brand university
of applied sciences

The Use of Advanced Analytics and Artificial Intelligence in Oil and Gas Trading: Opportunities and Challenges for International Energy Traders

Studiengang: MBA Digitales Business Management

Name: Ismael Sebastian Fontaine

E-Mail-Adresse: ismaefontaine12@gmail.com

Betreuer:in: Prof. Dr. Thomas Reimer

Datum: 26.01.2026

Table of Content

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average with Exogenous Variables
CNN	Convolutional Neural Network
CTRM	Commodity Trading and Risk Management
DL	Deep Learning
DM	Diebold–Mariano Test
EEMD	Ensemble Empirical Mode Decomposition
EGARCH	Exponential Generalized Autoregressive Conditional Heteroskedasticity
ES	Expected Shortfall
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GBM	Gradient Boosting Machine
HAR	Heterogeneous Autoregressive Model
HAR-RV	Heterogeneous Autoregressive Model of Realized Volatility
HFT	High-Frequency Trading
IEA	International Energy Agency
IMF	International Monetary Fund
LSTM	Long Short-Term Memory Network
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
OOS	Out-of-Sample
OPEC	Organization of the Petroleum Exporting Countries
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QLIKE	Quasi-Likelihood Loss Function
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RER	Relative Error Reduction
SLR	Systematic Literature Review
SVR	Support Vector Regression
TTF	Title Transfer Facility
VaR	Value-at-Risk
VAR	Vector Autoregression
WTI	West Texas Intermediate
XGBoost	Extreme Gradient Boosting

Tables

Explanation

I hereby declare under oath that I have written the present thesis independently, without using any unauthorized aids, and that I have used no sources other than those stated. All passages taken verbatim or in substance from the stated sources are clearly indicated as such.

If any aids (in particular IT- and AI-assisted tools) are used in the master's thesis, I declare that I have fully listed them in the thesis, including the respective product name, product version, and a description of the scope of the functions used.

I further affirm that I have prepared this thesis in accordance with the applicable examination regulations of Brand University of Applied Sciences. The thesis has not previously been submitted, either domestically or abroad, for review or assessment, and has not been published.

Place und Date: Vienna 26.01.2026

Signature: Ismael Sebastian Fontaine

Abstract

The research tests whether AI forecasting models provide better accuracy for oil and gas price and volatility time series data than traditional statistical methods. The research uses a standardized out-of-sample back testing framework to assess forecast accuracy and risk metrics which include Value-at-Risk (VaR) while testing how different market conditions affect the results. The research shows AI models produce significant performance gains for oil and gas markets during times of high market volatility, but their value for VaR estimation varies based on different contexts. AI systems improve forecasting accuracy, but organizations must conduct thorough validation processes and consider different operational states before using them to assess risk.

1 Introduction

1.1 Background and Motivation

The global oil and gas trading market functions as two distinct areas which intersect between physical markets that rely on supply and demand and storage and logistics and financial markets which operate through futures contracts and derivatives and market sentiment. The current situation provides researchers with a vast research environment which combines economic data with geopolitical information. The determination of crude prices requires the assessment of multiple factors which include international demand patterns and production capabilities and stock levels and OPEC announcements and overall economic conditions (EIA 2025). The unexpected nature of these factors together with instant changes in market conditions creates continuous difficulties for traders and analysts who attempt to make forecasts.

Recent years have shown that predictive models can be disrupted by unexpected events. The COVID19 pandemic caused an energy demand drop which reached its lowest point in history and resulted in negative WTI futures prices on April 2020. This episode demonstrated that traditional forecasting methods and decision-making processes encountered major limitations (IEA, 2020; EIA, 2021). Market disruptions from geopolitical events or coordinated OPEC releases have shown unpredictable market behaviour which exists at the same level of unpredictability that complex forecasting systems were designed to handle (Loutia, Mellios, & Andriosopoulos, 2016; IMF, 2022). Energy trading companies now use advanced analytics and AI and machine learning technologies to handle their new challenging market conditions because these tools help them predict market trends and make better decisions while developing effective risk management plans.

The top CTRM companies research findings demonstrate that they apply AI and advanced analytics to a greater extent than their market competitors do. The companies channel their financial resources into data science and data analysis which results in improved model performance through enhanced accuracy (Capgemini Invent, 2024; Reuters, 2024).

The adoption of Machine Learning (ML) and Deep Learning (DL) technologies by energy traders has increased for two decades because these technologies enable traders to identify patterns and find abnormal patterns within extensive time-series data (Financial Times, 2024; CTRM Center, 2025).

1.2 Problem Statement

The impact of artificial intelligence and advanced analytics on oil and gas trading remains uncertain despite their transformative effects on multiple industries. The knowledge rich market experiences inherent instability which results in increased model sensitivity to changes in market regimes and nonlinear patterns and extreme tail events. The econometric models such as ARIMA, GARCH, or HAR models have been a standard practice to forecast the price level and volatility (Bollerslev, 1986; Engle, 1982). The introduction of machine learning and deep learning techniques enables researchers to identify deeper nonlinear patterns together with market data adaptive responses.

Simultaneously, scholarly evidence is still inconclusive. As for the larger fore casting competitions M4 and M5, hybrid or statistically blended models, for example, have been shown to be as good, or even superior to, ML pure solutions, depending on sample data structure as well as evaluation scheme design (Makrida-kis, Spiliotis & Assimakopoulos, 2018, 2022). Similarly, experimental research using energy markets has yielded a mixed bag, with some announcing performance improvement using deep learning architectures while forecasting crude oil (Foroutan, Sa-madi & Mohammadi, 2024; Guo, Li & Zhang, 2023), while other work shows low transferability across markets.

As a result, the evidence gaps correspond to three fundamental categories, as stated below.

First, a comparative evidence gap exists there are insufficient transparent, reproducible studies that evaluate AI and ML models using consistent out-of-sample evaluation against robust econometric baselines.

Second, a translation evidence gap exists in risk outcomes marginal improvements in forecast accuracy do not always improve Value-at-Risk (VaR) exceptions or enhance back-testing coverage, both of which are necessary for real trading applications (Christoffersen, 1998; Campbell, 2005).

