representative index has diverse measurement scales, there-

Table 3.2: Summary Statistics for Group-A

Feature Code	**Correlation Coefficient**	**Shapiro-Wilk**
WTI	1^{**}	0.87176^{**}
Non-OPEC-P	0.8107^{**}	0.96818^{**}
OPEC-S	0.8603^{**}	0.92051^{**}
OECD-C	0.1459^{*}	0.9933
Non-OECD-C	0.9020^{**}	0.93295^{**}
IC	0.8702^{**}	0.98605^{*}
CC	0.8934^{**}	0.9346^{**}
PEC	0.2212^{**}	0.97463^{**}
OSC	-0.3671^{**}	0.97192^{**}
OECD-R	0.7516^{**}	0.75038^{**}
OPEC-R	0.8827^{**}	0.89231^{**}
CR	-0.7309^{**}	0.98633^{*}
RP	0.6101^{**}	0.67444^{**}
SPR	0.8653^{**}	0.88979^{**}
OPS	0.7340^{**}	0.96582^{**}
I-OPEC	0.5736^{**}	0.98826^{**}
I-Non-OPEC	0.3790^{**}	0.98356^{*}
RC	0.8168^{**}	0.90412^{**}
FP	0.9999^{**}	0.87141^{**}
GU	0.4995^{**}	0.92163^{**}
JU	0.4871^{**}	0.92025^{**}
EU	0.6695^{**}	0.97169^{**}
C-GDP	0.8672^{**}	0.82988^{**}
I-GDP	0.8988^{**}	0.85495^{**}
Inf	0.1739^{**}	0.86896^{**}
GE	0.0129	0.90889^{**}
GDP	0.8296^{**}	0.90625^{**}
HD	-0.0767	0.89082^{**}
CD	0.0948	0.81058^{**}

Notes:
* Significant at 5% level.
** Significant at 1% level.

fore, data corresponding to 28 index is pre-processed by scaling to (-1,1) as input for Multi-Layered perceptron neural network. The ranges of values in input layer for General Regression neural network are standardized by subtracting the median and dividing by the interquartile range. Table 3.2 presents summary statistics for the crude oil price and each factor considered under Group-A. The result of Shapiro-Wilk test indicate that most variables do not follow normal distribution. The correlation between various variables and WTI indicates that correlation coefficient of most variables is significant at 1% level.

3.7.2 Group B

The factors driving oil prices are classified into eight major classes: Speculations (2), Supply (3-4), Demand (5-8), Reserves (9-15), Inventory (16-18), Exchange Market (19-22), Stock Market (23) and Economy (24-26) as shown in Table 3.3. The features are selected on the basis of extensive literature review as discussed in Chapter 2. According to data availability, 25 representative index were chosen within eight classes. Data for WTI, supply, demand, inventory and reserves is from the Energy Information Administration , data for speculation is from the Commodity Futures Trading Commission (CFTC) and data for economy and U.S. Dollar exchange rate index is from EIA's STEO reports.

The data for each of the representative index is on monthly basis from January 1995 to November 2013. There were only 16 indexes data available on monthly basis, 5 on quarterly basis and 4 on yearly basis. For synchronization of data on monthly basis, quarterly and yearly data are converted to monthly data using interpolation. Due to non-availability of data before January 1995, the study has limited the analysis from January 1995 to November 2012. The study uses nearly 17 year (January 1995-November 2012) monthly data as the training and validation sample, whereas the 12 month ahead data (December 2012-November 2013) is used as testing sample.

The data for numerous representative index has diverse measurement scales, therefore, data corresponding to 25 index is pre-processed by scaling to (-1,1) as input for Multi-Layered perceptron neural network. The ranges of values in input layer for General Regression neural network are standard-

Table 3.3: Description of input variables under Group-B

Feature No.	**Feature Name**	**Feature Code**
2	Speculation	NCPP
3	Non-OPEC Production	Non-OPEC-P
4	OPEC Supply	OPEC-S
5	OECD Consumption	OECD-C
6	China Consumption	CC
7	Non-OECD Consumption	Non-OECD-C
8	India Consumption	IC
9	OPEC Spare Capacity	OSC
10	OECD Petroleum Stock	OPS
11	Reserve-Production Ratio	RP
12	Strategic Petroleum Reserves	SPR
13	OECD Reserves	OECD-R
14	OPEC Reserves	OPEC-R
15	China Reserves	CR
16	U.S. Petroleum Import From OPEC	I-OPEC
17	U.S. Petroleum Import From Non-OPEC	I-Non-OPEC
18	U.S. Refinery Capacity	RC
19	U.S. Dollar Exchange Rate Index	DER
20	GBP/USD	GU
21	JPY/USD	JU
22	EUR/USD	EU
23	Dow Jones Index	DJI
24	U.S. Gross Domestic Product	GDP
25	Consumer Price Index	CPI
26	Producer Price Index-Petroleum	EPPI

Table 3.4: Summary Statistics for Group-B

Feature Code	**Correlation Coefficient**	**Shaprio-Wilk Test**
WTI	1.0000	0.89334
NCPP	0.9168	0.89865
Non-OPEC-P	0.8203	0.9529
OPEC-S	0.8821	0.93077
OECD-C	-0.0316	0.9942
CC	0.8895	0.93186
Non-OECD-C	0.9007	0.9279
IC	0.8572	0.96032
OSC	-0.3772	0.96913
OPS	0.7348	0.96351
RP	0.7171	0.95136
SPR	0.8626	0.85435
OECD-R	0.7445	0.74296
OPEC-R	0.8678	0.86271
CR	-0.5203	0.92993
I-OPEC	0.2308	0.97735
I-Non-OPEC	0.4812	0.99448
RC	0.8086	0.90578
DER	0.1536	0.89732
GU	0.4398	0.92238
JU	0.4487	0.92278
EU	0.5699	0.97217
DJI	0.8160	0.9607
GDP	0.8546	0.91895
CPI	0.9158	0.93949
EPPI	0.9844	0.87898

Notes:
* Significant at 5% level.
** Significant at 1% level.

ized by subtracting the median and dividing by the interquartile range. Table 3.4 presents summary statistics for the crude oil price and each factor considered in the study. The result of Shapiro-Wilk test indicates that most variables do not follow normal distribution. The correlation between various variables and WTI indicates that correlation coefficient of most variables is significant at 1% level.

3.8 Concluding Remarks

This chapter illustrated the data mining process followed for the research work starting with defining business problem, data collection, data preparation, model building & evaluation to knowledge deployment. The building blocks for the two proposed algorithms MI^3 and I^2MI^2 for feature selection were discussed. The issues of data-preprocessing steps, number of hidden neurons and over-fitting related to neural network framework were discussed.

The factors considered within both groups are classified into various classes as supply, demand, inventory, reserves, economy, exchange market, stock market and many more. The correlation between various factors considered for the study shows significant relationship with oil prices. The test for normality indicates that most variables do not follow normal distribution. The factors are standardize according to specification required of each neural network.

Chapter 4

Feature Selection for Oil Price Prediction

4.1 Overview

In this chapter, the analysis and findings of the research work is discussed in detail. Oil prices are a key variable in evaluation of economic development, energy policy decisions and stock markets. Oil prices have been governed by various strategic indicators and it is important to identify key factors driving them to achieve high prediction performance for oil prices. There is vast literature accounting for non-linear or non-stationary type complexity of crude oil prices but the empirical literature is very far from any consensus on identifying key factors driving oil prices. Feature selection play an important role in data mining process to extract relevant and non-redundant features. Most of the feature selection methods are based on the assumption of conditional independence or on need of the number of features to be extracted. But still these methods can't provide the minimal set of features that are most relevant and non-redundant for the study. There is a lack of robust feature selection method to select relevant and non-redundant factors for oil price forecasting that can incorporate complexities of crude oil prices.

This chapter deals with analysis of the data using two new proposed MI^3 and I^2MI^2 algorithms as feature selection method to assess the non-linear dependencies between oil prices and input variables. The proposed algorithms are build on the pillars of information-theoretic concepts and are composed of two and three stages respectively. Interaction information and

mutual information concepts are used as building blocks for the proposed MI^3 and I^2MI^2 algorithms. The proposed algorithms are used to provide an insight on the explanatory power of factors and contribution of these factors in driving oil prices for Group-A and Group-B category. Section 4.2 discusses the theme based literature review. The proposed algorithms for the study are discussed in detail in section 4.3. The data analysis and findings for both group of datasets are explained in section 4.4. Section 4.5 provide concluding remarks for this chapter.

4.2 Literature Review

Crude oil prices are influenced by large number of factors, which are complex, noisy, and uncertain [11]. There is no single indicator (lags, future prices or macroeconomic variables) which can provide a complete picture of how prices can be determined. There are few indicators that can give us a snapshot of some fluctuation in oil prices and by modelling these significant snapshots, one can get a clear picture on direction of future oil prices. Xiaotong et al. [116] suggested to build a model for forecasting prices with influential indicators that cause changes in prices. In particular, feature selection plays an important part in data mining to select the set of features that are relevant and non-redundant for prediction. It is important to identify a minimal set of input variables that achieves highest prediction performance for oil prices. The empirical literature is very far from any consensus about selecting the appropriate features/ indicators that can explain the true characteristics of oil market. In economics literature, there are many studies that had examined the relationship between oil prices and macroeconomic variables but there seem be to no consensus on the extent to which these macroeconomic variables are related to oil prices. Further, all existing methods of predicting oil prices have accounted for non-linearity, non-stationary and time-varying structure of the oil prices [43] [70] [40] [88] [24] but seldom have focused on selecting significant features with high prediction power.

Most of the researchers have designed input vector for oil price forecasting based on judgemental criterion or trial and error procedures (conclusion drawn from literature review in chapter-2). To overcome this research gap, this thesis focuses on mining the key factors driving oil prices using robust feature selection methods for achieving high prediction performance.

Researchers in data mining had provided many feature selection methods such as Correlation based Feature Selection (CFS), Modified Relief (MR) and Mutual Information + Modified Relief (MR + MI) to enhance performance of data mining problems, while at the same time reducing the number of features used in learning process. No single learning algorithm is superior to other for all problems. These feature selection methods fail to select relevant features when data contains strong interacting features. The basic assumption of conditional independence of these feature selection methods degrades the performance of learners if features are strongly inter-connected. Most of the real world problems contain features that are strongly dependent on each other. There is a lack of robust feature selection method for designing the input vector for oil price forecasting that can incorporate non-linearity, non-stationary and time varying properties of oil prices, together with dependency between features. To overcome these research gaps, this thesis has focused on developing two new feature selection methods using mutual information and interaction information.

Recently, information-theoretic approaches are increasingly being used for studying the dependency and association of input variables for classification or predication problems. The thematic literature review has been discussed in this chapter to provide an insight on application of information-theoretic approaches whereas overall literature review is already discussed in detail in chapter 2. Amjady and Daraeepour [117] proposed a feature selection technique composed of modified relief and mutual information to design an input vector for price forecasting of PJM, Spanish and Ontario electricity markets. The set of candidate inputs from the proposed method were used as input for cascaded neural network that act as forecasting engine. Zhang et al. [118] proposed a path consistency algorithm based on conditional mutual information and showed its significant performance to infer gene regulatory networks (GRNs) for understanding the complex regulatory mechanisms in cellular systems. Mutual information is a generalization of correlation analysis as it captures non-linear relationships in a non-parametric way. Menezes et al. [119] used mutual information to provide more information on the relationship among stock markets of G7 countries as compared to Granger Causality and Vector Error Correction Models.

4.3 Feature Selection Methods: MI^3 & I^2MI^2 Algorithm

Inspired by the superiority of information-theoretic approaches, this chapter propose two new feature selection methods: MI^3 &I^2MI^2 algorithms. The term MI^3 stands for **M**utual **I**nformation **I**nteraction **I**nformation wherein mutual information and interaction information based irrelevance and redundancy filters are applied. This proposed algorithm consists of two stages. In the first stage, mutual information based irrelevance filter is proposed to select the most relevant features from the set of candidate inputs. The selected features from stage one are filter based on a threshold value. The filtered features are sent to stage two. In the second stage, interaction information based redundancy filter is propose to remove the redundant features from selected relevant candidates. The set of features having negative interaction information are used to filter redundant features. The selected features filtered from stage one and two form a set of relevant and non-redundant features for the study. These selected features are used as set of input variables to build neural networks for predicting oil prices. The ensemble model can provide an insight into the explanatory power of selected relevant and non-redundant features and their contribution in driving oil prices.

Without perturbing about the number of features to be extracted, on the natural domain, MI^3 algorithm eliminates more than $\frac{1}{2}$ number of features. Due to non-availability of data, it may not be possible to have forecast value for all reduced number of variables. Many researchers have argued that structural model fails to forecast oil prices due to non-availability of forecast values of right hand side indicators [30]. Therefore, there is a need to expand above algorithm further so as to provide a minimal set of most relevant and non-redundant features. This limitation of MI^2 algorithm is overcome by an extended version I^2MI^2 algorithm that comprise of three-stages to provide 100% relevant and non-redundant features for the study.

The three-stage algorithm comprise of one stage build on interaction information (I^2) based filter and two stages based on mutual information (MI^2) based filters. In the first stage, mutual information is computed between dependent variable and input variables. The variables are ranked according

to normalized mutual information and the irrelevant features are filtered out based on a threshold value. In the next stage, three-variable interaction information is computed to remove the redundant features among the candidate inputs till selected features are synergy in nature. This study incorporates the concept of interaction information to filter redundant input variables instead of correlation analysis or partial correlation analysis. Interaction information is favoured as compared to correlation analysis as it measures non-linear dependency among features. Further, in the third stage, mutual information is computed between the selected features from stage two and then ranked according to normalized mutual information. Depending on a threshold value, redundant features are filtered out according to relevance rank in stage one. The selected features then uses neural networks as forecasting engines for oil price series.

For better illustration, the step by step procedure followed for the proposed feature selection methods are discussed in detail below. MI^3 algorithm comprise of steps 1-9 and the extended version of MI^3 named as I^2MI^2 comprise of all steps 1-14.

Step 1 All N variables are linearly normalized in the range of [0,1] to eliminate the effect of different ranges of input variables.

Step 2 Relevance of each input variable to target variable (WTI price) is calculated by mean of mutual information.

Step 3 The relevance of each variable is normalized with respect to maximum relevance and further ranked accordingly.

Step 4 Keep n features with normalized relevance more than a pre-specified threshold $Th1$ and filter out $N - n$ remaining features. Step 1-4 completes the process of Stage one to provide the set S_1 of relevant features.

Step 5 Compute three-variable interaction information $I(Y, X_i, X_j)$ between target variable (Y) and relevant features (X_i, X_j) from set S_1. Filter all subset of features $\{Y, X_i, X_j\}$ having negative interaction information.

Step 6 Consider the most relevant feature X_{imax} (having maximum mutual information with target variable) from Step 3. If there exist any subset

(Y, X_{imax}, X_j) for which interaction information is negative, proceed to step 7. Otherwise add it to set S_2 and repeat step 6 for all X_i based on their relevance rank in stage one.

Step 7 If $MI(Y, X_{imax}) > MI(Y, X_j)$, then X_j is redundant feature and filtered out. Add X_{imax} to set S_2.

Step 8 Repeat steps 5-7 for higher order interaction till all subset of variables are in synergy i.e. $I(Y, X_i, X_j) > 0$.

Step 9 Steps 5-8 provide a set S_2 of non-redundant (synergy) and relevant features. This completes the procedure to be followed for selecting features based on MI^3 algorithm. MI^3 algorithm provides relevant and non-redundant features by following steps 1-9. The flowchart for stage one and two of proposed MI^3 algorithm is shown in Fig 4.1 and Fig 4.2 respectively. MI^3 algorithm can provide list of relevant and non-redundant features but an extended version called I^2MI^2 algorithm is proposed to provide the minimal representative of features to achieve high prediction performance for oil prices. This three-stage algorithm comprises of above mentioned steps together with steps 10-12 as mentioned below.

Step 10 Compute mutual information between the non-redundant (synergy) and relevant variables in set S_2 obtained from stage one and stage two.

Step 11 Based on a pre-specified threshold $Th2$, select the pairs of variable such that $I(X_i, X_j) \geq Th2$.

Step 12 Consider X_{imax} (the most relevant variable from Step 3) and add it to the final set S_3. If there exists any subset (X_{imax}, X_j) for which $I(X_{imax}, X_j) \geq Th2$, the feature X_j is considered as redundant feature and has to be filtered out by mutual information based redundancy filter. Otherwise, the feature X_j is added to final features set S_3.

Step 13 Repeat step 12 for all subsequent ranked features X_i. This completes the process of stage 3 to find more relevant features.

Step 14 The set S_3 of selected features is the final set of input variables obtained using the proposed three-stage I^2MI^2 Algorithm. The flowchart