The existing governance evidence gap exists because regulators and supervisory authorities require organizations to manage model risks through their development of clear documentation standards for model validation and monitoring processes which remain unstandardized for artificial intelligence systems.

The trading industry has adopted AI technology into its operational processes yet there exists no organized method to create repeatable results from empirical research which demonstrates how AI technology impacts forecasting and risk management.

To address this research gap, there will be conducted a robust systematic literature review (SLR) and use PRISMA principles to apply structured quantitative synthesis to evaluate performance results from different studies while investigating how market conditions and asset categories and assessment methods influence study outcomes.
Objectives and Scope

The main goal of this research work is to establish an understandable and reproducible process which uses scientific evidence to determine the conditions that affect advanced analytics and artificial intelligence (AI) systems ability to enhance forecasting accuracy and risk assessment in global oil and gas trading operations.

The research study investigates existing academic publications which have been published between 2010 and 2025 instead of creating new forecasting methods or performing original research evaluations. The time period under study shows how machine learning methods became more popular while energy markets experienced major changes which resulted from the oil price collapse in 2014 and the COVID-19 pandemic and current geopolitical developments.

The thesis contains multiple research objectives which serve as the main method to reach its main research goal.

- **The first objective** of the study involves systematic identification and review of empirical studies that use artificial intelligence machine learning and deep learning methods to predict oil and gas market prices and volatility and risk measurements
- **The second objective** of the study compares the forecasting accuracy of AI models with established econometric benchmark models which include ARIMA

GARCH and HAR through the analysis of out-of-sample testing data that researchers have published

- **The third objective** of the study evaluates whether AI models achieved greater forecasting accuracy which resulted in improved risk management results through their effects on Value-at-Risk performance and back-testing assessment.
- **The fourth objective** of the study examines how market regime and asset type (between oil and natural gas) and forecast horizon and evaluation design affect the performance of AI-based forecasting models.
- **The fifth objective** of the study focuses on creating managerial and governance-related guidelines which must be followed to ensure responsible AI usage in energy trading through model validation and transparency and human oversight.

The thesis defines its research boundaries through its specific scope. The study does not collect or generate new primary data, nor does it implement or recalibrate forecasting models. The study uses only peer reviewed empirical research together with selected high quality industry publications that contain quantitative performance assessments. The analysis is restricted to oil and gas markets and excludes other asset classes such as equities, cryptocurrencies, agricultural commodities, or metals, unless oil or gas forecasting is a central component of the study.

The research uses Systematic Literature Review (SLR) together with quantitative comparison techniques to achieve its goal of advancing academic research and practical applications. The MBA programme in Digital Business Management requires students to develop data analytics skills together with digital transformation abilities and strategic decision-making capacity, which matches the research's focus on transparency and reproducibility and structured evidence synthesis method.

1.3 Contributions and Structure of the Thesis

The thesis provides academic research and managerial practice with multiple important contributions to their respective fields.

The study fulfils an academic requirement because it establishes a research gap which it fills through its comprehensive analysis of scattered research findings about artificial intelligence and advanced analytics applications in oil and gas trading. The research shows inconsistent results for machine learning and deep learning models because studies test different data sets and evaluation methods and market conditions. The thesis uses the Systematic Literature Review (SLR) method to create a complete review of AI based models which show benefits over standard econometric models in some situations. The study establishes how different energy market forecasting methods perform when using artificial intelligence technology under various market conditions.

The thesis provides energy trading and risk management decision makers with useful information which they can use for upcoming practical applications. The study investigates AI driven enhancements which extend beyond forecasting accuracy because it assesses their effect on risk management performance specifically for Value-at-Risk (VaR) and model reliability. The research results enable trading organizations to evaluate AI technology adoption while demonstrating that validation and transparency and governance practices are essential for successful implementation of advanced analytics in trading operations.

The structure of the thesis is organised as follows. Chapter 2 reviews the relevant literature on energy trading, market microstructure, classical forecasting models, and the growing application of artificial intelligence in commodity markets. This chapter establishes the theoretical and empirical foundation for the research. Chapter 3 outlines the research methodology, including the Systematic Literature Review design, the PRISMA-based screening process, the inclusion and exclusion criteria, and the methods used for qualitative synthesis. Chapter 4 presents the empirical results of the literature review, including a descriptive overview of the included studies, comparative performance results, and regime and risk related findings. The results show how the research objectives were achieved through existing research while delivering benefits for forecasting, risk management, and AI governance in energy trading.

The thesis uses this structure to develop academic research which meets strict academic standards and addresses real world problems which businesses encounter when they work between energy markets and data analytics and digital transformation.

2 Literature Review

2.1 Energy Trading & Market Microstructure (Oil & Gas)

The worldwide markets for oil and gas are beyond doubt the most complex ecosystems on the planet due to the interlocking of highly developed financial trading with the physical movements of commodities (International Energy Agency [IEA], 2023). Since microstructure determines the routes via which information, supply shocks, and risk are transmitted across the system and price discovery takes place at the financial and physical layers, it is important to understand it (Energy Information Administration [EIA], 2024).

The presence of these structural elements makes it difficult to predict price movements because both market fundamentals and liquidity conditions and trader expectations and short-term trading activities impact price movements. The data generating process in oil and gas markets exhibits non-linear behaviour because it depends on different market regimes and experiences sudden structural changes. Classical econometric models which depend on established relationships and previous data patterns face challenges when market conditions and information dissemination patterns change rapidly.

2.1.1 Overview of Global Oil and Gas Trading

The global oil and gas trade is conducted across two markets: the physical market consisting of production, transport, storage, and consumption, and the financial market comprising futures, derivatives, swaps, and ETFs. The physical market is characterized by the situation of the major oil exporting countries such as OPEC+ intending to boost production, hence refinery capacity, stock levels, and transport hurdles (International Monetary Fund [IMF], 2023). The financial market is the playground of hedge funds, commodity traders, and energy companies that employ specific financial contracts either to hedge against risky or to speculate on price fluctuations (Capgemini Invent, 2024).

In general, crude oil product prices are linked to worldwide price markers such as Brent and West Texas Intermediate (WTI), whereas natural gas prices reflect more the regional price hubs like Henry Hub, or the Title Transfer Facility (TTF) in Europe (IEA, 2022). There is a relationship between these price markers, and one can make a profit by trading price differentials in different regions. Nonetheless, the 2022 European gas crisis (Reuters, 2023) has witnessed the most unstable ventures across the regions.

The COVID-19 pandemic made it obvious how risky this system really was. WTI futures prices on April 2020 reached an unprecedented low and were even sold at a negative price (-$37.63 per barrel) as a result of the pandemic which shrank demand and turned the market into a reservoir of crude oil with no storage options (EIA, 2021).

2.1.2 Market Microstructure in Energy Trading

Market microstructure describes the various systems, operations, and regulations through which trades are executed, and prices are determined (O'Hara, 2015). In the case of energy markets, it implies the participation of exchanges, clearing houses, and the over the counter (OTC) market for bilateral contracts. The trading of oil and gas is mainly characterized by four different aspects.

1. Two types of trading venues the physical and the financial: Price discovery is simultaneous and occurs in both the spot and the futures markets. There is evidence that futures prices are often the first to indicate changes in the spot prices as they too incorporate the expectations about future fundamentals (Luo & Ye, 2022). However, in times of stress, the relationship might get inverted due to the presence of information asymmetry and liquidity constraints (Pedraz & González-Pérez, 2020).

2. Liquidity and high frequency trading: The development of both electronic trading and algorithmic execution has played a key role in the reduction of spreads along with the increase in intraday trading. Miao (2024) reports that the share of algorithmic strategies, which influence the short-term volatility scenarios, in the total volume of both Brent and WTI futures is already over 60%.

3. Information asymmetry and order flow toxicity: There are certain actors in the market, e.g., producers or inventory managers, who have much better knowledge of the supply/demand situation. Their trading generates adverse selection effects, which increase short-term volatility and the width of the bid ask spread (Będowska-Sójka & Trück, 2023).

4. Storage and logistical frictions: Oil and gas are not financial assets, but rather physical commodities, that must be stored and moved. The storage capacity affects the cost of carry, which results in backwardation (spot > futures) during times of storage constraints, while storage abundances lead to con tango (futures > spot) (Kang & Ratti, 2022). Such structural frictions influence the volatility and raise the cost of inefficient hedging.

2.1.3 Integration of Physical and Financial Markets

The physical and financial markets are tightly connected to each other. Shocks in one market will instantaneously reach the other market through arbitrage and sentiment movement. As an example, the geopolitical disruptions, like the one caused by the Russian invasion of Ukraine in 2022, firstly restricted the physical supply but soon led to massive flows of speculation into energy futures and ETFs (Financial Times, 2023).

This phenomenon has been characterized as the financialization of commodities, where macroeconomic factors and investor mood gradually dictate the price more and more (Tang & Xiong, 2012; updated by IMF, 2023). Financialization makes liquidity better and hedging easier; however, it can also disconnect prices from the fundamental physical forces, thus creating a risk of a systemic event.

Another factor that links both layers is storage arbitrage. Traders, they consider whether to keep inventories or sell futures contracts based on the rates of interest, the costs of storage, and the anticipated price spreads. The EIA (2024) points out that such intertemporal choices account for the majority of short term contango, and backwardation cycles seen since 2015.

2.1.4 Implications for Forecasting and Risk Management

There is a great impact of the microstructure of oil and gas markets on forecasting, volatility modelling, and risk management.

First, the forecasting models based on macro fundamentals only (e.g., demand, inventories) overlook microstructure dynamics like order book liquidity or execution costs. The implementation of these signals, which are accessible from high frequency data, can make the model more responsive to regime shifts (Foroutan et al., 2024).

Second, the noise from microstructure must be taken into consideration in volatility modelling. Liquidity shocks usually result in short term volatility, not fundamental supply changes (Będowska-Sójka & Trück, 2023). Models such as GARCH-MIDAS or HAR-RV increasingly employ microstructure variables (realized variance, bid–ask spread) to forecast volatility more accurately (Zhang & Wang, 2023).

Thirdly, Dismantling the stability of the distribution of the dataset implies the failure of the risk measures such as Value-at-Risk (VaR). Flash crashes and liquidity dry ups are the main disruptions of the financial microstructure creating fat tails and hence, violating the Gaussian assumptions. Therefore, regulators reinforce robust back-testing and model risk governance (Board of Governors of the Federal Reserve System, 2011).

In the end, model governance can be portrayed as a vital factor. When trading companies are using AI based forecasting systems, such as the European Securities and Markets Authority (ESMA, 2024) calls for transparency, validation, and explainability requirements for algorithmic trading, their supervisory bodies are in the thick of things. Keeping microstructure awareness in the decision-making processes ensures that the models are still interpretable, auditable, and reliable even under duress.

2.1.5 Section Summary and Transition to Forecasting Models

The section demonstrates how oil and gas trading markets display complex structural systems which connect their physical commodity movements to their financial trading operations. The price formation process depends primarily on market microstructure components which include liquidity conditions and information asymmetry and storage limitations and speculative trading activities, while these elements determine both short-term price fluctuations and price changes that occur over extended periods.

The characteristics of energy markets demonstrate that these markets experience continuous changes between different operational modes and exhibit unpredictable market patterns and sudden market disruptions, which create instability that contradicts typical market stability assumptions, which assume that markets operate in a straight-line manner. Historical data-based forecasting methods which use fixed parameters to calculate average values will not successfully predict all market patterns. This section begins to explain the existing market environment by presenting (2.2) the important econometric methods which researchers have used to predict commodity price movements and market volatility through ARIMA, GARCH, and HAR models. This review establishes the theoretical foundation which enables assessment of advanced analytics and artificial intelligence-based forecasting methods through evaluation of their potential advantages and disadvantages.