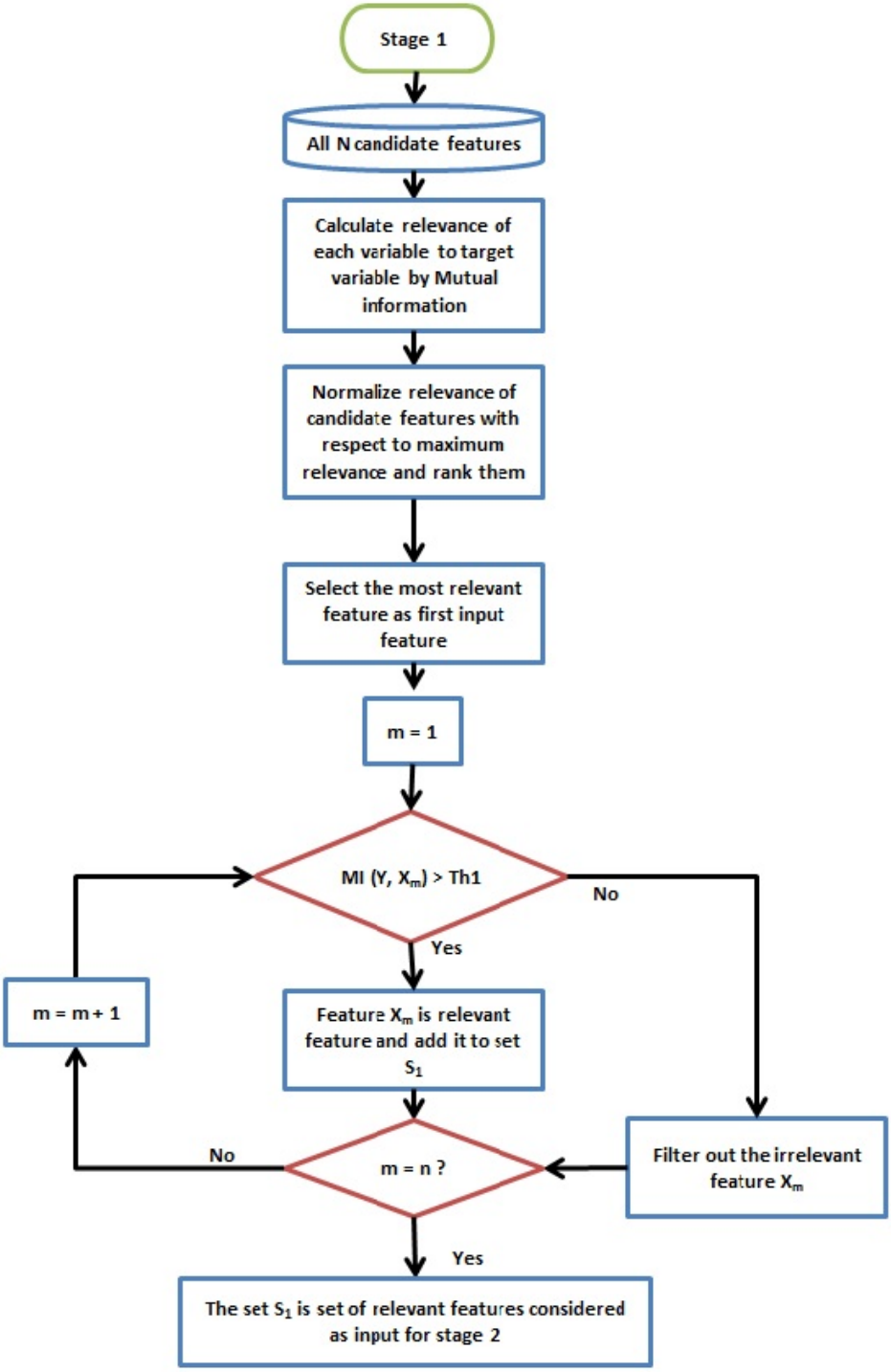


Figure 4.1: Flowchart of stage one of the proposed algorithm

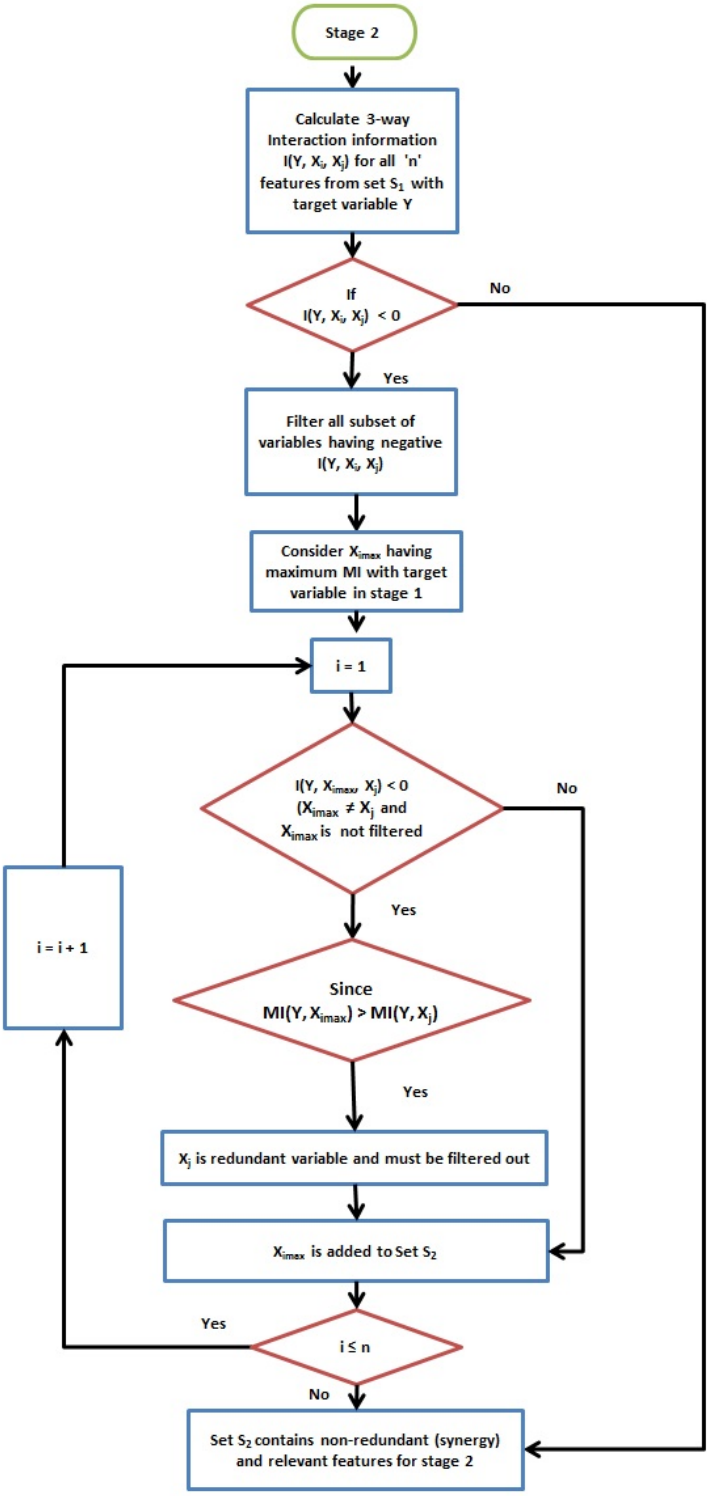


Figure 4.2: Flowchart of stage two of the proposed algorithm

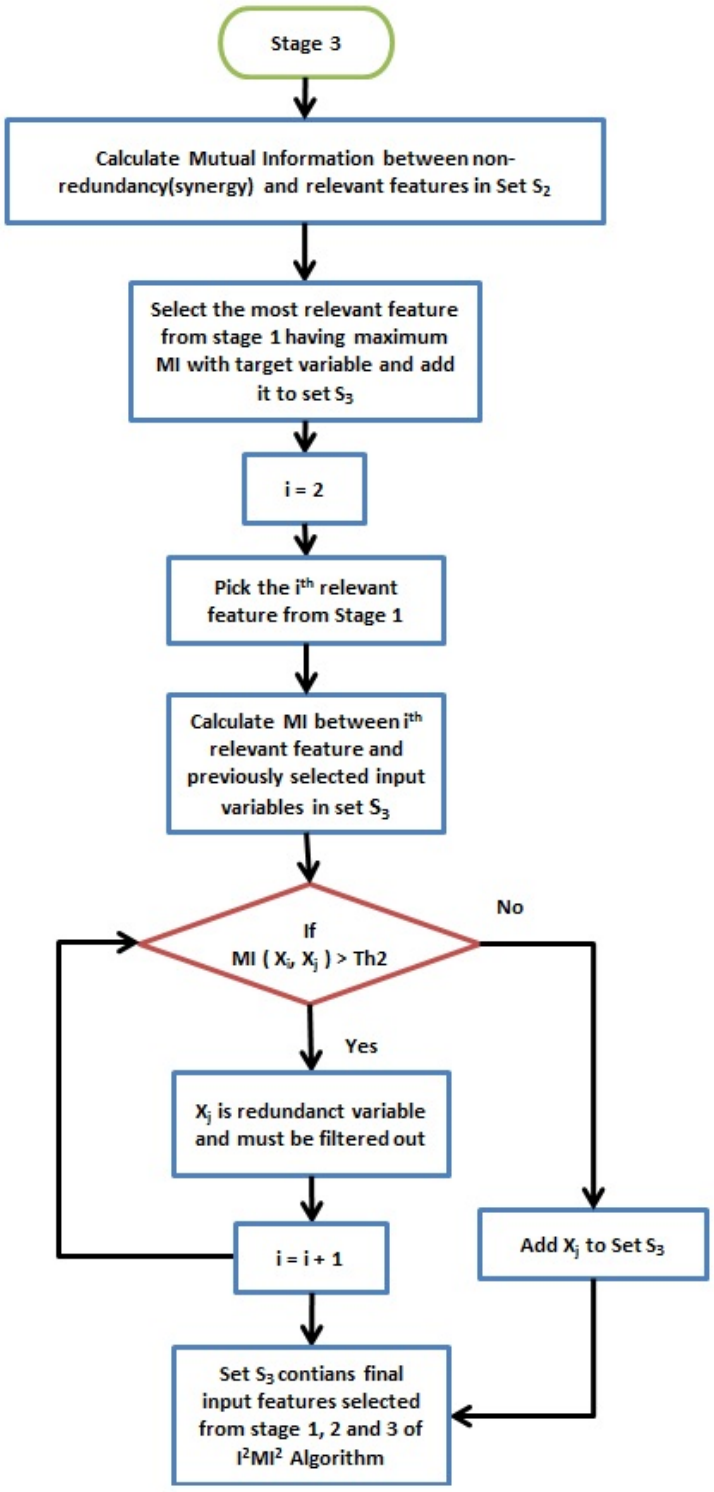


Figure 4.3: Flowchart of stage three of the proposed algorithm

for the third stage of proposed algorithm is shown in Fig 4.3. The selected features are considered as input to various forecasting engines.

4.4 Data Analysis

4.4.1 Feature Selection by MI^3 & I^2MI^2 Algorithm (Group A)

The step by step procedure of proposed I^2MI^2 algorithm is explained in detail below.

Stage 1 The goal of stage one is to provide relevant features based on mutual information irrelevant filter. In Table 4.1, the candidate features (column 1) with relevance rank (column 2) and their normalized relevance value (column 3) with respect to maximum mutual information with WTI spot prices are shown. Column 4 provides the feature number. A low threshold value $Th1$ is chosen to decide for independence between target and input candidate variables. According to the pre-specified threshold $Th1$, the variable GE, HD and CD are filtered out to provide set S_1 of relevant features. This completes the process of stage one of I^2MI^2 algorithm.

Stage 2 The three-variable interaction information $I(Y, X_i, X_j)$ between target variable (Y) and relevant features (X_i, X_j) from set S_1 is visually represented by Interaction Information Graph (IIG) in Fig 4.4. The yellow marks represent negative interaction (redundancy) and green marks represent positive interaction (synergy) between variables. For better illustration, Table 4.2 provides the list of pair of variables for which $I(Y, X_i, X_j) < 0$. The obtained results from the redundancy filter are shown in column 1-8.

Interaction information $I(Y, X_i, X_j)$ is a symmetric measure; it cannot derive the direction whether X_j inhibits the correlation between (Y, X_i) or X_i inhibits the correlation between (Y, X_j) . Therefore, it become difficult to filter the redundant variable from the set of relevant features (X_i, X_j) when $I(Y, X_i, X_j) < 0$. This limitation of interaction information is relieved by focusing on mutual information between target and input variables $I(Y, X_i)$. The algorithm in stage two starts with maximum relevance rank variable X_{imax} from Table 4.1. According to relevance ranked Table 4.1, the feature FP(18) is ranked first. Add X_{18} to set S_2. For the first relevance

Table 4.1: Selected features by the stage one irrelevance filter for WTI spot price market

Candidate features	**Relevance Rank**	**Normalized relevance value**	**Feature No.**
FP	1	1	18
IC	2	0.599377	6
OPEC-R	3	0.599377	10
C-GDP	4	0.599377	22
I-GDP	5	0.599377	23
OECD-R	6	0.59177	9
GDP	7	0.566123	26
RP	8	0.536065	12
Inf	9	0.514031	24
RC	10	0.511644	17
SPR	11	0.505644	13
Non-OECD-C	12	0.499905	4
CC	13	0.499545	5
OPEC-S	14	0.444139	2
Non-OPEC-P	15	0.420322	1
EU	16	0.371493	21
CR	17	0.332217	11
I-Non-OPEC	18	0.329361	15
OPS	19	0.314563	14
JU	20	0.303466	20
OSC	21	0.279744	8
GBP/USD	22	0.277578	19
I-OPEC	23	0.221318	16
OECD-C	24	0.214105	3
PEC	25	0.179782	7
CD	26	0.157316	28
HD	27	0.135986	27
GE	28	0.003725	25

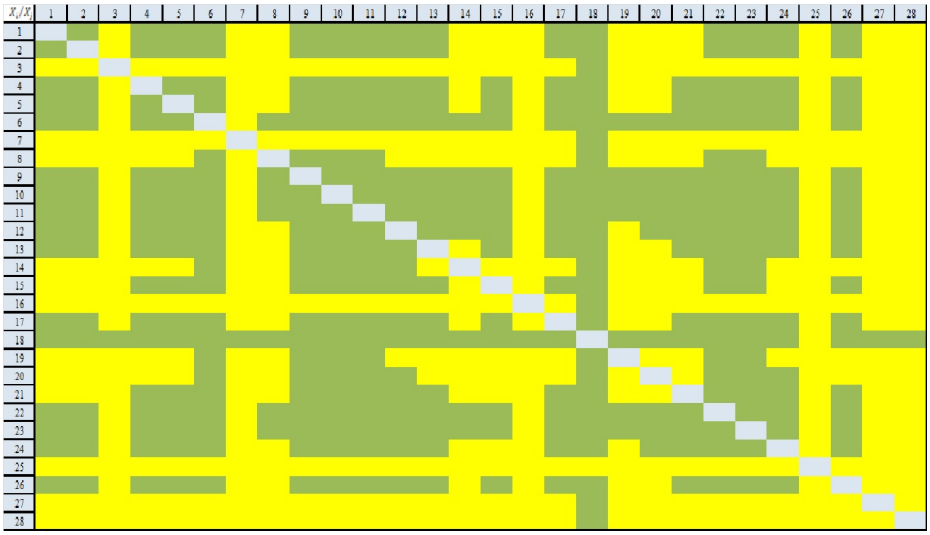


Figure 4.4: Interaction Information Graph for three variables

Table 4.2: List of pair of variables having negative interaction information

Column 1		Column 2		Column 3		Column 4		Column 5		Column 6		Column 7		Column 8	
1	3	4	3	7	26	13	3	16	7	19	28	24	16	27	8
1	7	4	7	7	27	13	7	16	8	20	1	24	19	27	9
1	8	4	8	7	28	13	8	16	9	20	2	24	25	27	10
1	14	4	14	8	1	13	14	16	10	20	3	24	27	27	11
1	15	4	16	8	2	13	16	16	11	20	4	24	28	27	12
1	16	4	19	8	3	13	19	16	12	20	5	25	1	27	13
1	19	4	20	8	4	13	20	16	13	20	7	25	2	27	14
1	20	4	25	8	5	13	25	16	14	20	8	25	3	27	15
1	21	4	27	8	7	13	27	16	15	20	13	25	4	27	16
1	25	4	28	8	12	13	28	16	17	20	14	25	5	27	17
1	27	5	3	8	13	14	1	16	19	20	15	25	6	27	19
1	28	5	7	8	14	14	2	16	20	20	16	25	7	27	20
2	3	5	8	8	15	14	3	16	21	20	17	25	8	27	21
2	7	5	14	8	16	14	4	16	22	20	19	25	9	27	22
2	8	5	16	8	17	14	5	16	23	20	21	25	10	27	23
2	14	5	19	8	19	14	7	16	24	20	25	25	11	27	24
2	15	5	20	8	20	14	8	16	25	20	26	25	12	27	25
2	16	5	25	8	21	14	13	16	26	20	27	25	13	27	26
2	19	5	27	8	24	14	15	16	27	20	28	25	14	27	28
2	20	5	28	8	25	14	16	16	28	21	1	25	15	28	1
2	21	6	3	8	26	14	17	17	3	21	2	25	16	28	2
2	25	6	7	8	27	14	19	17	7	21	3	25	17	28	3
2	27	6	16	8	28	14	20	17	8	21	7	25	18	28	4
2	28	6	25	9	3	14	21	17	14	21	8	25	19	28	5
3	2	6	27	9	7	14	24	17	16	21	14	25	20	28	6
3	4	6	28	9	16	14	25	17	19	21	15	25	21	28	7
3	5	7	1	9	25	14	26	17	20	21	16	25	22	28	8
3	6	7	2	9	27	14	27	17	25	21	19	25	23	28	9
3	7	7	3	9	28	14	28	17	27	21	20	25	24	28	10
3	8	7	4	10	3	15	1	17	28	21	25	25	26	28	11
3	9	7	5	10	7	15	2	18	25	21	27	25	27	28	12
3	10	7	6	10	16	15	3	19	1	21	28	25	28	28	13
3	11	7	8	10	25	15	7	19	2	22	3	26	3	28	14
3	12	7	9	10	27	15	8	19	3	22	7	26	7	28	15
3	13	7	10	10	28	15	14	19	4	22	16	26	8	28	16
3	14	7	11	11	3	15	16	19	5	22	25	26	14	28	17
3	15	7	12	11	7	15	19	19	7	22	27	26	16	28	19
3	16	7	13	11	16	15	20	19	8	22	28	26	19	28	20
3	17	7	14	11	25	15	21	19	13	23	3	26	20	28	21
3	19	7	15	11	27	15	24	19	14	23	7	26	25	28	22
3	20	7	16	11	28	15	25	19	15	23	16	26	27	28	23
3	21	7	17	12	3	15	27	19	16	23	25	26	28	28	24
3	22	7	19	12	7	15	28	19	17	23	27	27	1	28	25
3	23	7	20	12	8	16	1	19	20	23	28	27	2	28	26
3	24	7	21	12	16	16	2	19	21	24	3	27	3	28	27
3	25	7	22	12	19	16	3	19	24	24	7	27	4		
3	26	7	23	12	25	16	4	19	25	24	8	27	5		
3	27	7	24	12	27	16	5	19	26	24	14	27	6		
3	28	7	25	12	28	16	6	19	27	24	15	27	7		

Table 4.3: Filtered feature by redundancy filter in stage two of proposed algorithm

Filtered features by redundancy filter	**Feature No.**	**Feature Rank (Stage 1)**
Non-OPEC-P	1	15
OPEC-S	2	14
Non-OECD-C	4	24
CC	5	13
IC	6	2
OECD-R	9	6
OPEC-R	10	3
CR	11	17
RP	12	8
SPR	13	11
RC	17	10
FP	18	1
C-GDP	22	4
I-GDP	23	5
Inf	24	9
GDP	26	7

ranked variable X_{18}, there is a single set $\{Y, X_{18}, X_{25}\}$ having negative interaction information as evident from Table 4.2. The question that arises here is whether X_{25} inhibits the correlation between Y and X_{18} or X_{18} inhibits the correlation between Y and X_{25}. The redundant variable is filtered by comparing mutual information $I(Y, X_{18})$ with $I(Y, X_{25})$. The results obtained in Table 4.1 showed that mutual information $I(Y, X_{18}) > I(Y, X_{25})$. Therefore, the variable X_{25} is a redundant variable and must be filtered out from the list of relevant and non-redundant variables. Similarly, the process holds for next ranked feature X_6 from Table 4.1. There are six set of variables $\{Y, X_6, X_j\}$ having interaction information $I(Y, X_6, X_j) < 0$ $\forall\ j = \{3, 7, 16, 25, 27, 28\}$. The variable X_{25} is already filtered out, but there are five more variables having negative interaction information with X_6 and Y. Since mutual information $I(Y, X_6) > I(Y, X_j)\ \forall\ j$, therefore, X_6 is added to set S_2 and $X_j\ \forall\ j = \{3, 7, 16, 27, 28\}$ are filtered out by redundancy filter. Table 4.3 shows the list of relevant and non-redundant features selected from stage one and two.

The number of candidate inputs (N) are reduced from 28 to 16 in stage two; i.e. to 50% of the actual number of input features. The set of vari-

ables in Table 4.3 have been cross-validated for three-variable interaction information such that $I(Y, X_i, X_j) > 0$ always. The proposed I^2MI^2 Algorithm is superior as compared to (MR + MI) algorithm [117] in identifying the set of non-redundant features. The features selected from (MR + MI) algorithm are $2, 4, 8, 11, 18, 19, 20, 21$ respectively. Interaction information is computed for the pair of features selected from (MR + MI) algorithm and found to be negative for the set $\{Y, X_2, X_4\}$ and $\{Y, X_2, X_8\}$. The limitation of other competing feature selection methods lies in fact that they were not successful in providing 100% relevant and non-redundant features set. The set S_2 is the set of features selected by stage two of the proposed algorithm. Since three-variable interaction information has provided features for which $I(Y, X_i, X_j) > 0$, therefore, higher order interaction information is not required. This complete two stages for the proposed MI^3 algorithm. The final selected features from the proposed MI^3 algorithm are Non-OPEC Production (1), OPEC Supply (2), Non-OECD Consumption (4), China Consumption (5), India Consumption (6), OECD Reserves (9), OPEC Reserves (10), China Reserves (11), Reserve-Production Ratio (12), Strategic Petroleum Reserves (13), U.S Refinery Capacity (17), NYMEX Future Price (18), China GDP (22), India GDP (23), U.S Inflation (24) and U.S GDP (26).

Stage 3 The extended version I^2MI^2 algorithm is designed to filter out redundant features based on mutual information between selected features $I(X_i, X_j)$ from stage two. The algorithm in stage three starts with maximum relevance rank variable X_{18} from Table 4.1. By default, X_{18} is considered as part of final set S_3. Consider the next relevance rank variable X_6. According to the pre-specified threshold value $Th2$, variables from set S_2 are filtered out based on mutual information between selected features. Since $I(X_{18}, X_6) > Th2$, therefore, X_6 is filtered out based on redundancy filter. For the next relevance ranked variable X_m, calculate maximum mutual information $Max(MI)$ between X_m and previously selected candidates by redundancy filter. If mutual information $Max(MI) > Th2$ for any set, then X_m is filtered out by redundancy filter. Otherwise, X_m is added to the final selected features set S_3. The algorithm will run iteratively for all 16 selected variables from stage two.