2.2 Commodity Forecasting and Volatility Modelling

For a long time now, the oil and gas markets have relied on econometric and statistical models to forecast future prices. Basically, these models try to pick up the patterns of price levels, volatility, and behaviour of the market. The main function of these models in the market is to provide quantitative future price movements which in turn help the market participants to take trading, hedging and risk management decisions. However, though the artificial intelligence and deep learning techniques have become a trend in recent years, traditional forecasting methods are still considered as the major methods to follow because of their transparency and well-established statistical properties (Hyndman & Athanasopoulos, 2021). The discussion hereafter outlines the prevalent econometric models in the commodity markets forecasting, their merits as well as de-merits in periods of high volatility.

2.2.1 Price Level Forecasting Models

Energy market traditional forecasting often uses linear time series models, such as ARIMA, ARIMAX, and VAR, which are vector autoregression. These models illustrate trends, seasonality, and lagged interactions by breaking price changes into their auto-regressive and moving average components (Box & Jenkins, 1976).

- ARIMA / ARIMAX: ARIMA models are preferred for commodity forecasting largely because of their simplicity and the fact that non-stationary data can be treated by transformation. ARIMAX goes further to include outside variables (e.g., stocks, lending rates, and output). It has been found that ARIMA based models can compete in short-term forecasting, particularly when the market is fairly calm (Gupta & Modise, 2013).
- Vector Autoregression (VAR): With VAR, simultaneous forecasting of several time series is done. This is very suitable to the situation where crude oil, natural gas, exchange rates, and macroeconomic indicators are all interlinked. The effectiveness of VAR to capture the interaction of the energy markets has been demonstrated by multiple researchers, yet the accuracy of the method falls in times of changes in the underlying structure (Kilian & Zhou, 2020).

Nevertheless, the econometric models maintain their interpretability and transparency, but they suffer during the crises such as the COVID-19 pandemic in 2020 or geopolitical shocks when the structural changes caused the non-linearity and the sudden shifts in the data generating process.

2.2.2 Volatility Modelling

The process of forecasting volatility is essential in energy trading, primarily because of the extensive use of derivatives and the necessity for precise risk assessment. GARCH (Generalized Autoregressive Conditional Heteroskedasticity) and its extensions are still the main techniques in volatility prediction.

- GARCH Family (GARCH, EGARCH, GJR-GARCH): Bollerslev (1986) was the one who introduced the GARCH model which is capable of taking need of the volatility clustering phenomenon where periods of high volatility are eventually followed by high volatility. The other two models, EGARCH and GJR-GARCH also deal with the clustering but these two include aspects of asymmetries i.e., negative shocks like (supply disruptions) might have different volatility effects than positive shocks.
- HAR-RV (Heterogeneous Autoregressive Model of Realized Volatility): The models of HAR-RV are built upon the use of realized volatility which is obtained from intraday high frequency data, and they are more appropriate for the market which experience microstructure noise. Their superior performance in forecasting energy volatility is confirmed by the empirical studies, especially when heterogeneous time scales (weekly, monthly) are incorporated (Corsi, 2009).
- QLIKE and Realized Measures: Volatility models are more and more evaluated by the methods QLIKE (Quasi-Likelihood Loss Function) and realized variance because these methods can provide more valid comparisons if the market consists of fat tailed distributions (Patton, 2011).

2.2.3 Limitations of Classical Models

The application of ARIMA and GARCH and HAR models in commodity price and volatility forecasting remains common yet these classical econometric methods face multiple structural obstacles when used in present-day oil and gas markets. Energy markets experience their primary challenge because they undergo structural breaks and regime shifts which happen due to geopolitical events and supply disruptions and regulatory interventions and macroeconomic shocks. The oil price collapse in 2020 and the geopolitical tensions of 2022 demonstrate how historical relationships can change suddenly which makes past data-based parameter estimates untrustworthy (Kilian and Zhou 2020). The current model fails to handle nonlinear market behaviour because it does not treat such market conditions effectively. Energy markets exhibit complex price behaviour because market participants react to price changes through an unpredictable pattern of buying and selling activities. The research demonstrates that negative supply shocks result in higher volatility than positive demand-driven market shifts which standard linear models fail to accurately show (Będowska-Sójka and Trück 2023).

The research fails to adequately model three phenomena which include threshold effects and saturation dynamics and the integration of physical market operations with financial market operations. The growing importance of market microstructure effects creates difficulties for existing forecasting methods which use traditional techniques. The rising use of algorithmic and high frequency trading systems causes short-term price movements to depend on liquidity conditions and order flow and bid ask imbalances instead of only fundamental market information (Miao, 2024). Microstructure-driven signals present challenges to classical models which depend on lower-frequency data and simplified variance structures for their estimation.

Most econometric models use fixed specifications together with time invariant parameters which restrict their capability to adjust to markets that experience fast changing conditions. The issue is partially resolved through extensions that include state space and time varying parameter models, but these extensions lose effectiveness when interactions reach a high level of complexity. The explanatory and predictive capabilities of classical models experience a decline during times when uncertainty levels reach their peak.

The existing limitations of traditional econometric methods prove their applicability inside specific conditions. The existing limitations of traditional econometric methods lead researchers to pursue more flexible modelling methods, which include machine learning and deep learning, because these methods handle non-linear relationships and different operating conditions and extensive data sets.

2.2.4 Performance Benchmarks and Evaluation Methods

Commodity markets require precise performance measures for testing forecasting models because energy price data shows three specific patterns which include volatility clustering and fat tailed return distributions and frequent structural breaks. The evaluation metrics together with testing frameworks determine the actual significance of reported forecasting enhancements for real world trading and risk management purposes.

The energy forecasting literature continues to use traditional point forecast accuracy measures which include Root Mean Squared Error, Mean Absolute Error, and Mean Absolute Percentage Error. The RMSE metric punishes large forecast errors with greater severity because it calculates extreme price changes between two points in time. MAPE enables users to assess relative mistakes yet becomes unstable when energy markets experience stress and the price level reach near zero (Hyndman & Athanasopoulos, 2021).

Researchers use realized measure-based loss functions to predict market fluctuations. The Quasi-Likelihood loss function serves as a standard performance measure because it maintains its measurement error resistance to volatility proxy errors while allowing researchers to assess different volatility forecasting methods through standardized model evaluations (Patton, 2011). Researchers studying high frequency data typically use realized volatility or realized variance as their primary reference point because these metrics provide better intraday market movement representation than daily squared returns (Corsi, 2009).

The main method used to evaluate models depends on whether researchers conduct testing with the same data used for model development or with new data. In-sample performance tends to overstate predictive accuracy due to overfitting and therefore provides limited insight into real-world forecasting performance. The energy forecasting literature now focuses on OOS evaluation frameworks which researchers typically implement through rolling or expanding window testing methods. Rolling windows enable models to adjust to current market conditions while expanding windows use extended historical data to maintain long-term market trends.

Statistical tests are often applied to assess whether observed differences in forecast accuracy are statistically significant. The Diebold–Mariano (DM) test is a common tool used to evaluate different forecasts by measuring performance differences that exist beyond random chance (Diebold & Mariano, 1995). For energy markets this situation holds particular importance because even minor enhancements in forecasting proficiency result in significant financial advantages.

The assessment of forecasting models requires risk-based evaluation frameworks which extend beyond point forecast assessments. The Value-at-Risk (VaR) back-testing procedures use Kupiec's proportion of failures test together with Christoffersen's conditional coverage test to assess whether predicted risk levels match actual loss occurrences. The tests have special significance in regulated trading settings because they enable traders to determine accurate risk estimations while validating their trading models, which they need to comply with regulations and make capital allocation decisions (Christoffersen, 1998).

The combination of point forecast error measurements together with volatility-specific loss functions and statistical comparison tests and risk based back-testing frameworks establishes a complete system which researchers can use to assess forecasting models in energy markets. The multi-dimensional evaluation approach enables researchers to differentiate between statistically attractive outcomes and forecasting models, which provide actual benefits for risk management and decision-making in uncertain situations.

2.2.5 Transition to AI and Advanced Analytics

The structural limitations of classical econometric models which earlier sections described have created an increasing demand for advanced analytics and artificial intelligence techniques to forecast energy markets. The oil and gas markets exhibit nonlinear dynamics while displaying regime-dependent behaviour and experiencing complex interactions between their physical fundamentals and financial market forces. The characteristics of these markets create challenges which cause traditional models to break their fundamental assumptions and produce unreliable forecasts during market stress.

Machine learning methods such as Random Forests Support Vector Regression and Gradient Boosting Machines provide users with greater flexibility because the methods enable them to learn nonlinear relationships and interaction effects directly from the data without requiring specific parametric assumptions about the data-generating process (Hastie Tibshirani & Friedman 2009). The methods effectively handle the current modern energy trading environments which present challenges through their extensive feature sets and noisy input data and their complicated dependency structures.

Deep learning methods that use Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) enable systems to handle time-based data and structured information through their capability to depict time-based relationships and hierarchical data structures. LSTM architectures have become the standard choice for predicting energy prices and market volatility because they enable researchers to capture extended memory patterns while needing to track market changes. Through their design deep learning models enable researchers to combine different data types from macroeconomic data and market microstructure data and news reports that show sentiment analysis.

AI based models have theoretical benefits, but research findings show mixed results about their effectiveness compared to traditional econometric models. Some research shows that forecasting accuracy and volatility prediction have improved while other studies demonstrate that performance enhancements depend on specific situations which include market conditions and forecast periods and evaluation methods. The M4 and M5 forecasting competitions demonstrate that basic statistical methods and hybrid models can achieve performance that matches or exceeds complex machine learning methods in specific situations (Makridakis et al., 2018; 2022).

The use of AI based models creates multiple challenges which impact their ability to be interpreted and validated and governed through their evaluation processes. The complex nature of their systems creates difficulties for understanding economic principles, which makes it harder to follow regulatory requirements in high-risk situations like energy trading. The decision to use these forecasting methods requires more than just their forecasting performance, because their forecasting accuracy needs to be evaluated together with their capacity to withstand disruptions and their ability to deliver clear information and handle potential risks.

The existing empirical research needs a comprehensive and structured synthesis because it contains contradictory results and real-world implementation difficulties. A Systematic Literature Review with its quantitative comparison provides a better method to compare AI based models with traditional econometric methods, because it allows for standardized evaluation across different assets and market conditions and assessment methods. The research design introduced in Chapter 3 relies on this methodological framework, which allows researchers to evaluate the circumstances in which artificial intelligence delivers advantages to oil and gas forecasting and risk management.

3 Methodology (Systematic Literature Review & Quantitative Synthesis)

3.1 Energy Trading & Market Microstructure (Oil & Gas)

The research uses a Systematic Literature Review (SLR) method to conduct structured quantitative research which it uses for its comparative analysis. The objective is to summarize and compare results across studies in a transparent way without computing pooled meta-analytic estimates. The intent of this study is not to create new forecasting models but rather to systematically uncover, assess, and combine existing empirical research on artificial intelligence and advanced analytics in oil and gas trading. This method is justified as the literature in this area is extremely disjointed, differs significantly in methodological rigor, and often yield conflicting performance results across different types of models, assets, market conditions, and evaluation strategies.

The systematic literature review (SLR) adopts the PRISMA 2020 guidelines as its main principles, which require clearly stating one's process, being able to repeat it, and having a well-organized record of the whole process to come up with a right conclusion in research that has been synthesized (Page et al., 2021). Consequently, the investigation is conducted according to a specific protocol that defines the searches in the databases, the use of the exact search strings, the criteria for inclusion and exclusion, and the general strategy for screening. Such protocol makes the selection of literature systematic instead of arbitrary, thus minimizing the potential for bias and improving the internal validity of the review.