The final selected features from the proposed I^2MI^2 algorithm are Non-OPEC Production (1), OPEC Supply (2), Non-OECD Consumption (4),

China Consumption (5), China Reserves (11) and NYMEX future prices (18). The selected features (4, 5, and 11) show the enormous impact of emerging economies in driving crude oil prices. According to BP [5], Non-OECD consumption grew by 5.3%, in track with 10-year average. The results have proved the importance of Non-OECD consumption through recent change in data post 2009. Further, Non-OPEC production which constitutes production from developed countries is also influential in deciding direction of crude oil prices.

4.4.2 Numerical Results

Forecasting Results of MI^3 Algorithm

In this experiment, the performance of two-stage MI^3 Algorithm is evaluated with known feature selection techniques such as : Modified Relief (MR) [120],Correlation Feature Selection (CFS) [121] and Modified Relief + Mutual information (MR + MI) [117]. For this purpose, the proposed I^2MI^2 algorithm with three forecasting engines MLP (MI^3 + MLP), GRNN (MI^3 + GRNN) and CNN (MI^3 + CNN) is compared with:

- Single-stage feature selection techniques: MR + CNN, MR + MLP, MR + GRNN , CFS+MLP, CFS+GRNN, CFS+CNN
- Two-stage feature selection techniques: (MR + MI) + CNN [117], (MR + MI) + MLP, (MR +MI) + GRNN

The performance criterion used for comparing proposed MI^3 algorithm with other methods are RMSE (Root Mean Square Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). The performance criterion are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(WTI_{Actual} - WTI_{Predicted})^2} \tag{4.1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|WTI_{Actual} - WTI_{Predicted}| \tag{4.2}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|WTI_{Actual} - WTI_{Predicted}|}{WTI_{Actual}} \tag{4.3}$$

The RMSE, MAE and MAPE are represented in Table 4.4. The results provided the following observations:

- MI^3 algorithm has performed better with GRNN and MLP neural network. The performance of MI^3 algorithm with GRNN is s superior among 12 models in terms of RMSE & MAE.
- CFS has performed well with MLP but there are several limitations regarding relevance, redundancy and basic assumptions of conditional independence associated with CFS.
- (MR + MI) algorithm based feature selection together with MLP neural network seems to have performed best in terms of MAPE but the input variables were not equally distributed in this case. The neural structure in this case is fundamentally based on the single pillar i.e. NYMEX future prices. Therefore, MI^3 + GRNN based methodology proved to be best for inferring the explanatory power of sixteen selected factors in deciding future price path.
- CNN as a forecasting engine has performed worst in comparison to MLP & GRNN.

Table 4.4: Performance criterion for comparing MI^3 with different feature selection methods

Models	**RMSE**	**MAE**	**MAPE**
$MI^3 + GRNN$	0.32	0.25	0.92
$MI^3 + MLP$	0.60	0.46	1.68
$MI^3 + CNN$	2.27	1.73	5.07
$MR + MI + GRNN$	1.51	0.96	2.38
$MR + MI + MLP$	0.48	0.29	0.89
$MR + MI + CNN$	1.46	0.97	2.67
$MR + GRNN$	2.38	1.55	3.99
$MR + MLP$	1.27	0.85	2.81
$MR + CNN$	4.11	2.81	8.65
$CFS + GRNN$	1.82	1.15	2.94
$CFS + MLP$	0.39	0.29	0.92
$CFS + CNN$	3	2.21	7.11

Based on the above mentioned observations, this study is evaluating the proposed methodology to explain the explanatory power of these factors and their contribution in explaining direction of oil prices. The explanatory power for oil prices using sixteen selected features is 99.01%, indicating that the variable reduction is reasonable and will have no essential influence on subsequent analysis. It is clear that our proposed algorithm is able to

identify true drivers of oil prices and has used most relevant and non-redundant features as input to GRNN for supreme performance. Fig 4.5 describes the explanatory power of each factor selected by proposed MI^3 algorithm and GRNN (as forecasting engine). It is evident from Fig 4.5 that reserves as a factor played a significant role in deciding direction of oil prices for nearly 17 years. The relationship between future and spot price is a debatable issue and the results show that in long run, future prices have largely influenced spot oil prices. The results have shown the effect of emerging economies in influencing oil prices, with recent change in data post 2009.

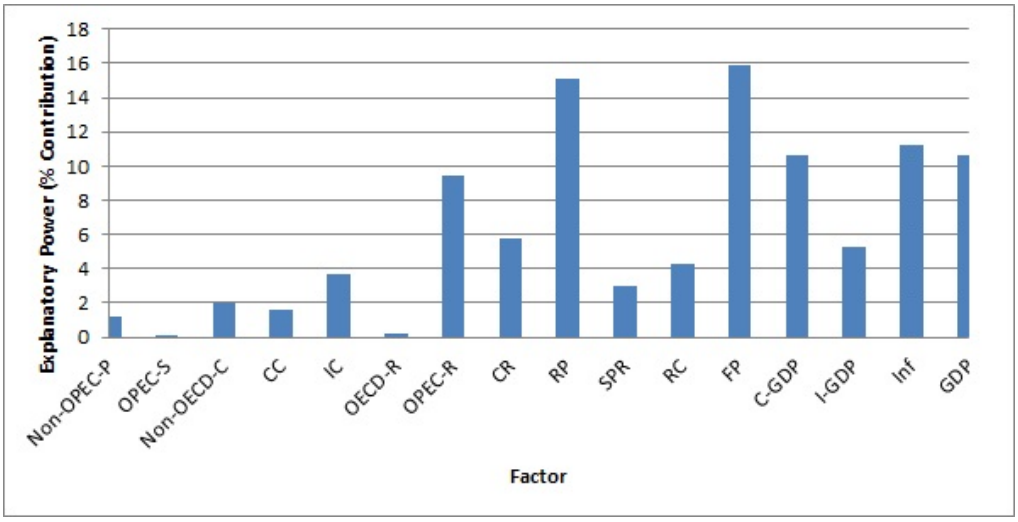


Figure 4.5: Explanatory power of 16 selected factors based on MI^3 Algorithm

Forecasting Results of I^2MI^2 Algorithm

In this experiment, the performance of three-stage I^2MI^2 Algorithm is evaluated with known feature selection techniques such as : Modified Relief (MR) [120],Correlation Feature Selection (CFS) [121] and Modified Relief + Mutual information (MR + MI) [117]. For this purpose, the proposed I^2MI^2 algorithm with three forecasting engines MLP (I^2MI^2 + MLP), GRNN (I^2MI^2 + GRNN) and CNN (I^2MI^2 + CNN) is compared with:

- Single-stage feature selection techniques: MR + CNN, MR + MLP, MR + GRNN , CFS+MLP, CFS+GRNN, CFS+CNN
- Two-stage feature selection techniques: (MR + MI) + CNN [117], (MR + MI) + MLP,(MR +MI) + GRNN

The performance criterion used for comparing proposed I^2MI^2 algorithm with other methods are RMSE (Root Mean Square Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). The performance criterion are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(WTI_{Actual} - WTI_{Predicted})^2} \tag{4.4}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|WTI_{Actual} - WTI_{Predicted}| \tag{4.5}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|WTI_{Actual} - WTI_{Predicted}|}{WTI_{Actual}} \tag{4.6}$$

Table 4.5: Performance criterion for comparing I^2MI^2 with different feature selection methods

Models	**RMSE**	**MAE**	**MAPE**
$I^2MI^2 + GRNN$	0.26	0.7	0.58
$I^2MI^2 + MLP$	0.4	0.28	0.87
$I^2MI^2 + CNN$	1.76	1.29	4.08
$MR + MI + GRNN$	1.51	0.96	2.38
$MR + MI + MLP$	0.48	0.29	0.89
$MR + MI + CNN$	1.46	0.97	2.67
$MR + GRNN$	2.38	1.55	3.99
$MR + MLP$	1.27	0.85	2.81
$MR + CNN$	4.11	2.81	8.65
$CFS + GRNN$	1.82	1.15	2.94
$CFS + MLP$	0.39	0.29	0.92
$CFS + CNN$	3	2.21	7.11

The RMSE, MAE and MAPE are represented in Table 4.5. The results provided the following observations:

- The five single-stage feature selection methods have not performed well for complex oil price data except the $CFS + MLP$ ensemble method. The reason behind the poor results from single-stage methods is the existence of redundancy in selected feature set. CFS and MR has performed poor with CNN because they only accounts for linear dependency; however, the relationship of oil prices with macroeconomic variables is too complex and non-linear in nature.

- Two-stage (MR + MI) + CNN algorithm as proposed by Amjady and Daraeepour [117] have not performed well compared to (MR + MI) + MLP and (MR + MI) + GRNN. The results showed superiority of GRNN compared to MLP and CNN forecasting engines.

- The proposed three-stage feature selection I^2MI^2 with GRNN as forecasting engine has performed best among 12 methods employed for forecasting WTI spot prices in terms of RMSE & MAPE. Although, CNN forecasting engine has not performed well but the proposed feature selection I^2MI^2 algorithm has performed significantly well with GRNN and MLP (as forecasting engine). The reason for the best performance lies in the fact that the final set of features selected by I^2MI^2 algorithm are 100% non-redundant (synergy) and highly relevant features for the study.

The study proposes I^2MI^2 algorithm based selected features as input to GRNN forecasting engine to predict oil prices. The proposed model is used to forecast in-sample and out-of-sample. In order to compare the model stability, 17-year monthly data (January 1994-December 2011) is used for training and validation, while the twelve-month ahead data (January 2012-December 2012) is used as testing sample.

The model is used to produce one-month and twelve-month ahead forecasts. The forecasting procedure for one-month ahead prediction begins by training the model with data from January 1994 to December 2011, and forecasting the value of January 2012. Then actual value of January 2012 was added, and the model is refitted to forecast the price of February 2012. The process is repeated till the value of December 2012 is forecast. The one-month ahead forecast from the proposed methodology is compared with monthly forecasts shown in EIA's STEO reports from January 2012 till December 2012. The twelve-month ahead forecast from the proposed methodology is compared with forecast shown in EIA's STEO January 2012 report.

Out-of-sample evaluations are shown in Table 4.6. The results proved the superiority of proposed methodology for both short-run (one-month ahead) and long-run (twelve-month ahead) time period in comparison to EIA's STEO forecasts as reported. The results showed that twelve-month ahead forecasts provide better result than forecasts proposed by EIA's STEO re-

Table 4.6: Out-of-sample forecast comparison

Comparison Model	RMSE	MAE	MAPE
One-month(Proposed)	2.71	1.86	1.98
One-month(STEO)	3.88	3.05	3.33
Twelve-month(Proposed)	1.63	1.27	1.32
Twelve-month(STEO)	9.8	8.36	9.32

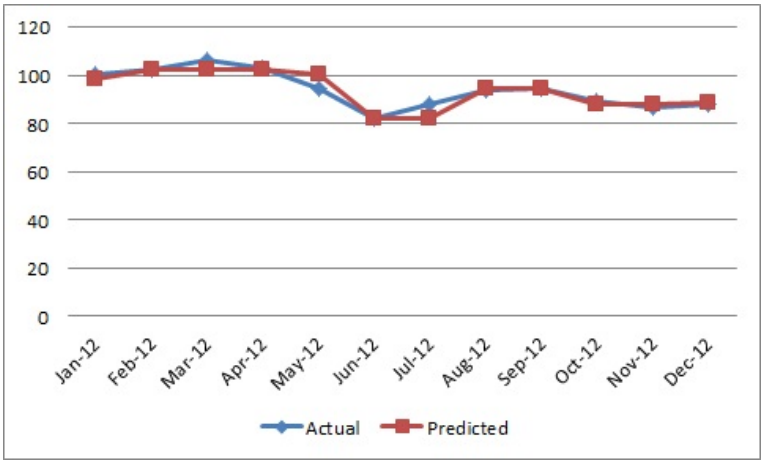


Figure 4.6: One-month out-of-sample forecast

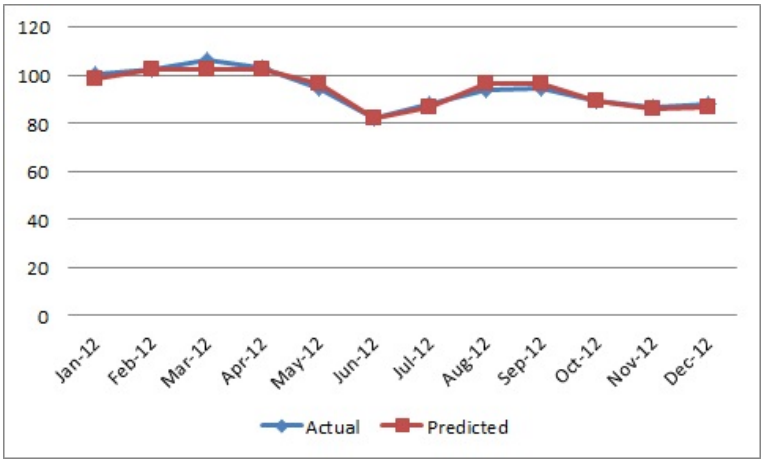


Figure 4.7: Twelve-month out-of-sample forecast

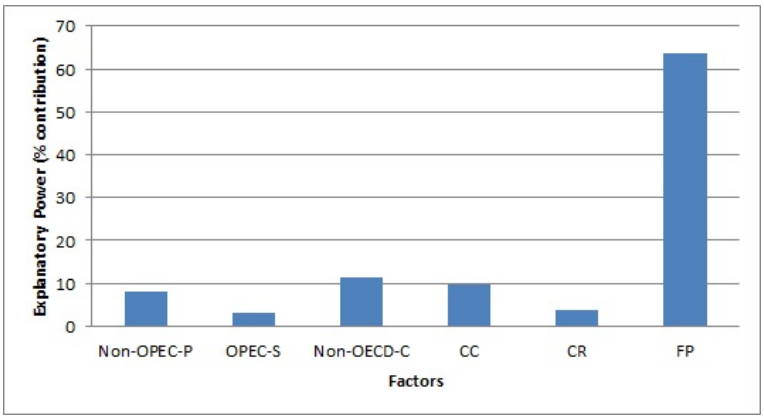


Figure 4.8: Explanatory power of selected features using proposed methodology for Group-A

port. The proposed methodology performed more accurately in long-run forecasting as compared to short-run. It is clear that our proposed model is able to identify true drivers of oil prices and used most relevant and non-redundant features as input to GRNN for supreme performance. Fig 4.6 and Fig 4.7 shows the real and fitted crude oil prices from January 2012-December 2012. The explanatory power for oil prices using six selected features is 99.08%, indicating that the variable reduction is reasonable and will have no essential influence on subsequent analysis. Fig 4.8 describes the explanatory power of each factor selected by proposed I^2MI^2 algorithm and GRNN as forecasting engine. The graph shows massive influence of future prices in influencing oil prices, followed by Non-OPEC production and Non-OECD consumption as new indicators driving them.

4.4.3 Feature Selection by MI^3 & I^2MI^2 Algorithm (Group B)

The proposed two-stage MI^3 algorithm is applied to identify the set of relevant and non-redundant features for Group-B dataset. The proposed three-stage I^2MI^2 algorithm is applied to identify the minimal set of relevant and non-redundant features to achieve high oil price prediction performance. The step by step procedure of MI^3 and I^2MI^2 algorithms are explained as follows. In Table 4.7, the candidate features (column 1) with relevance rank (column 2) and their normalized relevance value (column 3) with respect to maximum mutual information with WTI spot prices are shown. Column 4 shows feature number. According to a pre-specified threshold $Th1$, the variable JU, GU, OSC, OECD-C and I-OPEC are fil-

tered out to provide set S_1 of relevant features. Table 4.8 provides the list

Table 4.7: Selected features by the stage one irrelevance filter for WTI spot price market

Candidate Feature	**Relevance Rank**	**Normalized Relevance Value**	**Feature No.**
EPPI	1	1	26
OECD-R	2	0.895432	13
CPI	3	0.842058	25
OPEC-R	4	0.804443	14
GDP	5	0.789605	24
RP	6	0.748466	11
Non-OECD-C	7	0.714141	7
RC	8	0.709464	18
CC	9	0.707181	6
SPR	10	0.700856	12
OPEC-S	11	0.685697	4
DER	12	0.635516	19
NCPP	13	0.613563	2
IC	14	0.610674	8
Non-OPEC-P	15	0.610190	3
DJI	16	0.571768	23
I-Non-OPEC	17	0.487915	17
CR	18	0.480246	15
EU	19	0.479611	22
OPS	20	0.462465	10
JU	21	0.419529	21
GU	22	0.410434	20
OSC	23	0.375247	9
OECD-C	24	0.347442	5
I-OPEC	25	0.342035	16

of variables for which three-variable interaction information $I(Y, X_i, X_j)$ is negative. Applying the same procedure as discussed earlier, the redundant variables are filtered out. For the first relevance ranked variable X_{26}, there are three sets $\{Y, X_{26}, X_j\}$ where $j = \{5, 9, 16\}$ for which $I\{Y, X_{26}, X_j\} < 0$. Since mutual information $I(Y, X_{26}) > I(Y, X_j)$, therefore X_j are redundant variables and must be filter out. The process is repeated for all variables in Table 4.7. The list of relevant and non-redundant variables passing through stage one and stage two filter are presented in Table 4.9. Stage two has reduced number of candidate inputs (N) from 25 to 15; i.e. to approx 60%. The variables listed in Table 4.9 has been cross-validated such that selected

Table 4.8: List of pair of variables having negative interaction information

Column 1		Column 2		Column 3		Column 4		Column 5		Column 6		Column 7		Column 8	
2	5	5	7	7	21	10	3	16	4	17	20	20	16	22	8
2	9	5	8	7	22	10	4	16	5	17	21	20	17	22	9
2	10	5	9	8	5	10	5	16	6	17	22	20	18	22	10
2	16	5	10	8	9	10	6	16	7	17	23	20	19	22	12
2	17	5	11	8	10	10	7	16	8	18	5	20	21	22	16
2	19	5	12	8	16	10	8	16	9	18	9	20	22	22	17
2	20	5	13	8	17	10	9	16	10	18	10	20	23	22	19
2	21	5	14	8	20	10	16	16	11	18	16	20	24	22	20
2	22	5	15	8	21	10	17	16	12	18	20	21	2	22	21
2	23	5	16	8	22	10	18	16	13	18	21	21	3	22	23
3	5	5	18	8	23	10	19	16	14	18	22	21	4	23	3
3	9	5	19	9	2	10	20	16	15	19	2	21	5	23	5
3	10	5	20	9	3	10	21	16	17	19	5	21	6	23	8
3	16	5	21	9	4	10	22	16	18	19	9	21	7	23	9
3	17	5	22	9	5	10	23	16	19	19	10	21	8	23	10
3	20	5	23	9	6	11	5	16	20	19	16	21	9	23	16
3	21	5	24	9	7	11	16	16	21	19	17	21	10	23	17
3	22	5	25	9	8	12	5	16	22	19	20	21	12	23	19
3	23	5	26	9	10	12	9	16	23	19	21	21	16	23	20
4	5	6	5	9	12	12	16	16	24	19	22	21	17	23	21
4	9	6	9	9	16	12	20	16	25	19	23	21	18	23	22
4	10	6	10	9	17	12	21	16	26	20	2	21	19	24	5
4	16	6	16	9	18	12	22	17	2	20	3	21	20	24	9
4	17	6	20	9	19	13	5	17	3	20	4	21	22	24	16
4	20	6	21	9	20	13	16	17	4	20	5	21	23	24	20
4	21	6	22	9	21	14	5	17	5	20	6	22	2	25	5
4	22	7	5	9	22	14	16	17	8	20	7	22	3	25	16
5	2	7	9	9	23	15	5	17	9	20	8	22	4	26	5
5	3	7	10	9	24	15	16	17	10	20	9	22	5	26	9
5	4	7	16	9	26	16	2	17	16	20	10	22	6	26	16
5	6	7	20	10	2	16	3	17	19	20	12	22	7		

Table 4.9: Filtered features by redundancy filter in stage two of proposed algorithm

Candidate Feature	**Feature No.**	**Relevance Rank(Stage 1)**
Non-OPEC-P	3	15
OPEC-S	4	11
CC	6	9
Non-OECD-C	7	7
IC	8	14
RP	11	6
SPR	12	10
OECD-R	13	2
OPEC-R	14	4
CR	15	18
RC	18	8
DER	19	12
GDP	24	5
CPI	25	3
EPPI	26	1

features are in synergy. This complete two stages for the proposed MI^3 algorithm. The final selected features from the proposed MI^3 algorithm are Non-OPEC Production (3), OPEC Supply (4), China Consumption (6), Non-OECD Consumption (7), India Consumption (8), Reserve-Production Ratio (11), Strategic Petroleum Reserves (12), OECD Reserves (13), OPEC Reserves (14), China Reserves (15), U.S Refinery Capacity (18), U.S Dollar Exchange Rate Index (19), U.S Gross Domestic Product (24), Consumer Price Index (25) and Producer Price Index-Petroleum (26).