In case the studies identified show comparable quantitative results usually forecast accuracy metrics like RMSE, MAE, MAPE, QLIKE, or Value-at-Risk back testing statistics the dissertation employs structured quantitative synthesis. Relative Error Reduction (RER) or other performance metrics are utilized to create the relative performance indicators that enable the comparison of AI based models with traditional econometric baselines. The Relative Error Reduction (RER) functions as a performance metric which scientists use to evaluate artificial intelligence systems against established econometric benchmarks that show baseline performance. The research maintains separate study results without combining or merging multiple studies. The results present two types of output which include a structured quantitative analysis and a narrative explanation of results.

The Relative Error Reduction (RER) function serves as a performance metric to assess AI models against econometric standards which have documented enough baseline performance data. The study presents its findings through structured comparisons which also include narrative descriptions for results that need additional explanation. The studies are classified according to the significant dimensions such as model type, asset class (oil versus natural gas), forecast horizon, sample period, market regime (e.g., pre- or post-2020), or evaluation design (static versus rolling out-of-sample). This process guarantees that the synthesis is still systematic, even without comparative quantitative data.

To sum up, this research design not only meets the objectives laid down in Chapter 1 but also (a) maps the existing evidence of the highest transparency and reproducibility, (b) performance differences between AI and classical forecasting models are synthesized, and (c) harvesting insights that are important for forecasting, risk management, and model governance in energy trading firms. Further, it is consistent with the MBA programme’s focus on rigorous methodologies, digital transformation, and data informed decision making

3.2 Protocol Registration and Transparency

In order to guarantee a transparent and replicable research process, this dissertation adopts a review protocol that was drafted prior to the commencement of the literature search. The protocol is not registered on the PROSPERO or Open Science Framework platforms, which is the case with medical sciences but not so much with economics or data analytics. The protocol was outlined before the literature search and is available from the author on request.

The main elements of the Systematic Literature Review are defined in the protocol. These are the databases utilized for the search, the specific search terms, the duration of the study, and the rules for study selection. The predefinition of the rules not only contributes to a well-structured review but also minimizes the chance of arbitrary choices or hidden reporting. Furthermore, it secures the uniformity of judging all studies by the same standards.

The protocol elucidates the screening process by PRISMA 2020. Initially, all studies that could possibly be relevant are found out in academic databases and acknowledged industry sources. Then, repeated entries are eliminated, and titles with abstracts are scanned for the eligibility criteria. At the last step, the full text of the remaining studies is reviewed so as to be certain they give the required number of details and fulfil all the inclusion criteria.

Moreover, the protocol highlights the quality appraisal procedure. It consists of considering if the studies give their data in a clear manner, apply satisfactory econometric baselines, carry out out-of-sample testing, report on performance metrics in a transparent way, and perform robustness checks. By considering variances in methodological rigor, these prerequisites set a cautious tone in discussing findings.

The protocol has also specified the data extraction plan that contains all the variables collected from each study, such as asset type, forecasting horizon, model classes, performance measures, sample period, and evaluation design. The standardisation applied here allows for consistency and the generation of comparisons with real meaning across the studies. This way of doing it is the writing down of all the methodological by documenting all steps beforehand and including the codebook and screening materials in the appendix. The thesis thus adopts the practice of transparency recommended in evidence synthesis research. Such a writing down makes the review to be reproducible, complies fully with the PRISMA 2020 standards, and leaves a very clear methodological trail for the future be it for updates or extensions of the study.

3.3 Data Sources and Search Strategy

In order to cover all the important empirical literature, this thesis employs a multi database search strategy that involves academic platforms commonly cited in the fields of economics, finance, energy studies, and machine learning research. The issue here is that studies on forecasting in the oil and gas markets are often published in different disciplines and journals, so using just one database would mean overlooking some important works. Therefore, the main literature search is done through Open Alex, Scopus, Web of Science, ScienceDirect, and IEEE Xplore. Scopus and Web of Science and Open Alex have been chosen for they are three of the largest and most commonly used citation databases, providing access to a wide range of peer reviewed articles from top journals like Energy Economics, Applied Energy, and Energy Policy. ScienceDirect, as the full text platform of Elsevier, also allows access to a considerable amount of empirical studies that employ econometric and statistical models in the context of energy markets. IEEE Xplore further adds to these sources by taking in the research output from the fields of engineering and computer science, specifically studies that apply machine learning and deep learning techniques, which are becoming increasingly important for forecasting in the commodities market.

Apart from these main academic databases, additional resources like Google Scholar and industry publications are checked to guarantee that nothing pertinent is missed. Google Scholar is only relied upon as an extra screening tool since it does not possess the filtering and quality control features required for systematic inclusion. Industry and regulatory sources including the International Energy Agency (IEA), U.S. Energy Information Administration (EIA), and CTRM Center offer insights and background information but are excluded from the quantitative synthesis as they usually do not provide the quantitative benchmarking necessary for relative performance indicators.

The search technique is executed in a systematic and reproducible way, thus abiding by the PRISMA 2020 guidelines. Every search is performed with the use of specific keywords, Boolean logic, and database filters that were set beforehand. The period of the review is from the year 2010 till 2025, which indicates the usage of machine learning techniques in forecasting and capturing market disruptions like the 2014 price drop, the COVID-19 pandemic in 2020, and the 2022 geopolitical shock, all of which are very important for understanding model performance in various conditions. One of the search strings applied in Open Alex consists of the following: combinations of terminology referring to the asset (“oil”, “crude”, “natural gas”), the forecasting field (“forecast*”, “predict*”, “volatility”, “VaR”), and the econometric or AI based model classes that are widely used (“ARIMA”, “GARCH”, “HAR”, “machine learning”, “deep learning”, “LSTM”, “XGBoost”, “SVR”, “neural networks”). Synonymous search keywords are modified to each database.

Once the process of search is complete, filters come into picture as it could limit output only to peer-reviewed journal articles in English that have been published during the time period from 2010 through 2025. All studies that may be relevant are exported together with their metadata and obtained as full text PDFs. The resulting records are then merged into one dataset prior to duplicate elimination and screening, which is explained in the next chapters. This methodical search strategy not only secures a wide range of literature but also guarantees the literature to be transparent and reproducible, and it incorporates both conventional econometric techniques and the latest AI powered forecasting methods in the oil and gas market.

3.4 Inclusion and Exclusion Criteria

To make the review more focused around the studies that have a similar methodology and are closely related to the research questions, this thesis uses the already determined inclusion and exclusion criteria. These criteria were set prior to the commencement of the screening process, and they were employed uniformly throughout the whole review. The aim is to make sure that just the empirical studies with significant and comparable forecasting or risk modelling results for oil and gas markets are included in the final evidence base.

The inclusion criteria are primarily concerned with the asset type, model type, evaluation design, and presence of quantitative performance metrics. Only those studies are included that look at the crude oil, refined oil products, or natural gas markets and use either traditional econometric models (like ARIMA, GARCH, or HAR) or machine-learning and deep-learning methods (like Random Forests, Gradient Boosting, LSTMs, or hybrid neural networks). Moreover, a study is considered eligible only if it shows a performance comparison between at least one econometric baseline model and one AI or ML approach. This condition makes it possible to calculate the relative performance measures in a consistent way. Moreover, only those studies that provide clear out-of-sample (OOS) evaluation results using metrics such as RMSE, MAE, MAPE, QLIKE, or Value-at-Risk (VaR) statistics are included, as in-sample results alone do not have enough predictive validity for real world forecasting.

In order to retain the utmost scientific quality of the evidence base, the review considers exclusively those peer reviewed articles and reviews published in the English language from 2010 to 2025. This time frame not only covers the period when machine learning methods gained more and more importance in forecasting but also gives space for the representation of market dynamics in terms of both pre and post 2020.

The exclusion criteria apply the same reasoning and eliminate studies that do not comply with the methodological or topical standards of this review. Studies that deal with assets like metals, agricultural commodities, equities, or cryptocurrencies are not considered unless oil or gas forecasting is a major part of the study. Besides, theoretical works, conceptual research without empirical validation, and studies that present only in-sample results or vague quantitative performance measures are not to be included. Use of proprietary datasets without detailing the modelling process or the benchmark methods leads to exclusion too, for their findings cannot be compared reliably. Working papers, theses, conference abstracts, or non-peer reviewed material are not taken into account unless they are published by recognised peer reviewed outlets.

Exclusion finally extends to studies that utilize machine learning models but do not give details about the baseline model for comparison or only report qualitative results, since such studies preclude quantitative synthesis. The same is true for those papers whose full text is not available or those in which critical information such as sample period, model specification, or evaluation design is missing.

The consistent application of these inclusion and exclusion criteria results in the review obtaining an evidence base that is coherent, comparable, and methodologically strong. This provides a solid foundation for the screening process described in Chapter 3.5 and for the quantitative comparison carried out in Chapter 4.

3.5 Screening Process (PRISMA)

In the present review, no studies were excluded at the full text stage (n = 0). This process includes three major phases, namely, identification, screening, and eligibility assessment. Each phase reduces the number of potential relevant studies in a clear and repeatable way.

During the identification phase, all the publications found in the chosen electronic databases are uploaded into one combined database. This includes all the results from Open Alex, Scopus, Web of Science, ScienceDirect, and IEEE Xplore, plus additional documents that were found via Google Scholar or checking references. At this phase, no decisions are taken about the relevance of the studies; the aim is to gather all the studies that could possibly be included in the review.

The next step is the screening stage in which the duplicates are taken off, and the rest of the titles and abstracts are compared with the previously set up inclusion and exclusion criteria. This stage removes studies of no interest at all, such as those dealing with non-energy commodities, papers without empirical forecasting results, or studies that do not compare econometric models with machine learning or deep learning techniques. The studies that have been selected in this initial assessment move on to full text review.

The last stage is the eligibility assessment where the full text of each study is thoroughly scrutinized to see if it has enough methodological details and empirical results for inclusion.

During this whole process, the causes for exclusion are recorded to keep the operation clear and to have a trace. The PRISMA flow diagram in Chapter 4 depicts a summary of all studies that were identified, screened, excluded, and included. This diagram shows the movement of the results from the initial search to the final selection of the studies that were subjected to synthesis and quantitative comparison. Documenting in this way is important for showing the methodological strength of the review and for the selection process of the studies to be replicated or checked by other researchers.

3.6 Quality Appraisal

In order to maintain the highest standards of the evidence base, every research article accepted through the screening process is subjected to an organized quality assessment. The main aim of this assessment is not only to eliminate studies, but also to determine the degree of reliability, transparency, and comparability of the reported results. This method is in line with the best practices in systematic literature reviews and it opens up the findings for a cautious interpretation, since the possible methodological shortcomings are taken into consideration.

The quality assessment is directed towards a number of important points that are most pertinent to forecasting and risk modelling studies in the energy market sector. The first point is that the data transparency is rated by the research, and the evaluation is based on the extent to which the research openly indicates its data sources, sample period, data preparation steps, and any transformations done. In case of studies that do not give enough information about their datasets, they will be treated as less reliable, since their results cannot be independently verified or replicated.

Secondly, the review looks into the appropriateness of the benchmark models employed in the different studies. Since the thesis is about the comparison of AI powered models and traditional econometrics, it is necessary for the studies to contain classic benchmark models like the random walk, ARIMA, GARCH, or HAR for the purpose of comparison. These benchmarks allow for a level playing field in terms of comparison. Studies that apply weak, poorly formulated, or ambiguous benchmarks are marked as they might inflate the performance gains attributed to the AI methods.

Third, the study designs are particularly emphasized in the evaluation of the studies. The evaluation considers whether true out of sample testing is implemented, whether rolling or expanding windows are employed, and whether forecast horizons are explicitly stated. In sample results are not enough for quality assessment as they do not represent real world forecasting conditions and are more prone to overfitting. Therefore, only studies that provide out of sample results and describe their evaluation framework clearly are favoured.