Stage 3 initialized with maximum relevant rank variable EPPI (16). Since the mutual information $I(X_{16}, X_{13}) > Th2$, therefore, X_{13} is filtered out by redundancy filter. The process run iteratively for next ranked variables as discussed earlier. The final selected features from the proposed I^2MI^2 algorithm are Non-OPEC Production (3), OPEC Supply (4), China Reserves (15), U.S. Dollar Exchange Rate (19), Consumer Price Index (25) and Producer Price Index (26). The study used these selected features as input variables to various forecasting engines.

4.4.4 Numerical Results

Forecasting Results for MI^3 Algorithm

The performance of proposed feature selection algorithm is compared with Correlation based Feature Selection (CFS), Modified Relief (MR) and Modified Relief + Mutual Information (MR + MI) feature selection methods. The performance criterion used for comparing MI^3 algorithm with other algorithms are RMSE, MAE and MAPE. The performance criterion are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(WTI_{Actual} - WTI_{Predicted})^2} \tag{4.7}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|WTI_{Actual} - WTI_{Predicted}| \tag{4.8}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|WTI_{Actual} - WTI_{Predicted}|}{WTI_{Actual}} \tag{4.9}$$

The performance criterion needed for comparison are represented in Table 4.10. The results provided the following observations:

Table 4.10: Performance criterion for comparing MI^3 with different feature selection methods

Models	**RMSE**	**MAE**	**MAPE**
$MI^3 + CNN$	3.77	2.80	6.76
$MI^3 + GRNN$	1.61	0.82	2.01
$MI^2 + MLP$	3.43	2.43	5.77
$(MR + MI) + CNN$	3.31	2.5	5.82
$(MR + MI) + GRNN$	1.32	0.94	2.55
$(MR + MI) + MLP$	3.79	2.95	7.18
$MR + CNN$	3.23	2.33	5.41
$MR + GRNN$	1.39	0.95	2.63
$MR + MLP$	4.3	3.09	7.46
$CFS + CNN$	3.46	2.61	6.39
$CFS + GRNN$	1.93	1.4	3.76
$CFS + MLP$	4.04	2.99	6.76

- MI^3 algorithm with GRNN performed best in terms of all three performance criterion followed by $(MR + MI)$ and CFS based feature selection together with GRNN model.
- MLP has performed worst in comparison to GRNN and CNN for all feature selected from different methods.

The explanatory power for oil prices using fifteen selected features is 98.02% indicating that the variable reduction is reasonable and will have no essential influence on subsequent analysis. The proposed algorithm is able to select most relevant and non-redundant features from Group-B that performs better in deriving directions of oil prices. Fig 4.9 describe the explanatory power of each factor selected by proposed MI^3 algorithm and GRNN as forecasting engine. It shows that economic players such as producer price index and consumer price index played a significant role in driving oil prices compared to demand-supply factors. Exchange market is also playing an important role together with strategic petroleum reserves. U.S GDP is playing major role in long run and is influencing oil prices directly.

Forecasting Results for I^2MI^2 Algorithm

The performance of proposed feature selection algorithm is compared with Correlation based Feature Selection (CFS), Modified Relief (MR) and Modified Relief + Mutual Information (MR + MI) feature selection methods. The performance criterion used for comparing I^2MI^2 algorithm with other

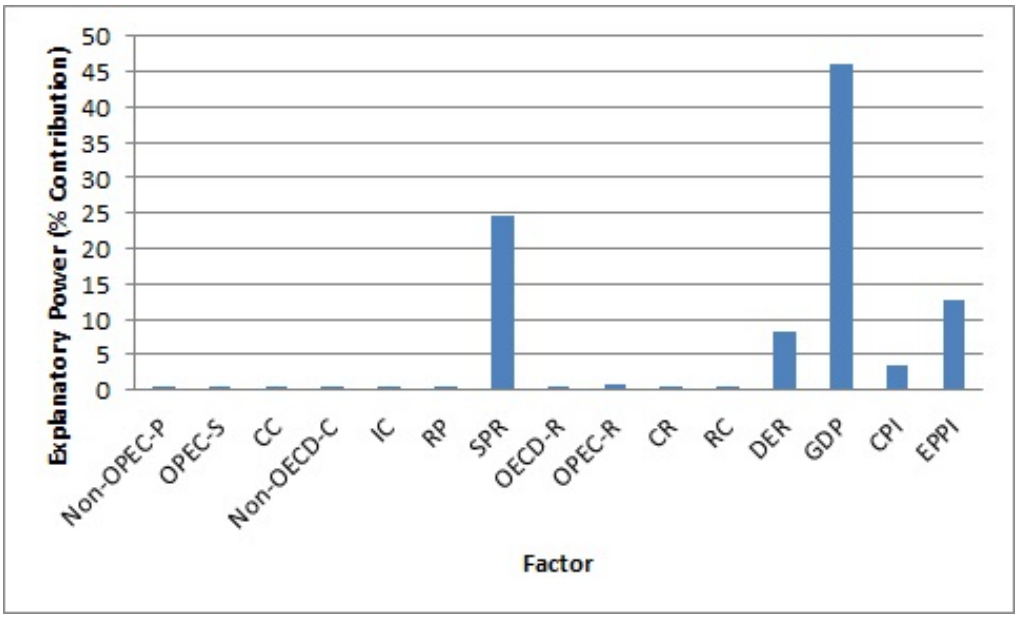


Figure 4.9: Explanatory power of 15 selected factors based on MI^3 Algorithm

algorithms are RMSE, MAE and MAPE. The performance criterion are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(WTI_{Actual} - WTI_{Predicted})^2} \tag{4.10}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|WTI_{Actual} - WTI_{Predicted}| \tag{4.11}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|WTI_{Actual} - WTI_{Predicted}|}{WTI_{Actual}} \tag{4.12}$$

The performance criterion needed for comparison are represented in Table 4.11. The results provided the following observations:

- The proposed algorithm I^2MI^2 with GRNN as forecasting engine has performed the best among all other feature selection methods. I^2MI^2 + GRNN model have lowest RMSE, MAE and MAPE as 1.29, 0.96 and 2.51 respectively.
- The reason for the best performance lies in the fact that the final selected features from proposed algorithm are 100% non-redundant and relevant for the study.
- Two stage (MR + MI) with CNN as forecasting engine as proposed

Table 4.11: Performance criterion for comparing I^2MI^2 with different feature selection methods

Models	**RMSE**	**MAE**	**MAPE**
$I^2MI^2 + CNN$	3.06	2.37	6.86
$I^2MI^2 + GRNN$	1.29	0.96	2.51
$I^2MI^2 + MLP$	4.35	3.19	7.41
$(MR + MI) + CNN$	3.31	2.5	5.82
$(MR+MI)+GRNN$	1.32	0.94	2.55
$(MR + MI) + MLP$	3.79	2.95	7.18
$MR + CNN$	3.23	2.33	5.41
$MR + GRNN$	1.39	0.95	2.63
$MR + MLP$	4.3	3.09	7.46
$CFS + CNN$	3.46	2.61	6.39
$CFS + GRNN$	1.93	1.4	3.76
$CFS + MLP$	4.04	2.99	6.76

by Amjady and Daraeepour [117] has not performed better than proposed algorithm.

- The single stage feature selection algorithms with MLP as forecasting engine have not performed well for complex oil price data. The reason behind their poor performance is the existence of redundancy in selected feature set.

The study propose I^2MI^2 algorithm based selected features as input to GRNN forecasting engine to forecast crude oil prices. The proposed model is used to forecast in-sample and out-of-sample. Firstly, in order to compare the model's capability with other models, nearly 17-year (January 1995-November 2012) monthly data is used for training and validation, while the twelve-month ahead data from December 2012-November 2013 is used as testing sample. The model is used to produce one and twelve-month out-of-sample forecasts from November 2012 till November 2013. The forecasting procedure for one-month ahead prediction begins by training the model with data from January 1995 to November 2012, and forecasting the value of December 2012. Then actual value of December 2012 was added, and the model is refitted to forecast the price of January 2013. The process is repeated till the value of November 2013 is forecast. The one-month ahead forecast from the proposed methodology is compared with monthly forecasts shown in EIA's STEO reports from December 2012 till November 2013. The twelve-month ahead forecast from the proposed methodology is compared with forecast shown in EIA's STEO December 2013 report.

Out-of sample evaluations are shown in Table 4.12. The results proved the superiority of proposed methodology for long-run (twelve-month ahead) time period in comparison to EIA's STEO forecasts as reported. The proposed methodology performed more accurately in long-run forecasting as compared to short-run. It is clear that our proposed model is able to identify true drivers of oil prices and uses most relevant and non-redundant features as input to GRNN for supreme performance.

Table 4.12: Out-of-sample forecast comparison

Comparison Models	**RMSE**	**MAE**	**MAPE**
One-month(Proposed)	5.3	4.25	4.36
One-month(STEO)	3.51	2.8	2.90
Twelve-month(Proposed)	7.13	5.91	6.02
Twelve-month(STEO)	9.62	7.81	7.74

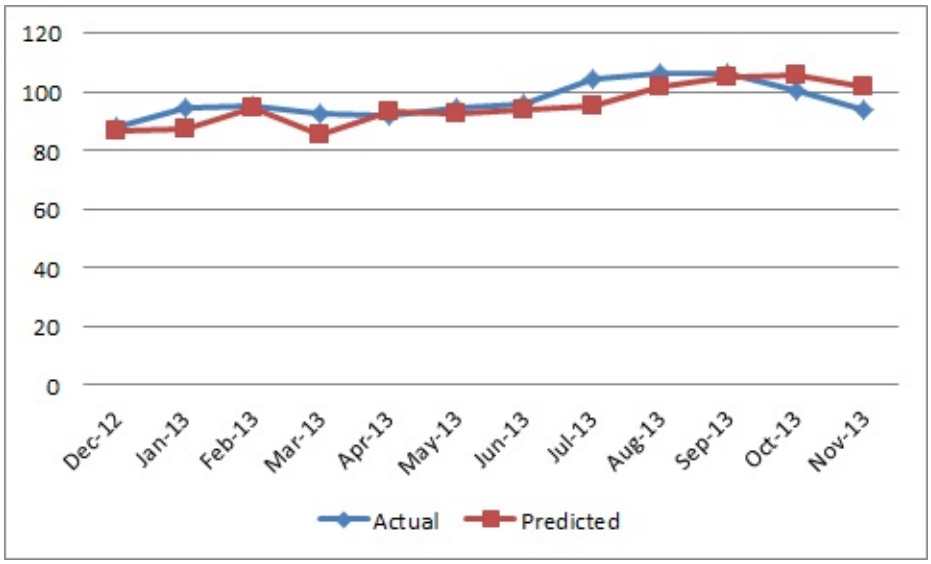


Figure 4.10: One-month out-of-sample forecast

The forecasts provided by the proposed methodology performed significantly better than EIA's STEO forecasts for twelve-month ahead compared to EIA's STEO December 2012 report forecasts.

Finally, the proposed methodology has provided the minimal set of input variables that performed significantly well in predicting oil prices compared to STEO forecasts.

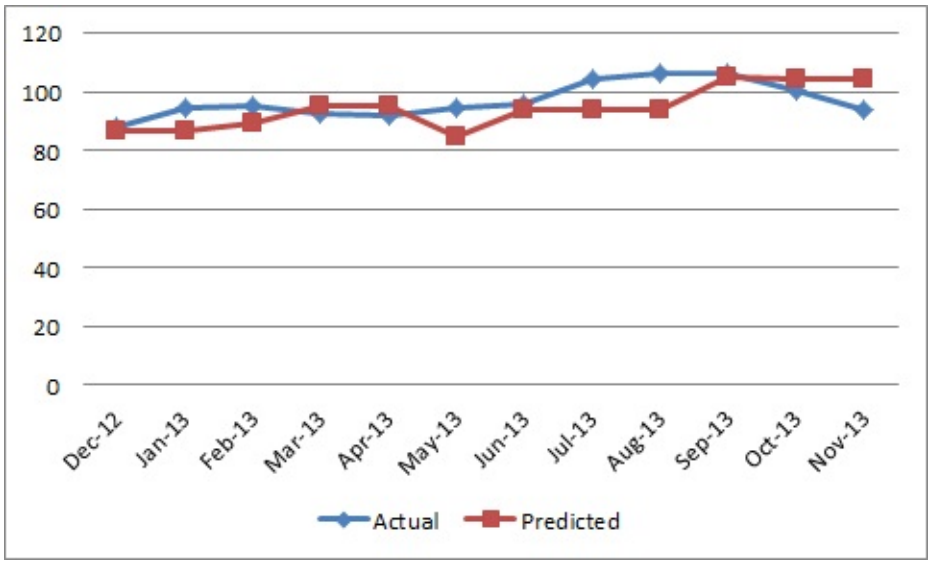


Figure 4.11: Twelve-month out-of-sample forecast

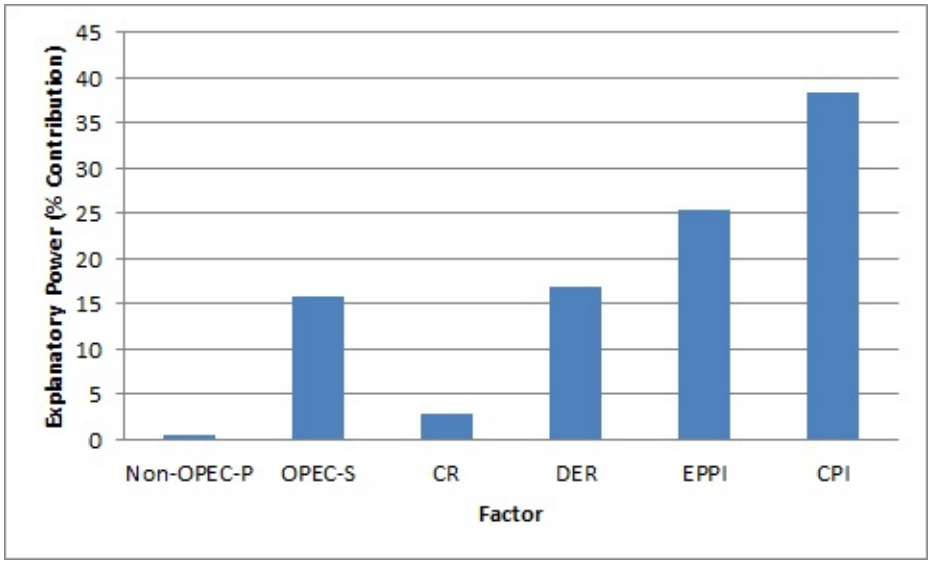


Figure 4.12: Explanatory power of selected features using proposed methodology for Group-B

The results showed that the features selected by proposed I^2MI^2 algorithm when used as input variables provide better forecasting performance for oil prices. Fig 4.10 shows the actual and predicted one-month ahead crude oil price from November 2012- November 2013 while Fig 4.11 shows the actual and predicted twelve-month ahead crude oil price for the same period. The explanatory power for oil prices using six selected features is 98.82% indicating that the variable reduction is reasonable and that it will have no essential influence on subsequent analysis. Fig 4.12 describe the explanatory power of each factor selected by proposed I^2MI^2 algorithm and GRNN as forecasting engine.

4.5 Concluding Remarks

Crude oil prices have been historically governed by uncertain political, economic and financial indicators. There is no single indicator which can provide a complete picture on how prices can be determined. Nor a simple combination of input indicators can provide accurate and robust price forecasts. In particular, feature selection plays a key role in identifying key drivers of oil prices to achieve high prediction performance. Due to lack of competent feature selection techniques based on associations and dependencies of indicators, two novel feature selection methods called MI^3 and I^2MI^2 algorithm are proposed for inferring non-linear dependence between oil prices and strategic indicators driving them. These algorithms are build on the pillars of interaction information and mutual information as measure of redundancy(or synergy) and relevance.

Experiments on both groups shows that MI^3 algorithm eliminate more than $\frac{1}{2}$ of the number of features. The proposed algorithms quickly identifies most relevant and non-redundant features for the study. Many researchers have argued that structural models fail to forecast oil prices due to non-availability of forecast value of right hand side indicators. Therefore, there is need to refine MI^3 algorithm so as to provide minimal set of most relevant and non-redundant features. This limitation of MI^3 is overcome by introducing an extended version I^2MI^2 algorithm consisting of three stages. Experiments on both groups shows that I^2MI^2 algorithm eliminates more than $\frac{1}{4}$ of the number of features.

The selected features from proposed algorithm are used as input to three

forecasting engines: Multi-Layered Perceptron (MLP) , General Regression Neural Network (GRNN) and Cascade Correlation Neural Network (CNN). The study is designed to examine input variables as defined in Group A and Group B to identify the core components driving crude oil prices from decades. The proposed algorithm is been examined on WTI spot prices and compared with some of the known feature selection methods. The performance criterion used for comparing proposed I^2MI^2 algorithm with other methods are RMSE (Root Mean Square Error), MAE(Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). The performance criterion reveals improved forecasting ability of features selected by I^2MI^2 algorithm. This algorithm is superior in comparison to other feature selection methods as it not only assesses non-linear dependencies but also provides 100% relevant and non-redundant (synergy) features for the study. The explanatory power for oil prices using selected features by proposed I^2MI^2 algorithm is above 98% for both groups, indicating that the variable reduction is reasonable and that it will have no essential influence on subsequent analysis. The results show emerging economies as an influential factor driving oil prices. The study examined shift in influence of OECD consumption to Non-OECD consumption as a key indicator driving oil prices. The results indicated that the economic players play a significant role in driving oil prices as compared to demand-supply framework. Exchange market and Strategic Petroleum Reserves also play a leading role that drive oil prices. U.S GDP is a major player in long run for influencing oil prices.

Chapter 5

Aftermath of 2008 Financial Crisis

5.1 Overview

In this chapter, I^2MI^2 algorithm based feature selection approach is applied to identify the main factors driving oil prices before and after 2008 global financial crisis. The selected features before and after 2008 financial crisis are used as input to neural networks that act as forecasting engines. A thematic literature review is presented to provide an insight on studies related to after-effects of extreme events on oil prices. This chapter is focus on analysing explanatory power of factors and their contribution in driving oil prices during the booming and downturn period. Further, the study evaluates the selected factors for predicting oil prices for one-month ahead and twelve-month ahead forecast horizon. The forecasts from the proposed methodology are compared with EIA's STEO January 2013 onwards forecast reports.

5.2 Literature Review

According to Energy Information Administration (EIA), geopolitical and economic events had strong impact on crude oil markets for over 40 years. Oil prices rose to $30 per barrel by end of 1996 but due to Asian financial crisis in quarter 1 of 1997, prices declined drastically and were at $16 per barrel by end of 1998. The impact of Asian financial crisis on the behaviour of international oil prices was enormous. Asian financial crisis has weakened the strong long-run relationship between spot prices and future

prices, complicating the objective of predicting movement of oil prices [122]. Olowe [123] investigated the month-of-the-year effect in the crude oil market using GARCH models and showed that the Asian financial crisis has an impact on oil price return series. Charles and Darné [124] suggested to take into account these geopolitical and economic events to improve modelling volatility of prices.

Further, as a consequence of OPEC cuts in production targets by 1.7 mmbpd, oil prices increased to more than $35 per barrel in late 2000. Kohl [125] examined OPEC behaviour from 1998-2001 and concluded that political factors are found to be important in OPECs shift from in-strategy market share to target price in 1999. Guidi et al. [126] showed that OPEC policy decisions influence oil markets, leading to oil price shocks which further affect market returns. Lin and Tamvakis [127] showed OPEC announcements do effect oil price expectation and volatility. Schmidbauer and Rösch [128] fitted GARCH model to crude oil price returns where dummy variable indicated the day of an OPEC announcement. The results showed that there is a post-announcement effect on return expectations and pre-announcement effect on return volatility. Researchers [129] [130] [131] shows that OPEC production decisions are important and its influence has evolved through time leading to change in oil price system. The impact of such extreme events is of prime importance as they had affected the objective of increasing the predication accuracy in crude oil prices. It is important to identify key factors that drive crude oil prices during the time frame of happening of such extreme event so that structural forecasting models can be designed in a better way.

In quarter 3 of 2001, when oil prices were around $34 per barrel, a terrorist activity (9/11 attack) led to an increase in volatility of oil prices, that soared oil prices above $54 per barrel in late 2004. International stock market had experienced large (permanent or temporary) shocks in response to September 11, 2001, terrorist attacks. Further, oil prices steadily rose for several years and in July 2008 stood at a record high of $145 per barrel due to low spare capacity. The question that arises is whether this rise in oil price is due to squeeze in availability or are there any other political or economic indicators to blame? Further, due to global financial crisis of 2008, oil prices plunged to around $43 per barrel by end of 2008. In quarter 1 of 2009, OPEC slashed production targets by 4.2 mmbpd and thus oil

prices rose from $43 per barrel to $91 per barrel by end of 2011.

One of the reasons for strong shift in oil price during 2009 is the effects of shocks to liquidity in the BRIC countries [89]. Bhar and Malliaris [132] concluded that price increases during financial crisis of 2007-2009 were so substantial that additional factors other than demand and supply were needed to explain such drastic shifts. Fan and Xu [133] used break test to divide the price fluctuations in oil markets after 2000 into three stages: January 2000–March 2004, March 2004–June 2008 and June 2008–September 2009. The study has shown that in different time periods, the main drivers of oil prices changed and their direction and degree of influence will change over time. Chai et al [45] showed that U.S dollar index remains an important factor of oil prices but its impact was strengthened after the financial crisis.

Oil prices are dependent on numerous macroeconomic indicators but there effects are subject to structural changes [134]. Oil prices have increased drastically since 2004 and reached the peak in July 2008. Consequently oil prices began to drop sharply. This chapter focuses on finding the main drivers of oil price before and after global financial crisis (quarter 3, 2008). The different mechanism driving oil prices during falling and rising period are analysed to provide an insight into the explanatory power of factors for oil price trend and their contribution to oil price volatility. Existing literature has accounted for non-linearity, non-stationary and time-varying structure of the oil prices but seldom focused on selecting significant features with high predicting power subject to structural change. There is a need to select appropriate features/ indicators explaining the characteristics of oil markets during booming and downturn period. In particular, feature selection plays an important role in determining key drivers of oil prices for each sub-period of structural change. The empirical literature is very far from consensus as to how, why and to what extent these macroeconomic variables drive oil prices subject to structural change. In empirical literature, the design of input vector of oil price forecast model was carried out on judgemental criteria or trial and error procedures. To overcome above mention research gaps, this chapter focuses on designing the input vector for oil price forecasting before and after financial crisis using I^2MI^2 algorithm and general regression neural networks.

5.3 Numerical Results

For analysing the different mechanism in the falling and rising period of oil prices, two sub-periods are considered for Group-B: January 2004-July 2008 and August 2008-December 2012, before and after 2008 financial crisis, respectively. For each sub-period, I^2MI^2 algorithm is applied to select minimal set of relevant and non-redundant factors that leads to high prediction performance for oil prices. Neural network models are used as forecasting engines to analyse the explanatory power of selected features and their contribution in driving oil prices. The proposed methodology is used to forecast the new characteristics of oil prices one-month and twelve-month ahead before and after the crisis. The step by step procedure to find the set of relevant and non-redundant features driving oil prices are as follows.

5.3.1 Sub-period 1: January 2004-July 2008

The correlation coefficient between WTI prices and all factors is shown in Table 5.1. It revealed the existence of significant relationship between factors selected for the study before 2008 financial crisis. The step by step procedure of proposed I^2MI^2 algorithm used for sub-period 1 are explained below.

Stage 1 The candidate features (column 1) with the relevance rank (column 2) and their normalized relevance rank value (column 3) with the respect to maximum mutual information with oil prices are shown in Table 5.2. Column 4 provides the feature number. The goal of stage one is to provide relevant features based on mutual information irrelevance filter. Based on a low threshold value $Th1$, variables I-OPEC, OECD-C, Non-OPEC-P and CR can be filtered out by relevance filter. There is no rule of thumb for fixing threshold value. Based on judgemental criterion, one can set up threshold value. The relevant features are then added to set S_1.

Stage 2 The goal of stage two is to provide non-redundant and relevant features based on redundancy filter. The three-variable interaction information between target variable and features selected from set S_1 is computed. For better illustration, Table 5.3 provides the list of pair of variables for which $I(Y, X_i, X_j) < 0$. The results obtained from the redundancy filter are shown in column 1-8.

Table 5.1: Correlation coefficient for Group-B before and after crisis

Feature Code	**Correlation-Before Crisis**	**Correlation-After Crisis**
WTI	1.0000**	1.0000**
NCPP	0.8975**	0.5331**
Non-OPEC-P	0.1858	0.5021**
OPEC-S	0.7568**	0.7066**
OECD-C	−0.3597**	−0.2487
CC	0.6518**	0.5970**
Non-OECD-C	0.8078**	0.6344**
IC	0.5208**	0.4351**
OSC	−0.0400	−0.3708**
OPS	0.4242**	−0.5168**
RP	0.8437**	0.4524**
SPR	0.7682**	−0.1870
OECD-R	−0.7991**	0.2564
OPEC-R	0.8513**	0.6218**
CR	−0.6751**	0.5835**
I-OPEC	0.3900**	−0.3385*
I-Non-OPEC	−0.2821*	−0.4070**
RC	0.8342**	−0.2637
DER	−0.8921**	−0.6938**
GU	0.8196**	−0.5506**
JU	−0.3220*	0.5286**
EU	0.7784**	−0.5576**
DJI	0.9427**	0.8383**
GDP	0.7942**	0.7215**
CPI	0.9129**	0.7518**
EPPI	0.9718**	0.9291**

Notes:
* Significant at 5% level.
** Significant at 1% level.

Table 5.2: Relevance rank based on stage one of proposed algorithm for sub-period 1: January 2004-July 2008

Candidate Feature	**Relevance Rank**	**Normalized Relevance Rank**	**Feature No.**
EPPI	1	1.000000	26
CPI	2	0.991436	25
DJI	3	0.958593	23
NCPP	4	0.891660	2
GDP	5	0.850607	24
SPR	6	0.777306	12
GU	7	0.733397	20
Non-OECD-C	8	0.722326	7
EU	9	0.708129	22
DER	10	0.677827	19
RP	11	0.670974	11
OPEC-R	12	0.670974	14
RC	13	0.654327	18
OECD-R	14	0.622873	13
OPS	15	0.611521	10
CC	16	0.604156	6
OSC	17	0.568176	9
OPEC-S	18	0.513150	4
IC	19	0.500051	8
JU	20	0.500051	21
I-Non-OPEC	21	0.467731	17
I-OPEC	22	0.406204	16
OECD-C	23	0.376868	5
Non-OPEC-P	24	0.267796	3
CR	25	0.207891	15

Since interaction information $I(Y, X_i, X_j)$ is a symmetric measure; it cannot derive the direction whether X_j inhibits the correlation between (Y, X_i) or X_i inhibits the correlation between (Y, X_j) . Therefore, it become difficult to filter the redundant variable from the set of relevant features (X_i, X_j) when $I(Y, X_i, X_j) < 0$. In this thesis, this limitation of interaction information is relieved by focusing on mutual information between target and input variables $I(Y, X_i)$. The algorithm in stage two starts with maximum relevance rank variable X_{imax} from stage one. The variable EPPI(26) is ranked first as evident from Table 5.2. Add X_{26} to set S_2. For the first relevance ranked variable X_{26}, there are seven set $\{Y, X_{26}, X_j\}$ where $j = \{3, 4, 5, 8, 16, 17, 21\}$ for which interaction information is negative i.e. $I(Y, X_i, X_j) < 0$. The question that arises here is whether X_j inhibits the correlation between Y and X_{26} or X_{26} inhibits the correlation between Y and X_j. The redundant variable is filtered by comparing mutual information $I(Y, X_{26})$ with $I(Y, X_j)$ for each j. The results thus obtained in Table 5.2 show that mutual information $I(Y, X_{26}) > I(Y, X_j)$ for each j. Therefore, the variables $X_j \; \forall \; j = \{3, 4, 5, 8, 16, 17, 21\}$ are redundant variables and must be filtered out from the list of relevant and non-redundant variables.

Similarly, the process holds for next ranked variable X_{25} from Table 5.2. There are five set of features having interaction information $I(Y, X_{25}, X_j) < 0$ where $j = \{3, 5, 8, 16, 17\}$. Since mutual information $I(Y, X_{25}) > I(Y, X_j)$ $\forall \; j = \{3, 5, 8, 16, 17\}$, therefore, X_{25} is added to set S_2 and X_j where $j = \{3, 5, 8, 16, 17\}$ are filtered out by redundancy filter. Table 5.4 show the list of relevant non-redundant features selected from stage one & two. The number of candidate inputs (N) are reduced from 25 to 11 in stage two; i.e. to less than 50% of the actual number of input variables. The set of variables in Table 5.4 has been cross-validated for three-variable interaction information such that $I(Y, X_i, X_j) > 0$ always. The set S_2 is the set of features selected by stage two of the proposed algorithm. Since three-variable interaction information has provided features for which $I(Y, X_i, X_j) > 0$, therefore, higher order interaction information is not required.

Stage 3 The algorithm in stage three starts with maximum relevance rank variable X_{26} from Table 5.2. By default, X_{26} is considered as part of final set S_3. Now, consider the next relevance rank feature X_{25}. According to

Table 5.3: List of pair of variables having negative interaction information

Column 1		Column 2		Column 3		Column 4		Column 5		Column 6		Column 7		Column 8	
2	3	4	20	6	24	9	18	12	22	16	19	19	6	22	4
2	4	4	21	6	25	9	19	13	3	16	20	19	8	22	5
2	5	4	22	7	3	9	20	13	4	16	21	19	9	22	6
2	6	4	24	7	4	9	21	13	5	16	22	19	10	22	7
2	8	4	26	7	5	9	22	13	6	16	23	19	16	22	8
2	16	5	2	7	6	9	23	13	8	16	24	19	17	22	10
2	17	5	3	7	8	10	3	13	10	16	25	19	18	22	12
2	21	5	4	7	9	10	4	13	16	16	26	19	20	22	13
3	2	5	6	7	10	10	5	13	17	17	2	19	21	22	15
3	4	5	7	7	16	10	6	13	20	17	3	19	22	22	16
3	5	5	8	7	17	10	7	13	21	17	4	20	4	22	17
3	6	5	9	7	20	10	8	13	22	17	5	20	5	22	18
3	7	5	10	7	21	10	9	14	3	17	6	20	6	22	19
3	8	5	11	7	22	10	11	14	4	17	7	20	7	22	21
3	9	5	12	8	2	10	12	14	5	17	8	20	8	23	3
3	10	5	13	8	3	10	13	14	6	17	9	20	9	23	5
3	11	5	14	8	4	10	14	14	8	17	10	20	10	23	6
3	12	5	15	8	5	10	15	14	10	17	11	20	12	23	8
3	13	5	16	8	6	10	16	14	16	17	12	20	13	23	9
3	14	5	17	8	7	10	17	14	17	17	13	20	15	23	10
3	15	5	18	8	9	10	18	14	21	17	14	20	16	23	16
3	16	5	19	8	10	10	19	15	3	17	15	20	17	23	17
3	17	5	20	8	11	10	20	15	4	17	16	20	18	23	21
3	18	5	21	8	12	10	21	15	5	17	18	20	19	24	3
3	19	5	22	8	13	10	22	15	6	17	19	20	21	24	4
3	20	5	23	8	14	10	23	15	8	17	20	21	2	24	5
3	21	5	24	8	15	10	24	15	9	17	21	21	3	24	6
3	22	5	25	8	16	11	3	15	10	17	22	21	4	24	8
3	23	5	26	8	17	11	4	15	12	17	23	21	5	24	10
3	24	6	3	8	18	11	5	15	16	17	24	21	6	24	16
3	25	6	4	8	19	11	6	15	17	17	25	21	7	24	17
3	26	6	5	8	20	11	8	15	20	17	26	21	8	24	21
4	2	6	7	8	21	11	10	15	21	18	3	21	9	25	3
4	3	6	8	8	22	11	16	15	22	18	4	21	10	25	5
4	5	6	9	8	23	11	17	16	2	18	5	21	11	25	8
4	6	6	10	8	24	11	21	16	3	18	6	21	12	25	16
4	7	6	11	8	25	12	3	16	4	18	8	21	13	25	17
4	8	6	12	8	26	12	4	16	5	18	9	21	14	26	3
4	9	6	13	9	3	12	5	16	6	18	10	21	15	26	4
4	10	6	14	9	4	12	6	16	7	18	12	21	16	26	5
4	11	6	15	9	5	12	8	16	8	18	16	21	17	26	8
4	12	6	16	9	6	12	9	16	9	18	17	21	18	26	16
4	13	6	17	9	7	12	10	16	10	18	19	21	19	26	17
4	14	6	18	9	8	12	15	16	12	18	20	21	20	26	21
4	15	6	19	9	10	12	16	16	13	18	21	21	22		
4	16	6	20	9	12	12	17	16	14	18	22	21	23		
4	17	6	21	9	15	12	18	16	15	19	3	21	24		
4	18	6	22	9	16	12	20	16	17	19	4	21	26		
4	19	6	23	9	17	12	21	16	18	19	5	22	3		

the pre-specified threshold value $Th2$, variables from set S_2 are filtered out based on mutual information between features. Since mutual information $I(X_{26}, X_{25}) > Th2$, therefore, X_{25} is filtered out by redundancy filter. For the next relevant ranked feature X_m, calculate maximum mutual information $Max(MI)$ between X_m and previously selected candidates in set S_3 by redundancy filter. If $Max(MI) > Th2$ for any set, then X_m is filter out by redundancy filter. Otherwise, X_m is added to the final selected features set S_3. The algorithm will run iteratively for all 11 selected variables from stage two.

The final selected features from the proposed I^2MI^2 algorithm are EPPI (26), NCPP (2), SPR (12), DER (19) and RP(11). Thus, five out of twenty five variables were selected to represent fluctuations in oil prices before the crisis. The selected features are used as input variables to neural networks forecasting engines.

Table 5.4: Filtered features by redundancy filter in stage two of proposed algorithm for sub-period 1

Filtered feature by redundancy filter	**Feature No.**	**Feature Rank**
EPPI	26	1
CPI	25	2
DJI	23	3
NCPP	2	4
GDP	24	5
SPR	12	6
Non-OECD-C	7	8
DER	19	10
RP	11	11
OPEC-R	14	12
OECD-R	13	14

Forecasting Results

To forecast crude oil prices, the study propose I^2MI^2 based features selected as input to GRNN forecasting engine. The performance of proposed feature selection algorithm with GRNN forecasting engine is evaluated based on RMSE, MAE and MAPE. The proposed ensemble model is used to forecast in-sample and out-of-sample. Firstly, in order to compare the model's capability with other models, nearly 4.4-year (January 2004-July 2008) monthly data is used for training and validation. In-sample evaluation are shown in Table 5.5. The model is used to produce one

Table 5.5: In-sample performance of proposed methodology

Model	**RMSE**	**MAE**	**MAPE**
$I^2MI^2 + GRNN$	3.55	2.74	4.13

and twelve-month ahead out-of-sample forecasts from August 2008 till July 2009. To evaluate the performance of our model, we compare it with forecasts shown in EIA's STEO reports from August 2008 onwards. Out-of sample evaluations are shown in Table 5.6. The proposed methodology performed better in terms of MAE for one-month ahead forecasts as compared to EIA's STEO forecasts but not in terms on RMSE and MAPE. Fig 5.1 shows one-month ahead out-of-sample forecasts of proposed methodology. It is evident from Table 5.6 that the proposed model performed superior as compared to STEO model for twelve-month ahead forecasts during

extreme complex and volatility phase of oil prices. Fig. 5.2 shows twelve-month ahead out-of-sample forecasts of proposed methodology and it also shows that the model does very well based on input variables selected by proposed algorithm as compared to EIA's STEO forecasts. The proposed methodology performed more accurately in long-run forecasting as compared to short-run when the market is too complex and highly volatile. The explanatory power for oil prices using five selected features is 97.6% before the crisis, indicating that the variable reduction is reasonable and that it will have no essential influence on subsequent analysis. The results show that the features selected by proposed I^2MI^2 algorithm when used as input variables provide better forecasting performance for oil prices. Fig 5.3 describe the explanatory power of each factor selected by proposed I^2MI^2 algorithm and GRNN as forecasting engines.

Table 5.6: Out-of-sample forecast comparison

Model	**RMSE**	**MAE**	**MAPE**
One-month(Proposed)	8.24	9.74	13.27
One-month(STEO)	6.85	9.91	10.82
Twelve-month(Proposed)	31.9	34.85	63.30
Twelve-month(STEO)	67.59	62.49	122.81

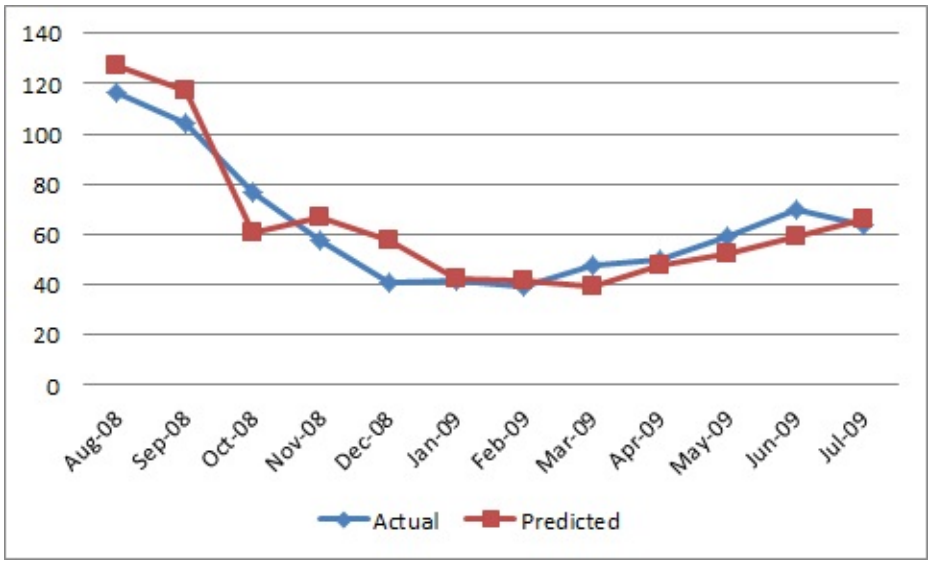


Figure 5.1: One-month out-of-sample forecast

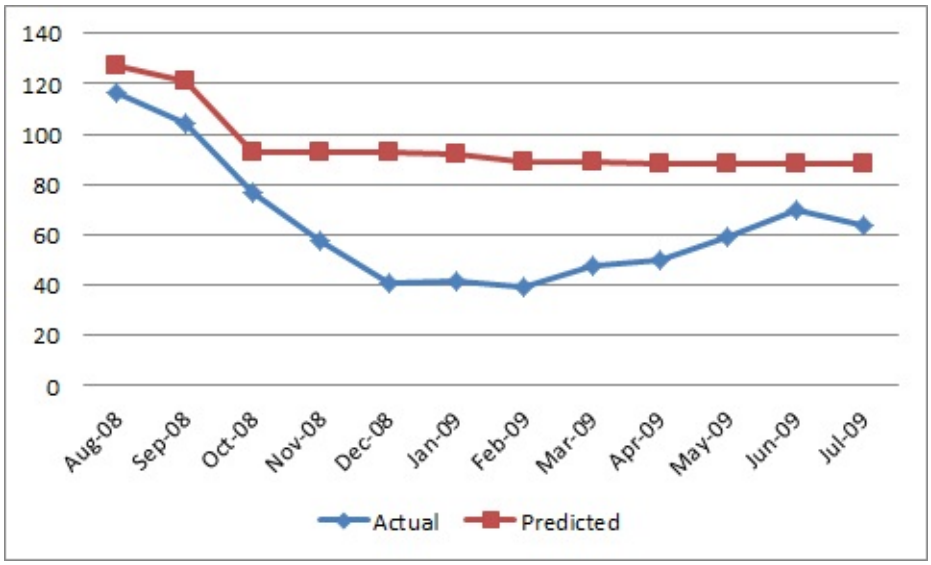


Figure 5.2: Twelve-month out-of-sample forecast

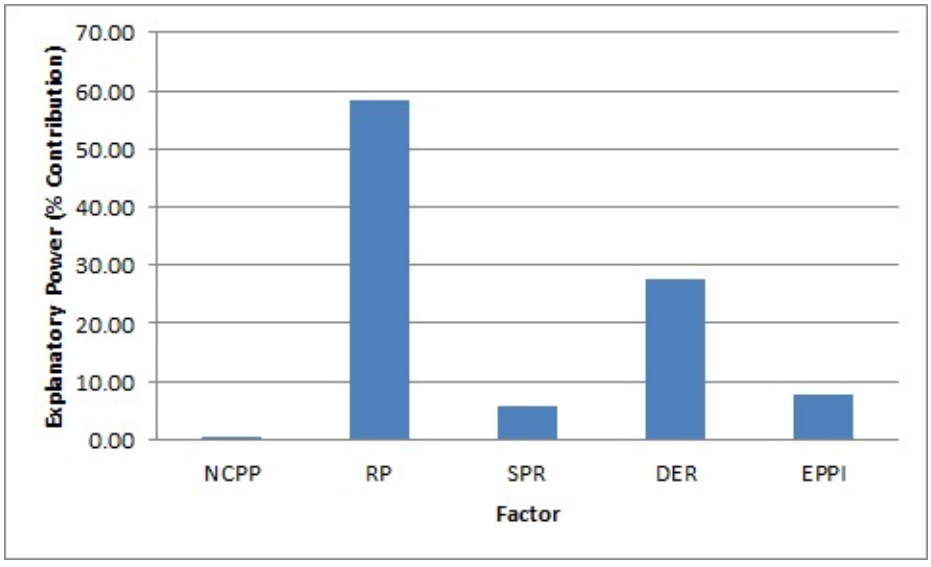


Figure 5.3: Explanatory power of 5 selected features before crisis

5.3.2 Sub-period 2: August 2008-November 2012

The correlation coefficient between WTI prices and all factors is shown in Table 5.1. It revealed the existence of significant relationship between factors selected for the study after 2008 financial crisis. The results of three stages of proposed I^2MI^2 algorithm are explained below.

Stage 1 The candidate features (column 1) with the relevance rank (column 2) and their normalized relevance rank value (column 3) with the respect to maximum mutual information with oil prices are shown in Table 5.7. Column 4 provides the feature number. The goal of stage one is

Table 5.7: Relevance rank based on stage one of proposed algorithm for sub-period 2: August 2008-November 2012

Candidate Feature	**Relevance Rank**	**Normalized relevance value**	**Feature No.**
EPPI	1	1	26
CPI	2	0.870164	25
DJI	3	0.814552	23
CC	4	0.692591	6
Non-OECD-C	5	0.692591	7
GDP	6	0.692591	24
IC	7	0.65298	8
OECD-R	8	0.645248	13
OPEC-S	9	0.614923	4
SPR	10	0.612197	12
OPEC-R	11	0.601805	14
RC	12	0.591455	18
OSC	13	0.578302	9
JU	14	0.578302	21
RP	15	0.543014	11
EU	16	0.543014	22
NCPP	17	0.508963	2
Non-OPEC-P	18	0.508963	3
I-Non-OPEC	19	0.508963	17
OPS	20	0.476068	10
DER	21	0.476068	19
GU	22	0.476068	20
OECD-C	23	0.383585	5
CR	24	0.365721	15
I-OPEC	25	0.354618	16

to provide relevant features based on mutual information irrelevance filter. Based on a low threshold value $Th1$, variables OECD-C, CR and I-OPEC

Table 5.8: List of pair of variables having negative interaction information

Column 1		Column 2		Column 3		Column 4		Column 5		Column 6		Column 7		Column 8	
2	3	4	7	7	4	10	11	14	20	17	19	20	8	22	20
2	4	4	8	7	5	10	12	14	22	17	20	20	9	22	21
2	5	4	9	7	9	10	13	15	2	17	21	20	10	22	23
2	6	4	10	7	10	10	14	15	3	17	22	20	11	22	24
2	7	4	12	7	16	10	15	15	5	17	23	20	12	22	25
2	8	4	16	7	17	10	16	15	9	17	24	20	13	22	26
2	9	4	17	7	19	10	17	15	10	17	25	20	14	23	2
2	10	4	18	7	20	10	18	15	16	17	26	20	15	23	3
2	11	4	19	7	22	10	19	15	17	18	2	20	16	23	4
2	12	4	20	7	23	10	20	15	19	18	3	20	17	23	5
2	13	4	21	8	2	10	21	15	20	18	4	20	18	23	7
2	14	4	22	8	3	10	22	15	22	18	5	20	19	23	8
2	15	4	23	8	4	10	23	16	2	18	8	20	21	23	9
2	16	5	2	8	5	10	24	16	3	18	9	20	22	23	10
2	17	5	3	8	9	10	25	16	4	18	10	20	23	23	16
2	18	5	4	8	10	10	26	16	5	18	12	20	24	23	17
2	19	5	6	8	16	11	3	16	6	18	16	20	25	23	19
2	20	5	7	8	17	11	5	16	7	18	17	20	26	23	20
2	21	5	8	8	18	11	10	16	8	18	19	21	2	23	21
2	22	5	9	8	19	11	16	16	9	18	20	21	3	23	22
2	23	5	10	8	20	11	17	16	10	18	21	21	4	24	2
2	24	5	11	8	21	11	19	16	11	18	22	21	5	24	3
2	25	5	12	8	22	11	20	16	12	19	2	21	6	24	5
2	26	5	13	8	23	11	22	16	13	19	3	21	8	24	6
3	2	5	14	8	24	12	2	16	14	19	4	21	9	24	8
3	4	5	15	9	2	12	3	16	15	19	5	21	10	24	9
3	5	5	16	9	3	12	4	16	17	19	6	21	16	24	10
3	6	5	17	9	4	12	5	16	18	19	7	21	17	24	16
3	7	5	18	9	5	12	9	16	19	19	8	21	18	24	17
3	8	5	19	9	6	12	10	16	20	19	9	21	19	24	19
3	9	5	20	9	7	12	16	16	21	19	10	21	20	24	20
3	10	5	21	9	8	12	17	16	22	19	11	21	22	24	21
3	11	5	22	9	10	12	18	16	23	19	12	21	23	24	22
3	12	5	23	9	12	12	19	16	24	19	13	21	24	25	2
3	13	5	24	9	15	12	20	16	25	19	14	22	2	25	3
3	14	5	25	9	16	12	22	16	26	19	15	22	3	25	5
3	15	5	26	9	17	13	2	17	2	19	16	22	4	25	10
3	16	6	2	9	18	13	3	17	3	19	17	22	5	25	16
3	17	6	3	9	19	13	5	17	4	19	18	22	6	25	17
3	18	6	4	9	20	13	10	17	5	19	20	22	7	25	19
3	19	6	5	9	21	13	16	17	6	19	21	22	8	25	22
3	20	6	9	9	22	13	17	17	7	19	22	22	9	26	3
3	21	6	10	9	23	13	19	17	8	19	23	22	10	26	5
3	22	6	16	9	24	13	20	17	9	19	24	22	11	26	10
3	23	6	17	10	2	13	22	17	10	19	25	22	12	26	16
3	24	6	19	10	3	14	2	17	11	19	26	22	13	26	17
3	25	6	20	10	4	14	3	17	12	20	2	22	14	26	19
3	26	6	21	10	5	14	5	17	13	20	3	22	15	26	20
4	2	6	22	10	6	14	10	17	14	20	4	22	16	26	22
4	3	6	24	10	7	14	16	17	15	20	5	22	17		
4	5	7	2	10	8	14	17	17	16	20	6	22	18		
4	6	7	3	10	9	14	19	17	18	20	7	22	19		

can be filtered out by relevance filter. These variables can be filtered out in stage two automatically (without depending on threshold value) if redundancy exist. The relevant features from stage one are added to set S_1.

Stage 2 The three-variable interaction information between target variable and features in set S_1 is computed. For better illustration, Table 5.8 provides the list of pair of variables for which $I(Y, X_i, X_j) < 0$. The results obtained from the redundancy filter are shown in column 1-8.

Interaction information $I(Y, X_i, X_j)$ is a symmetric measure; it cannot derive the direction whether X_j inhibits the correlation between (Y, X_i) or X_i inhibits the correlation between (Y, X_j). Therefore, it become difficult to filter the redundant variable from the set of relevant features (X_i, X_j) when $I(Y, X_i, X_j) < 0$. This limitation of interaction information is re-

lieved by focusing on mutual information between target and input variables $I(Y, X_i)$.

The algorithm in stage two starts with maximum relevance rank variable X_{imax} from stage one. The variable EPPI(26) is ranked first as evident from Table 5.7. Add X_{26} to set S_2. For the first relevance ranked variable X_{26}, there are eight sets $\{Y, X_{26}, X_j\}$ where $j = \{3, 5, 10, 16, 17, 19, 20, 22\}$ for which interaction information is negative i.e. $I(Y, X_i, X_j) < 0$. The question that arises here is whether X_j inhibits the correlation between Y and X_{26} or X_{26} inhibits the correlation between Y and X_j. The redundant variable is filtered by comparing mutual information $I(Y, X_{26})$ with $I(Y, X_j)$ for each j. The results obtained in Table 5.7 show that mutual information $I(Y, X_{26}) > I(Y, X_j)$ for each j. Therefore, the variables X_j $\forall\ j = \{3, 5, 10, 16, 17, 19, 20, 22\}$ are redundant variables and must be filtered out from the list of relevant non-redundant variables. Similarly, the process holds for next ranked variable X_{25} from Table 5.7. There are eight set of variables having interaction information $I(Y, X_{25}, X_j) < 0$ where $j = \{2, 3, 5, 10, 16, 17, 19, 22\}$. Since mutual information $I(Y, X_{25}) > I(Y, X_j)$ $\forall\ j = \{2, 3, 5, 10, 16, 17, 19, 22\}$, therefore, X_{25} is added to set S_2 and X_j where $j = \{2, 3, 5, 10, 16, 17, 19, 22\}$ are filtered out by redundancy filter. Table 5.9 shows the list of relevant and non-redundant features selected from stage one & two. The number of candidate inputs (N) are reduced in

Table 5.9: Filtered features by redundancy filter in stage two of proposed algorithm for sub-period 2

Filtered feature by redundancy filter	**Feature No.**	**Feature Rank**
EPPI	26	1
CPI	25	2
DJI	23	3
CC	6	4
OECD-R	13	8
SPR	12	10
OPEC-R	14	11
RP	11	15
CR	15	24

stage two from 25 to 9; i.e. to less than 45% of the actual number of input variables. The set of variables in Table 5.9 have been cross-validated for three-variable interaction information such that $I(Y, X_i, X_j) > 0$ always.

The set S_2 is the set of features selected by stage two of the proposed algorithm. Since three-variable interaction information has provided features for which $I(Y, X_i, X_j) > 0$, therefore, higher order interaction information is not required.

Stage 3 The algorithm in stage three starts with maximum relevance rank variable X_{26} from Table 5.7. By default, X_{26} is considered as part of final set S_3. Consider the next relevance rank variable X_{25}. According to the pre-specified threshold value $Th2$, variables from set S_2 are filtered out based on mutual information. Since $I(X_{26}, X_{25}) > Th2$, therefore, X_{25} is filtered out by redundancy filter. For the next subsequent relevant ranked variable X_m, calculate maximum mutual information $Max(MI)$ between X_m and previously selected candidates by redundancy filter. If mutual information $Max(MI) > Th2$ for any set, then X_m is filtered out by redundancy filter. Otherwise, X_m is added to the final selected features set S_3. The algorithm will run iteratively for all 9 selected variables from stage two. The final selected features from the proposed I^2MI^2 algorithm are EPPI (26), DJI(23), CC (6) and CR(15). Thus, four out of twenty five variables were selected to represent main changes in oil prices after the crisis. The selected features are then used as input variables to neural networks forecasting engines.

Forecasting Results

The study proposes I^2MI^2 based features selected as input to GRNN forecasting engine to forecast crude oil prices. The performance of proposed feature selection algorithm with GRNN forecasting engine is evaluated based on RMSE, MAE and MAPE. The proposed ensemble model is used to forecast in-sample and out-of-sample. Firstly, in order to compare the model's capability with other models, nearly 4.4-year (August 2008-November 2012) monthly data is used for training and validation. In-sample evaluation is shown in Table 5.10. The model is used to produce one and twelve-month

Table 5.10: In-sample performance of proposed methodology

Model	**RMSE**	**MAE**	**MAPE**
I^2MI^2+GRNN	4.41	3.41	4.31

ahead out-of-sample forecasts from December 2012 till November 2013. To evaluate the performance of our model, it is compared with EIA's STEO

econometric model forecast reported from December 2012 onwards. Out-of sample evaluations are shown in Table 5.11. The results show superior

Table 5.11: Out-of-sample forecast comparison

Model	**RMSE**	**MAE**	**MAPE**
One-month(Proposed)	2.64	2.01	2.12
One-month(STEO)	2.86	3.51	2.90
Twelve-month(Proposed)	6.47	6.30	6.27
Twelve-month(STEO)	9.81	8.36	8.31

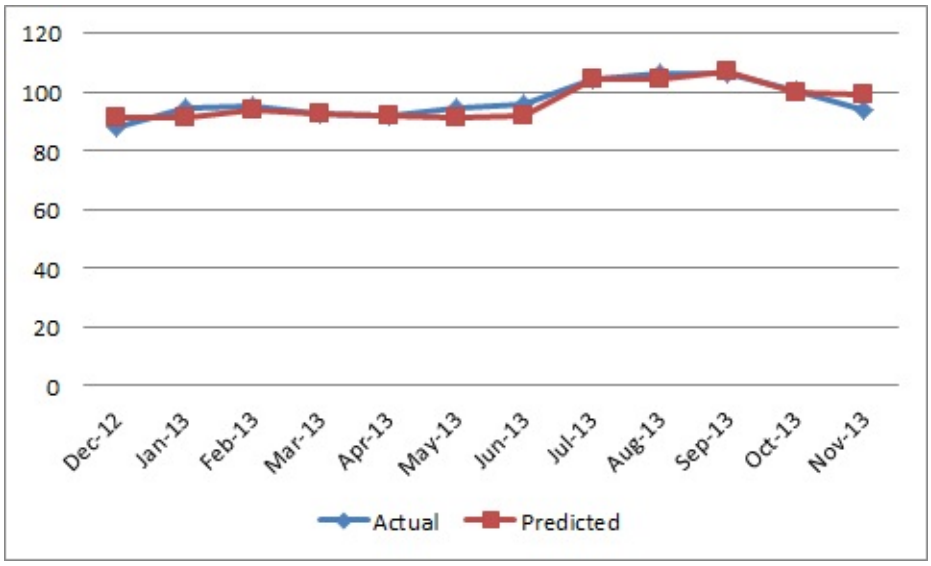


Figure 5.4: One-month out-of-sample forecast

performance of our proposed model in comparison to EIA's STEO model for both one-month and twelve-month ahead forecasts. The MAPE for the whole period (December 2012-November 2013) is 6.27 while RMSE and MAE are 6.47 and 6.30 respectively for twelve-month ahead time period. Similarly, the MAPE is 2.12 while RMSE and MAE are 2.64 and 2.01 for one-month ahead forecast horizon. Our model performed well in both in-sample and out-of-sample forecast horizons.

Fig 5.4 shows real versus predicted values for one-month ahead using proposed methodology. Fig 5.5 shows twelve-month ahead out-of-sample forecasts of our model. It is evident from both graphs that the model does very well based on input variables selected by proposed algorithm. The explanatory power of oil prices using four selected features is 93.8% after the crisis, indicating that the variable reduction is reasonable and that it

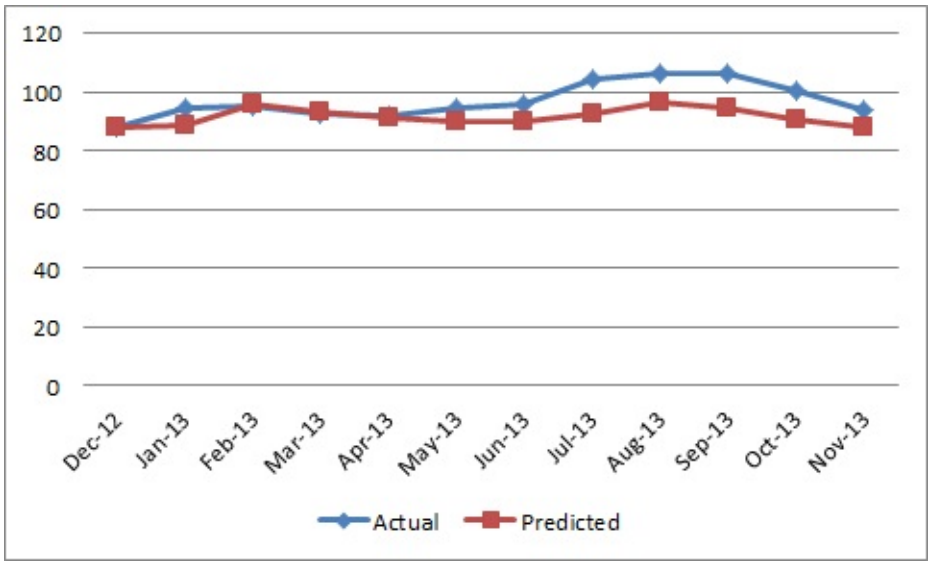


Figure 5.5: Twelve-month out-of-sample forecast

will have no essential influence on subsequent analysis. Fig 5.6 shows the percentage contribution of each selected factor in influencing oil prices post 2008 financial crisis.

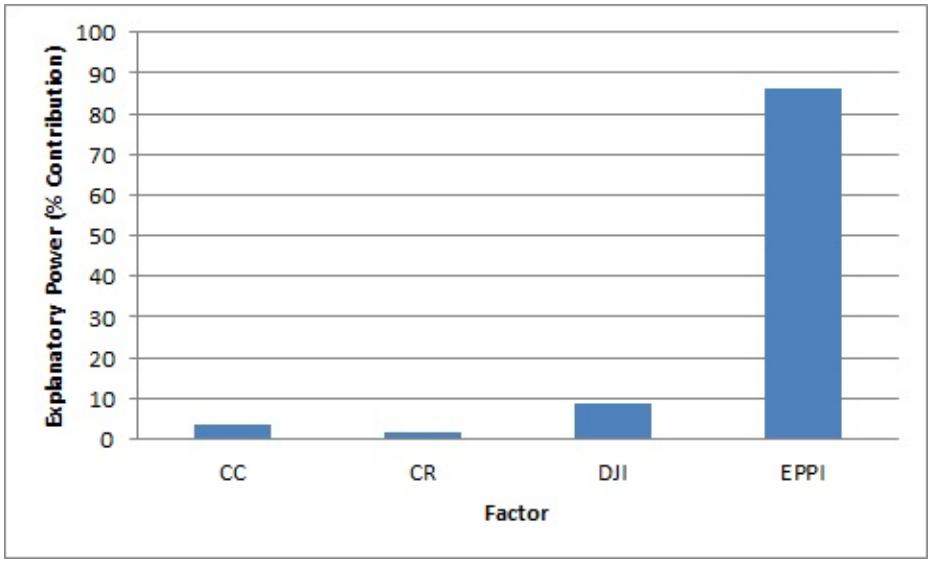


Figure 5.6: Explanatory power of 4 selected features after crisis

5.4 Factors Contribution to Oil Prices Before and After 2008 Financial Crisis

The explanatory power of each factor and the change in order before and after crisis is shown in Fig. 5.7. The importance of 11 variables (OPEC-

S, Non-OPEC-P, CC, Non-OECD-C, IC, OSC, OECD-R, OPEC-R, CR, RC, JU) increases, 10 variables (NCPP, Non-OECD-C, OPS, RP, SPR, I-OPEC, I-Non-OPEC, DER, GU, EU, GDP) decreases and for 4 variables (EPPI, CPI, DJI, OECD-C) remain unchanged. The analysis reveals that various driving factors show some new characteristics after the financial crisis. Same is discussed as follows:

- EPPI and CPI have taken up first two position before and after crisis. Speculation position has declined significantly after crisis due to high fluctuation in oil prices.
- Influence of Non-OECD consumption has increased after crisis but OECD consumption remains at same pace.
- The explanatory powers of China consumption and China reserves have increases and they both have emerged as an important variables driving oil prices.
- The explanatory power of strategic petroleum reserves and reserve-production ratio's have weaken after crisis.
- Global economic recession weaken US dollar together with GU and EU exchange market. On the other hand, JU exchange market power increased post crisis.
- The explanatory power of imports from OPEC declined whereas import from Non-OPEC increased. Due to disturbance in oil market as OPEC cuts target production, U.S is heading for sustainable solutions.

The number of variables were reduced from 25 to 11 in stage two for before crisis analysis by retaining only non-redundant variables. Similarly, the number of variables were reduced from 25 to 9 in stage 2 for after crisis analysis. The minimal set of variables that performed significantly well in finding direction of oil prices consists of only 5 and 4 variables as input to neural networks for before and after crisis analysis.

The explanatory power for oil prices using five selected features is 93.8% after the crisis, indicating that the variable reduction is reasonable and that it will have no essential influence on subsequent analysis. Overall, before the crisis, NCPP, EPPI, DER, SPR and RP were the major players

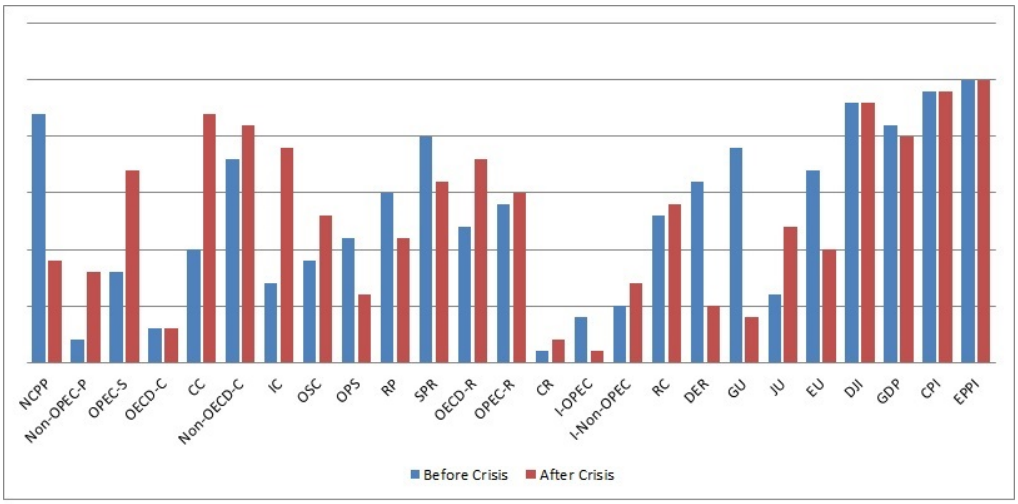


Figure 5.7: Variable ranking before and after crisis based on stage one of proposed algorithm

that influence oil prices volatility. Before the crisis, DER was the major factor boosting change in oil prices together with RP. SPR played a major role in influencing oil prices due to disturbance created by cuts in OPEC production or OPEC news. On the contrary, the original mechanism of crude oil market was destroyed by 2008 financial crisis and the relationship of EPPI and DER with oil prices strengthened after crisis. China consumption and its reserves emerged as important influencing variables in recent times. The supply-demand framework has weaken after crisis and the influence of emerging economies has increased.

5.5 Concluding Remarks

Oil prices are dependent on numerous indicators but there influence is subject to happening of geopolitical and economic events. There is need to identify appropriate features explaining the characteristics of oil markets during booming and downturn period. Feature selection can help in identifying the input variables before and after financial crisis. I^2MI^2 algorithm is applied with GRNN as forecasting engine to examine the explanatory power of selected features and their contribution in driving oil prices.

The explanatory power of selected features by I^2MI^2 algorithm before and after crisis is above 98%. Reserves and speculations were main players before the crisis and the overall mechanism was broken due to 2008 global financial crisis. The contribution of emerging economy (China) emerged

as important variable in explaining the directions of oil prices. EPPI and CPI remain the building blocks before and after crisis while influence of Non-OECD consumption rises after the crisis.

Chapter 6

Conclusions & Recommendations

6.1 Introduction

Over the past few decades, a rich set of data analysis techniques, including methods and algorithms, are been applied to select relevant features for oil price forecasting. There have been enormous development in data mining community in recent years but most of the studies related to oil price predictions have been concentrated in developing econometric time series models. This thesis has dealt with the challenging problem and tries to propose MI^2 and I^2MI^2 algorithm to provide the set of relevant and non-redundant features that can increase oil price forecasting performance. The effectiveness of the proposed algorithm with neural network as forecasting engine has been examined and evaluated with several state-of-art feature selection methods.

Section 6.2 discusses conclusions based on data analysis and evaluations. Section 6.3 provides certain recommendations for future researchers. The limitations of this study are discussed in section 6.4. Section 6.5 provides an overview of general contribution of this study in field of data mining and applied energy. Section 6.6 discusses the future scope of this study for researchers.

6.2 Conclusions

Based on the research study and analysis the following conclusions have been reached.

- **Impact of emerging economies**
  The proposed algorithm proved the importance of emerging economies (Non-OECD Consumption, China Consumption and China Reserves) in driving oil prices. The results from both groups identified China Reserves as a key variable for deriving the future price path. The proposed algorithm has performed well in figuring out the new changes in relationship between WTI and various factors. China consumption and its reserves have emerged as influential factors driving oil prices post 2008 financial crisis with drastic increase in their respective percentage contributions.

- **Recent change in impact of Non-OECD consumption compared to OECD Consumption as influential factor**
  According to British petroleum [5], Non-OECD consumption grew by 5.3% in track with 10-year average. It is evident from recent reports from EIA or BP that OECD Consumption tends to fall, while Non-OECD Consumption is projected to increase. The proposed algorithm selects Non-OECD Consumption as an key factor in driving oil prices. This confirms the superiority of the proposed algorithm in figuring out the recent change in data and identifying most relevant and non-redundant features for the study.

- **Non-OPEC Production & OPEC Supply as an emerging factor driving oil prices**
  Periodically, fluctuations in oil prices have been repercussion of OPEC news regarding cuts in production targets or changes in OPEC policies. OPEC announcements regarding change in policies or shift in production targets lead to change in oil prices. The impact of OPEC Supply still dominates the fluctuations in oil prices and the proposed methodology works well in identifying such relevant indicators driving oil prices. With recent change incorporated by importers to become exporters, U.S and China strive to optimize their domestic resources and become self-sustained to meet their own requirements. With increase in Non-OPEC production, the influence of older giants (OPEC) is diminishing as the most influential factors driving

oil prices. The proposed algorithm is effective in finding the most relevant features for the forecasting of oil prices.

- **NYMEX future prices is not a sole indicator**
  Many institutions-including central banks and international organizations are currently using NYMEX future prices as a key indicator for deriving the directions of spot prices. The results have shown there are number of external factors, which are driving crude oil prices. The explanatory power of NYMEX future prices is around 16% whereas 84% accounts for other factors that influenced oil prices over 17-year time period (result obtained through MI^3 algorithm + GRNNN based methodology). Further, the explanatory power of NYMEX future prices is around 62% while 38% share is explained by external factors (result obtained through I^2MI^2 + GRNNN based methodology). The results proved the importance of identifying key factors for deriving the future path of oil prices and not to focus on a single indicator.

- **Petrodollar Effect**
  U.S Dollar Index has become the main factor driving oil prices, its percentage contribution to WTI price fluctuations is around 17% of total share. U.S dollar is the basic reserve currency and more than 80% of all international currency transactions involves dollar. Oil is traded in global market and most of the trade has operated and continue to operate in dollars, even if U.S is not the trade partner. Since oil prices are defined in terms of dollars by most oil exporters, and as a result, oil importing countries also pay in dollars. This petrodollar cycle is an important factors which is highlighted by I^2MI^2 algorithm based methodology.

- **Effect of CPI and EPPI in influencing oil prices**
  Periodically, the relationship between CPI, as measure of inflation, and oil prices has been causal. Oil currently provides the majority of human energy requirements. Economy of any country can't run without oil and any fluctuations in oil prices effects economy directly. When oil prices are high, it leads to slower economic growth. High oil prices leads to high inflation initiating slower economic growth. The percentage contribution of CPI in influencing oil prices is around 37%. EPPI reflects the change of energy market and thus, the percentage

contribution of EPPI has been 25% for 17-year time period. CPI has the largest contribution followed by EPPI and DER as key drivers of oil prices.

- **Role of Speculation and Reserves before Crisis**
  Before the crisis, the effect of speculation in deriving oil prices increases with upward trend and shows strong explanatory power. After the crisis, its position weakens owing to high risk in crude oil markets and traders becoming susceptible of investing in oil markets. The proposed methodology is able to identify shifts in factors driving oil prices before and after 2008 financial crisis with high explanatory power. The role of reserves before the crisis seems to be enormous but weakens after the crisis. The importance of reserves before the crisis was repercussion of cuts in OPEC production targets or changes in OPEC policies. But after the crisis, increase in Non-OPEC Production indirectly weakens the effect of reserves on oil prices.

- **Original mechanism broke due to 2008 financial crisis**
  The overall mechanism of oil market broke after the crisis with EPPI, DJI, CC and CR being the influential factors driving oil prices. These four variables define the minimal set of input variables that can derive the direction of oil prices with high explanatory power post 2008 financial crisis. The effect of speculations and reserves together with EPPI is broken completely due to financial crisis. It shows that the influential mechanism of various factors on oil prices changed due to happening of geopolitical and economic events.

- **Superiority of proposed algorithms**
  Experiments showed that I^2MI^2 algorithm quickly identifies most relevant and non-redundant set of features. It has provided the minimal representative set of features which are more accurate in predicting oil prices as compared to other competing feature selection methods.

- **Number of features to be extracted**
  Without perturbing about the number of features to be extracted, on the natural domain, MI^3 and I^2MI^2 algorithm eliminated more than $\frac{1}{2}$ and $\frac{1}{4}$ of the features. The features thus selected through more refined version of MI^3 algorithm i.e. I^2MI^2 are 100% relevant and non-redundant features.

- **Application of proposed algorithm in varied disciplines**
  No single learning algorithm is superior to all others for all problems. But the proposed algorithm can provide the most relevant and non-redundant set of features for data mining problems. Practitioners can choose which algorithm to apply depending on their objective of the study. Armed with such insight, I^2MI^2 algorithm can enhance the performance of data mining problems, while at same time can achieve significant reduction in the number of features used in the study. I^2MI^2 algorithm can provide the minimal representative set of features for regression problems in business, biostatistics, applied energy and many more disciplines.

- I^2MI^2 **Algorithm is fully automatic algorithm**
  It doesn't require user to specify any number of features to be extracted or to specify any threshold. It operated on original feature set and doesn't incur the high computational cost associated with repeatedly invoking the learning algorithm.

- **Conditional Independence Assumption**
  Most of the feature selection methods assume that features are conditionally independent within the class. Due to existence of dependency within the features, these feature selection method could not perform well. This limitation is overcome in the proposed feature selection using the concept of interaction information. Conditional dependence is a measure of redundancy for complex real world problems. This research gap is taken care of in this study using proposed I^2MI^2 feature selection method.

6.3 Recommendations

The following recommendations are made to identify key drivers of oil prices.

- **Feature Selection for data mining problems**
  Forecasting oil price has never been an easy task, though it is important for so many economic policies. Using NYMEX future prices as stand-alone indicator for spot oil prices is not recommendable as there are high number of external factors influencing oil prices. Also, the input variables for any study cannot be selected based on judgemental criterion or trail and error method. It is a principal task to

identify key factors driving oil prices through feature selection algorithm before proceeding for model building and evaluation.

- **Influence of Geopolitical and Economic Events**
  The influence of input variables is assumed to be constantly driving oil prices in most studies. There is shift in influence of input variables driving oil prices subject to happening of geopolitical and economic events. These events are essential part of data analysis. Researchers are recommended to test for structural change in data due to happening of extreme events and further, proceed for model building and evaluation.

6.4 Limitations

- **Availability of data**
  Though Energy Information Administration (EIA) provides a comprehensive database for most of influential factors that can be incorporated in study but there are many more factors that are required to be accumulated for such important studies. There is no stability for the factors that can be considered for the study. Data is available for some factors in weekly or monthly term while some are available on quarterly or yearly basis. Also, researchers can predict oil prices more accurately if the forecasts of key factors is available over long term.

- **Selection of input variable for study**
  There is no exhaustive list of features that can be considered for study. Researchers have different view on finding relationship of oil price with supply-demand or with inventory, but these are not the only factors driving oil prices. The initial step of defining the dataset for the research ia a crucial step.

6.5 Contributions

This research leads to general contributions to the field of data mining and applied energy. The contributions are as follows:

- The study presented a new three stage I^2MI^2 algorithm for feature selection method, that performs very competitively as compared to

several state-of-the-art feature selection methods. The study present both theoretical and empirical contributions. (Chapter-4)

- The study presents a new two stage MI^3 algorithm for oil price prediction that simultaneously improves the predictive performances for oil price predictions using significant input variables. (Chapter-4)

- The new proposed algorithms provides 100% non-redundant and relevant features than previous feature selection methods for applications in varied disciplines. The explanatory power of key indicators influencing oil price market before and after financial crisis is presented. (Chapter-4).

- The study presents a new ensemble learning algorithm ($I^2MI^2 + GRNN$) for prediction of oil prices with extensive empirical evaluation with EIA's STEO econometric model (Chapter-4 & 5).

- A framework which can be used for predicting future value of oil prices depending upon movements in key factors driving the oil prices. (Chapter-5)

- The novel I^2MI^2 algorithm , which can be seen as a realization and an application of the proposed framework. Our experiments on real world problems show that the proposed algorithm performs very competitively as compared to other competitive ensemble models, and that it can provide optimized performance for real world complex problems. (Chapter-5)

6.6 Future Scope of the Study

The direction for future researchers are as follows:

- Detailed research can be carried out in subsequent studies by other scholars to quantify each factor for deriving directions of oil prices. Once these factors are quantified through separate research, the relevant and non-redundant features can be selected using proposed I^2MI^2 algorithm. Currently, this thesis has arrived at the fundamental stage of providing relevant and non-redundant features for any dataset.

- The study has been carried out to provide an insight into two major concerns: explanatory power of factors for oil price trend and their contribution to oil price prediction. The study can be expanded to provide information regarding transmission mechanism that follows between oil prices and factors.

- Future researchers can use number of other artificial intelligent forecasting engines with the proposed feature selection method to achieve high prediction performance for oil prices.

6.7 Concluding Remarks

The study has identified key factors influencing the direction of oil prices. China consumption and its reserves emerged as influential factors driving oil prices post 2008 financial crisis. The recent change in impact of Non-OECD consumption is highlighted in influencing oil prices as compared to OECD Consumption. OPEC Supply is dominating the fluctuations in oil prices due to sudden change in production targets or policies. With recent increase in Non-OPEC production, the influence of OPEC as the most influential factor driving oil prices is diminishing. NYMEX future price is not a stand-alone instrument for predicting spot oil prices but there are high number of external factors that are required to be identified. Since oil is traded in global market and most of the trade has operated and continue to operate in dollars, U.S Dollar Index remains an influential factor driving oil prices. Speculation and reserves played an important role in driving oil prices while CPI and EPPI have largest contribution as key drivers of oil prices after crisis.

The study showed the superiority of I^2MI^2 algorithm in comparison to other feature selection methods. Certain recommendations regarding the importance of each step in data mining process is highlighted in this chapter. Researchers are recommended to test for structural change in data due to happening of geopolitical and economic events. The contribution of the study in field of data mining and applied energy is presented together with future scope of the study.

Bibliography

[1] I. E. Agency, *World Energy Outlook, 2012*. OECD/IEA, 2012.

[2] P. Commission *et al.*, "Report of the committee on India Vision 2020," 2002.

[3] D. I. Stern, "The role of energy in economic growth," *Annals of the New York Academy of Sciences*, vol. 1219, no. 1, pp. 26–51, 2011.

[4] P. K. Narayan and R. Smyth, "Energy consumption and real GDP in G7 countries: new evidence from panel cointegration with structural breaks," *Energy Economics*, vol. 30, no. 5, pp. 2331–2341, 2008.

[5] B. Petroleum, "Statistical review of world energy 2013," *British Petroleum*, 2013.

[6] "Security of global oil flows: Risk assessment for india," November 2011.

[7] A. H. Demirbas and I. Demirbas, "Importance of rural bioenergy for developing countries," *Energy Conversion and Management*, vol. 48, no. 8, pp. 2386–2398, 2007.

[8] P. Sadorsky, "Oil price shocks and stock market activity," *Energy Economics*, vol. 21, no. 5, pp. 449–469, 1999.

[9] L. Kilian and C. Park, "The impact of oil price shocks on the US stock market," *International Economic Review*, vol. 50, no. 4, pp. 1267–1287, 2009.

[10] C. Benning and E. Pichersky, "Harnessing plant biomass for biofuels and biomaterials," *The Plant Journal*, vol. 54, no. 4, pp. 533–535, 2008.

[11] E. G. de Souza e Silva, L. F. Legey, and E. A. de Souza e Silva, "Forecasting oil price trends using wavelets and hidden markov models," *Energy Economics*, vol. 32, no. 6, pp. 1507–1519, 2010.

[12] E. Andreou, N. Pittis, and A. Spanos, "On modelling speculative prices: the empirical literature," *Journal of Economic Surveys*, vol. 15, no. 2, pp. 187–220, 2001.

[13] M. E. H. Arouri, A. Lahiani, A. Lévy, and D. K. Nguyen, "Forecasting the conditional volatility of oil spot and futures prices with structural breaks and long memory models," *Energy Economics*, vol. 34, no. 1, pp. 283–293, 2012.

[14] B. Xu and J. Ouenniche, "A data envelopment analysis-based framework for the relative performance evaluation of competing crude oil prices' volatility forecasting models," *Energy Economics*, vol. 34, no. 2, pp. 576–583, 2012.

[15] K. He, L. Yu, and K. K. Lai, "Crude oil price analysis and forecasting using wavelet decomposed ensemble model," *Energy*, vol. 46, no. 1, pp. 564 – 574, 2012.

[16] H. Mohammadi and L. Su, "International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models," *Energy Economics*, vol. 32, no. 5, pp. 1001–1008, 2010.

[17] T. Zeng and N. R. Swanson, "Predictive evaluation of econometric forecasting models in commodity futures markets," *Studies in Nonlinear Dynamics and Econometrics*, vol. 2, no. 4, pp. 159–177, 1998.

[18] 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*, pp. 213–216, IEEE, 2010.

[19] 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*, pp. 103–108, 2008.

[20] M. D. Chinn, M. LeBlanc, and O. Coibion, "The predictive content of energy futures: an update on petroleum, natural gas, heating oil and gasoline," tech. rep., National Bureau of Economic Research, 2005.

[21] J. Alvarez-Ramirez, E. Rodriguez, E. Martina, and C. Ibarra-Valdez, "Cyclical behavior of crude oil markets and economic recessions in the period 1986–2010," *Technological Forecasting and Social Change*, vol. 79, no. 1, pp. 47–58, 2012.

[22] K. Movagharnejad, B. Mehdizadeh, M. Banihashemi, and M. S. Kordkheili, "Forecasting the differences between various commercial oil prices in the persian gulf region by neural network," *Energy*, vol. 36, no. 7, pp. 3979 – 3984, 2011.

[23] 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*, pp. 1582–1585, IEEE, 2011.

[24] S. Dées, A. Gasteuil, R. Kaufmann, and M. Mann, "Assessing the factors behind oil price changes," 2008.

[25] Z. Jinliang, T. Mingming, and T. Mingxin, "Effects simulation of international gold prices on crude oil prices based on WBNNK model," in *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*, vol. 4, pp. 459–463, 2009.

[26] S. N. Abdullah and X. Zeng, "Machine learning approach for crude oil price prediction with artificial neural networks-quantitative (ann-q) model," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1–8, 2010.

[27] N. Krichene, "Crude oil prices: trends and forecast," *IMF Working Papers*, pp. 1–23, 2008.

[28] S. V. Chernenko, K. B. Schwarz, and J. H. Wright, *The information content of forward and futures prices: Market expectations and the price of risk*. Board of Governors of the Federal Reserve System, 2004.

[29] A. W. He, J. T. Kwok, and A. T. Wan, "An empirical model of daily highs and lows of west texas intermediate crude oil prices," *Energy Economics*, vol. 32, no. 6, pp. 1499–1506, 2010.

[30] R. S. Pindyck, "The long-run evolution of energy prices," 1999.

[31] E. Schwartz and J. E. Smith, "Short-term variations and long-term dynamics in commodity prices," *Management Science*, vol. 46, no. 7, pp. 893–911, 2000.

[32] M. Mazraati and S. Jazayeri, "Oil price movements and production agreements," *OPEC review*, vol. 28, no. 3, pp. 207–226, 2004.

[33] S. Moshiri and F. Foroutan, "Forecasting nonlinear crude oil futures prices," *The Energy Journal*, vol. 27, no. 4, pp. 81–96, 2006.

[34] A. Hou and S. Suardi, "A nonparametric GARCH model of crude oil price return volatility," *Energy Economics*, vol. 34, no. 2, pp. 618–626, 2012.

[35] S. H. Kang, S. M. Kang, and S. M. Yoon, "Forecasting volatility of crude oil markets," *Energy Economics*, vol. 31, no. 1, pp. 119–125, 2009.

[36] Y. Wei, Y. Wang, and D. Huang, "Forecasting crude oil market volatility: Further evidence using GARCH-class models," *Energy Economics*, vol. 32, no. 6, pp. 1477–1484, 2010.

[37] B.-N. Huang, C. Yang, and M. Hwang, "The dynamics of a nonlinear relationship between crude oil spot and futures prices: A multivariate threshold regression approach," *Energy Economics*, vol. 31, no. 1, pp. 91–98, 2009.

[38] L. Lixia, "Nonlinear test and forecasting of petroleum futures prices time series," *Energy Procedia*, vol. 5, pp. 754–758, 2011.

[39] K. M. Kisswani and S. A. Nusair, "Non-linearities in the dynamics of oil prices," *Energy Economics*, 2012.

[40] J. D. Hamilton, "Nonlinearities and the macroeconomic effects of oil prices," *Macroeconomic Dynamics*, vol. 15, no. S3, pp. 364–378, 2011.

[41] M. Ye, J. Zyren, and J. Shore, "Forecasting crude oil spot price using OECD petroleum inventory levels," *International Advances in Economic Research*, vol. 8, no. 4, pp. 324–333, 2002.

[42] M. Ye, J. Zyren, and J. Shore, "A monthly crude oil spot price forecasting model using relative inventories," *International Journal of Forecasting*, vol. 21, no. 3, pp. 491–501, 2005.

[43] M. Ye, J. Zyren, and J. Shore, "Forecasting short-run crude oil price using high-and low-inventory variables," *Energy Policy*, vol. 34, no. 17, pp. 2736–2743, 2006.

[44] M. Zamani, "An econometrics forecasting model of short term oil spot price," in *6th IAEE European Conference*, Citeseer, 2004.

[45] J. Chai, J. E. Guo, L. Meng, and S. Y. Wang, "Exploring the core factors and its dynamic effects on oil price: An application on path analysis and BVAR-TVP model," *Energy Policy*, vol. 39, no. 12, pp. 8022–8036, 2011.

[46] T. Xiong, Y. Bao, and Z. Hu, "Beyond one-step-ahead forecasting: Evaluation of alternative multi-step-ahead forecasting models for crude oil prices," *Energy Economics*, vol. 40, pp. 405–415, 2013.

[47] W. Qunli, H. Ge, and C. Xiaodong, "Crude oil price forecasting with an improved model based on wavelet transform and RBF neural network," in *Information Technology and Applications, 2009. IFITA '09. International Forum on*, vol. 1, pp. 231–234, 2009.

[48] A. Lin, "Prediction of international crude oil futures price based on GM (1, 1)," in *Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on*, pp. 692–696, IEEE, 2009.

[49] Y. Bao, Y. Yang, T. Xiong, and J. Zhang, "A comparative study of multi-step-ahead prediction for crude oil price with support vector regression," in *Computational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on*, pp. 598–602, 2011.

[50] 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*, vol. 3994 of *Lecture Notes in Computer Science*, pp. 444–451, Springer Berlin Heidelberg, 2006.

[51] 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, no. 5, pp. 2623–2635, 2008.

[52] 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, no. 4, pp. 275 – 286, 2012.

[53] R. Jammazi and C. Aloui, "Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling," *Energy Economics*, vol. 34, no. 3, pp. 828 – 841, 2012.

[54] K. L. Chang, "Volatility regimes, asymmetric basis effects and forecasting performance: An empirical investigation of the WTI crude oil futures market," *Energy Economics*, vol. 34, no. 1, pp. 294 – 306, 2012.

[55] 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, no. 3, pp. 889 – 904, 2008.

[56] A. Ghaffari and S. Zare, "A novel algorithm for prediction of crude oil price variation based on soft computing," *Energy Economics*, vol. 31, no. 4, pp. 531 – 536, 2009.

[57] M. T. Vo, "Regime-switching stochastic volatility: Evidence from the crude oil market," *Energy Economics*, vol. 31, no. 5, pp. 779–788, 2009.

[58] L. Liu and J. Wan, "A study of shanghai fuel oil futures price volatility based on high frequency data: Long-range dependence, modeling and forecasting," *Economic Modelling*, vol. 29, no. 6, pp. 2245–2253, 2012.

[59] X. Zhang, K. Lai, and S.-Y. Wang, "A new approach for crude oil price analysis based on empirical mode decomposition," *Energy Economics*, vol. 30, no. 3, pp. 905–918, 2008.

[60] W. M. Fong and K. H. See, "A markov switching model of the conditional volatility of crude oil futures prices," *Energy Economics*, vol. 24, no. 1, pp. 71–95, 2002.

[61] D. Huang, B. Yu, F. J. Fabozzi, and M. Fukushima, "CAViar-based forecast for oil price risk," *Energy Economics*, vol. 31, no. 4, pp. 511–518, 2009.

[62] M. R. Amin-Naseri and E. A. Gharacheh, "A hybrid artificial intelligence approach to monthly forecasting of crude oil price time series," in *The Proceedings of the 10th International Conference on Engineering Applications of Neural Networks, CEUR-WS284*, pp. 160–167, 2007.

[63] K. He, K. K. Lai, and J. Yen, "Crude oil price prediction using slantlet denoising based hybrid models," in *Computational Sciences and Optimization, 2009. CSO 2009. International Joint Conference on*, vol. 2, pp. 12–16, 2009.

[64] K. He, K. K. Lai, and J. Yen, "Morphological component analysis based hybrid approach for prediction of crude oil price," in *Computational Science and Optimization (CSO), 2010 Third International Joint Conference on*, vol. 1, pp. 423–427, 2010.

[65] K. He, C. Xie, S. Chen, and K. K. Lai, "Estimating VaR in crude oil market: A novel multi-scale non-linear ensemble approach incorporating wavelet analysis and neural network," *Neurocomputing*, vol. 72, pp. 3428 – 3438, 2009.

[66] M. Panella, L. Liparulo, F. Barcellona, and R. D'Ecclesia, "A study on crude oil prices modeled by neurofuzzy networks," in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pp. 1–7, 2013.

[67] G. Frey, M. Manera, A. Markandya, and E. Scarpa, "Econometric models for oil price forecasting: A critical survey," in *CESifo Forum*, vol. 10, pp. 29–44, Institute for Economic Research at the University of Munich, 2009.

[68] J. C. Reboredo, "How do crude oil prices co-move?: A copula approach," *Energy Economics*, vol. 33, no. 5, pp. 948–955, 2011.

[69] 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*, pp. 103–108, IEEE, 2008.

[70] Y. Pang, W. Xu, L. Yu, J. Ma, K. K. Lai, S. Wang, and S. Xu, "Forecasting the crude oil spot price by wavelet neural networks using oecd petroleum inventory levels," *New Mathematics and Natural Computation*, vol. 07, no. 02, pp. 281–297, 2011.

[71] S. Dées, P. Karadeloglou, R. K. Kaufmann, and M. Sanchez, "Modelling the world oil market: Assessment of a quarterly econometric model," *Energy Policy*, vol. 35, no. 1, pp. 178–191, 2007.

[72] 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*, vol. 2, pp. V2–465–V2–468, IEEE, 2010.

[73] 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, no. 2, pp. 421 – 430, 2012.

[74] A. Murat and E. Tokat, "Forecasting oil price movements with crack spread futures," *Energy Economics*, vol. 31, no. 1, pp. 85–90, 2009.

[75] W. Yang, A. Han, K. Cai, and S. Wang, "ACIX model with interval dummy variables and its application in forecasting interval-valued crude oil prices," *Procedia Computer Science*, vol. 9, pp. 1273–1282, 2012.

[76] X. Zhang, L. Yu, S. Wang, and K. K. Lai, "Estimating the impact of extreme events on crude oil price: An EMD-based event analysis method," *Energy Economics*, vol. 31, no. 5, pp. 768–778, 2009.

[77] Y. He, S. Wang, and K. K. Lai, "Global economic activity and crude oil prices: A cointegration analysis," *Energy Economics*, vol. 32, no. 4, pp. 868–876, 2010.

[78] H. Bu, "Price dynamics and speculators in crude oil futures market," *Systems Engineering Procedia*, vol. 2, pp. 114–121, 2011.

[79] Y. Wei, "Forecasting volatility of fuel oil futures in China: GARCH-type, SV or Realized Volatility models?," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 22, pp. 5546–5556, 2012.

[80] C. C. Wu, H. Chung, and Y. H. Chang, "The economic value of co-movement between oil price and exchange rate using copula-based GARCH models," *Energy Economics*, vol. 34, no. 1, pp. 270–282, 2012.

[81] 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, no. 2, pp. 107–119, 2002.

[82] M. Malliaris and S. G. Malliaris, "Forecasting energy product prices," in *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, vol. 5, pp. 3284–3289 vol. 5, 2005.

[83] A. Khashman and N. Nwulu, "Intelligent prediction of crude oil price using support vector machines," in *Applied Machine Intelligence and Informatics (SAMI), 2011 IEEE 9th International Symposium on*, pp. 165–169, 2011.

[84] J.-R. Zhu, "A new model for oil futures price forecasting based on cluster analysis," in *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on*, pp. 1–4, 2008.

[85] M. Kaboudan, "Compumetric forecasting of crude oil prices," in *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, vol. 1, pp. 283–287, IEEE, 2001.

[86] A. Alexandridis and E. Livanis, "Forecasting crude oil prices using wavelet neural networks," *Published in the proc. of 5th FSDET ($\Phi\Sigma\Delta ET$), Athens, Greece*, vol. 8, 2008.

[87] T. J. Considine and E. Heo, "Price and inventory dynamics in petroleum product markets," *Energy Economics*, vol. 22, no. 5, pp. 527–548, 2000.

[88] D. H. Kim, "What is an oil shock? Panel data evidence," *Empirical Economics*, vol. 43, no. 1, pp. 121–143, 2012.

[89] R. A. Ratti and J. L. Vespignani, "Crude oil prices and liquidity, the BRIC and G3 countries," *Energy Economics*, 2013.

[90] H. Li and S. Xiaowen Lin, "Do emerging markets matter in the world oil pricing system? Evidence of imported crude by China and India," *Energy Policy*, vol. 39, no. 8, pp. 4624–4630, 2011.

[91] Y.-J. Zhang and Z.-Y. Wang, "Investigating the price discovery and risk transfer functions in the crude oil and gasoline futures markets: Some empirical evidence," *Applied Energy*, vol. 104, pp. 220–228, 2013.

[92] S. A. Basher, A. A. Haug, and P. Sadorsky, "Oil prices, exchange rates and emerging stock markets," *Energy Economics*, vol. 34, no. 1, pp. 227–240, 2012.

[93] A. Mahdi, A. Hussain, and D. Al-Jumeily, "Adaptive neural network model using the immune system for financial time series forecasting," in *Computational Intelligence, Modelling and Simulation, 2009. CSSim '09. International Conference on*, pp. 104–109, 2009.

[94] J. Xiao, C. He, and S. Wang, "Crude oil price forecasting: A transfer learning based analog complexing model," in *Business Intelligence and Financial Engineering (BIFE), 2012 Fifth International Conference on*, pp. 29–33, IEEE, 2012.

[95] S. Yousefi, I. Weinreich, and D. Reinarz, "Wavelet-based prediction of oil prices," *Chaos, Solitons & Fractals*, vol. 25, no. 2, pp. 265 – 275, 2005.

[96] X. Guo, D. Li, and A. Zhang, "Improved support vector machine oil price forecast model based on genetic algorithm optimization parameters," *AASRI Procedia*, vol. 1, pp. 525–530, 2012.

[97] 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*, pp. 674–679, IEEE, 2013.

[98] C. E. Shannon, "Prediction and entropy of printed english," *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

[99] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, 1954.

[100] A. J. Bell, "The co-information lattice," *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan*, pp. 921–926, 2003.

[101] L. Kyung Joo, A. Y. Chi, Y. Sehwan, and J. Jongdae Jin, "Forecasting korean stock price index (KOSPI) using back propagation neural network model, bayesian chiao's model, and SARIMA model.," *Academy of Information & Management Sciences Journal*, vol. 11, no. 2, pp. 53 – 62, 2008.

[102] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 1. springer New York, 2006.

[103] T. Stengos, "Nonparametric forecasts of gold rates of return," in *Nonlinear Dynamics and Economics: Proceedings of the Tenth International Symposium on Economic Theory and Econometrics*, pp. 393–406, 1996.

[104] S. E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," 1989.

[105] L. Zhang, P. B. Luh, and K. Kasiviswanathan, "Energy clearing price prediction and confidence interval estimation with cascaded neural networks," *Power Systems, IEEE Transactions on*, vol. 18, no. 1, pp. 99–105, 2003.

[106] A. AlFuhaid, M. El-Sayed, and M. Mahmoud, "Cascaded artificial neural networks for short-term load forecasting," *Power Systems, IEEE Transactions on*, vol. 12, no. 4, pp. 1524–1529, 1997.

[107] S. Kouhi and F. Keynia, "A new cascade NN based method to short-term load forecast in deregulated electricity market," *Energy Conversion and Management*, vol. 71, no. 0, pp. 76 – 83, 2013.

[108] X. Peng, "The application of improved neural network in hydrocarbon reservoir prediction," in *2012 International Conference on Graphic and Image Processing*, pp. 87683I–87683I, International Society for Optics and Photonics, 2013.

[109] D. F. Specht, "A general regression neural network," *Neural Networks, IEEE Transactions on*, vol. 2, no. 6, pp. 568–576, 1991.

[110] P. H. Sherrod, "DTREG predictive modeling software. users manual," 2008.

[111] M. T. Leung, A.-S. Chen, and H. Daouk, "Forecasting exchange rates using general regression neural networks," *Computers & Operations Research*, vol. 27, no. 11, pp. 1093–1110, 2000.

[112] X.-H. Zhang, Q.-J. Wang, J.-J. Zhu, and H. Zhang, "Application of general regression neural network to the prediction of LOD change," *Chinese Astronomy and Astrophysics*, vol. 36, no. 1, pp. 86–96, 2012.

[113] E. Intelligence, "The international crude oil market handbook," 2009.

[114] K. Miller, M. Chevalier, and J. Leavens, "The role of WTI as a crude oil benchmark," *Purvin & Gertz for CME Group*, 2010.

[115] R. Silvério and A. Szklo, "The effect of the financial sector on the evolution of oil prices: Analysis of the contribution of the futures market to the price discovery process in the wti spot market," *Energy Economics*, 2012.

[116] L. Xiaotong, S. Shaohui, and L. Taohua, "Prediction of world crude oil price with the method of missing data," 2013.

[117] N. Amjady and A. Daraeepour, "Design of input vector for day-ahead price forecasting of electricity markets," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12281–12294, 2009.

[118] X. Zhang, X.-M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.-K. Hao, Z.-P. Liu, and L. Chen, "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information," *Bioinformatics*, vol. 28, no. 1, pp. 98–104, 2012.

[119] R. Menezes, A. Dionisio, and H. Hassani, "On the globalization of stock markets: An application of vector error correction model, mutual information and singular spectrum analysis to the G7 countries," *The Quarterly Review of Economics and Finance*, 2012.

[120] H. Liu and H. Motoda, *Computational methods of feature selection*. Chapman and Hall/CRC, 2007.

[121] M. A. Hall, *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[122] S. Hammoudeh and H. Li, "The impact of the Asian crisis on the behavior of US and international petroleum prices," *Energy Economics*, vol. 26, no. 1, pp. 135–160, 2004.

[123] R. A. Olowe, "Oil price volatility, global financial crisis and the month-of-the-year effect.," *International Journal of Business & Management*, vol. 5, no. 11, 2010.

[124] A. Charles and O. Darné, "Large shocks and the september 11th terrorist attacks on international stock markets," *Economic Modelling*, vol. 23, no. 4, pp. 683–698, 2006.

[125] W. L. Kohl, "OPEC behavior, 1998–2001," *The Quarterly Review of Economics and Finance*, vol. 42, no. 2, pp. 209–233, 2002.

[126] M. G. Guidi, A. Russell, and H. Tarbert, "The effect of OPEC policy decisions on oil and stock prices," *OPEC Review*, vol. 30, no. 1, pp. 1–18, 2006.

[127] S. X. Lin and M. Tamvakis, "OPEC announcements and their effects on crude oil prices," *Energy Policy*, vol. 38, no. 2, pp. 1010–1016, 2010.

[128] H. Schmidbauer and A. Rösch, "OPEC news announcements: Effects on oil price expectation and volatility," *Energy Economics*, vol. 34, no. 5, pp. 1656–1663, 2012.

[129] V. Brémond, E. Hache, and V. Mignon, "Does OPEC still exist as a cartel? An empirical investigation," *Energy Economics*, vol. 34, no. 1, pp. 125–131, 2012.

[130] R. Demirer and A. M. Kutan, "The behavior of crude oil spot and futures prices around OPEC and SPR announcements: An event study perspective," *Energy Economics*, vol. 32, no. 6, pp. 1467–1476, 2010.

[131] K. Hanabusa, "The effect of 107th OPEC ordinary meeting on oil prices and economic performances in Japan," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 3, pp. 1666–1672, 2012.

[132] R. Bhar and A. G. Malliaris, "Oil prices and the impact of the financial crisis of 2007–2009," *Energy Economics*, vol. 33, no. 6, pp. 1049–1054, 2011.

[133] Y. Fan and J.-H. Xu, "What has driven oil prices since 2000? A structural change perspective," *Energy Economics*, vol. 33, no. 6, pp. 1082–1094, 2011.

[134] G. Prat and R. Uctum, "Modelling oil price expectations: Evidence from survey data," *The Quarterly Review of Economics and Finance*, vol. 51, no. 3, pp. 236–247, 2011.

Profile of the Author

Neha Sehgal is a Ph.D. candidate at College of Management & Economic Studies, University of Petroleum & Energy Studies. She joined Jindal Global Business School, O. P. Jindal Global University in July 2012 as a core faculty in the research cluster of management and quantitative techniques. She has previously worked with SYSTAT as statistical analyst. She earned her Masters in Statistics from Banasthali University in 2008.

Her research focuses on the data mining for price predictions, energy economics and time series econometrics. In 2013, she was awarded the Best Paper Award from International Conference on Management & Infrastructure for her research. She has presented her research in various international conferences across India, UK and China. She is keen to work towards exploiting the hidden patterns in data and generating novel algorithms to be applied in varied disciplines.