In the fourth place, the review checks if the performance indices are expressed clearly and in detail that is adequate. Reports are accepted only if they show standard and widely acknowledged indicators like RMSE, MAE, MAPE, QLIKE, or Value at Risk back testing results. Reports that show results only in an ambiguous way, disclose the meter stick used in a very selective manner, or do not give numerical values of the performance are less prioritized for synthesis and structured quantitative synthesis.

The last factor of quality appraisal considers whether robustness checks or cautious interpretation are used by the studies. This could be changing model specifications, varying forecast horizons, or examining sub samples of different market regimes. Studies that go on to prove that their results are stable under different conditions are given higher ratings, since they produce weaker and less doubtful evidence.

All these factors lead to the qualitative quality evaluation of each literature as a criterion. The evaluation is documented and subsequently utilized for supporting cautious interpretation and result interpretation. This method doesn't merely exclude studies based on quality but rather it helps to portray the differences in methodological rigor in the synthesis while making sure that the claims made in Chapters 4 and 5 are properly supported by the variations in quality of the studies.

3.7 Data Extraction and Codebook

Once the last group of studies has been recognized through the examination and evaluation of their quality, relevant information is systematically extracted, and a predefined data extraction scheme is used. The intention of this step is to create uniformity, interoperability, and visibility among all the studies that are included and to make the data available for the synthesis for the synthesis and the quantitatively structured comparison

The data-extraction technique is based on a defined codebook that outlines the variables to be collected from each study beforehand. This method minimizes the possibility of selective reporting and guarantees that all studies are judged by the same standards. For each study that is included, bibliographic data such as author names, year of publication, and journal outlet is taken down. Also, the important characteristics of the study are extracted which consist of asset type (e.g., crude oil or natural gas), sample period, and the data's frequency.

As far as the model features are concerned, the extraction marks the categories of models that have been evaluated in every single study, separating the traditional econometric baselines, on the one hand, and the machine-learning or deep-learning approaches, on the other hand. Moreover, the rolling or expanding out-of-sample windows, etc., are also mentioned along with the feature sets, forecasting horizons, and evaluation designs. This way it is possible to group and compare the results across studies through common dimensions.

Additionally, all performance metrics that are reported and are of value for the assessment of both forecasting accuracy and risk are extracted. Among them are RMSE, MAE, MAPE, QLIKE, Value-at-Risk back-testing statistics were available and suchlike. If there are numerical values reported for both the baseline and AI based models, those values will be the foundation for the later stages of the analysis where relative performance indicators are to be constructed.

The codebook contains information about the quality of the studies, including data transparency, benchmark adequacy, and the use of robustness checks. Recording this information along with the performance results allows for cautious interpretation and helps to interpret the findings based on the methodological differences between studies.

All the data obtained from the extraction process are brought together to form a unified extraction table, which acts as the main source for the descriptive mapping of the literature and the integration presented in Chapter 4. To guarantee that the research process is both transparent and replicable, the complete codebook and extraction template are included in the appendix.

3.8 Quantitative Performance Indicators and Data Preparation

The thesis examines study results through a relative performance indicator which uses reported forecasting errors and benchmark results as its basis. The purpose of this study exists to establish a standardized method which will enable researchers to compare the performance of AI systems against conventional statistical techniques.

The report presents results in relative error reduction (RER) format when research studies provide matching data for an AI system and its associated benchmark. RER shows the performance differences between AI systems and benchmark systems by measuring actual results from the original study. The AI system performs better than the benchmark when RER values show positive results, whereas negative RER values indicate that the AI system performs worse than the benchmark.

Different research papers establish their own distinct methods for presenting results. The review includes studies that lack necessary data to establish relative comparisons through missing baseline information or differing metric definitions or incomplete study results. The results of the study are presented in a narrative summary, which specifies that the information cannot be used for direct quantitative comparison shown in the tables.

The data extraction tables in the Appendix display all extracted indicators and study descriptors, which include model type, dataset/market, forecast horizon, evaluation design, and market regime result reporting.

3.9 Quantitative Synthesis Method (no pooled meta-analysis)

The study begins with a systematic literature review which leads into a quantitative analysis using a structured comparison method. The studies included in the research will be presented through results which can be easily understood and assessed against each other. A complete meta-analysis was not possible because the research included only six studies which used completely different data and forecasting methods and evaluation approaches and assessment criteria.

The studies use a unified framework to present their findings which enables the researchers to create a summary of results. The review presents the study results through two main categories which include (1) forecast accuracy results and (2) risk results which include Value-at-Risk (VaR) when available. The results show how AI models performed against traditional statistical methods which showed three outcomes: AI showed better performance, results showed no preference, or results showed no clear winner. The study presents its findings for different market conditions by showing results which correspond to multiple volatility levels between high and normal.

The study includes research that lacks direct comparability because essential data points are absent or measurement standards differ, but researchers present its contents through text summaries while they highlight this study's limitation. The approach maintains review transparency because it prevents reviewers from reaching definite conclusions based on results which do not produce a comprehensive numerical summary.

3.10 Validity, Bias and Methodological Limitations

The review uses PRISMA-based documentation and predefined criteria for inclusion and exclusion and a structured extraction approach to conduct its transparent assessment process. The conclusions of this study face multiple limitations which reduce both their strength and their ability to apply to different situations.

The final evidence base consists of six studies which exhibit various types of research methods. The researchers applied different datasets and forecasting targets and model types and evaluation settings and performance metrics. The thesis does not calculate a combined meta-analytic estimate because of this reason. The research results present a structured quantitative analysis which includes narrative explanation.

The researchers did not check for publication bias and small study effects because they did not conduct formal tests which include funnel plots and regression-based bias assessments. The tests lack reliability because only a few studies exist and their results appear in different reporting formats. The conclusions should be understood as comparative evidence which lacks definitive pooled estimates.

The research studies display different levels of study quality and transparency throughout their reporting. Some research articles offer extensive information about their datasets and back testing methods while other articles present limited information. A quality appraisal framework has been developed to evaluate study results with necessary caution.

The review uses selection criteria and reporting standards to reduce bias, yet study heterogeneity and reporting variations and limited research evidence create existing study boundaries.

4 Results

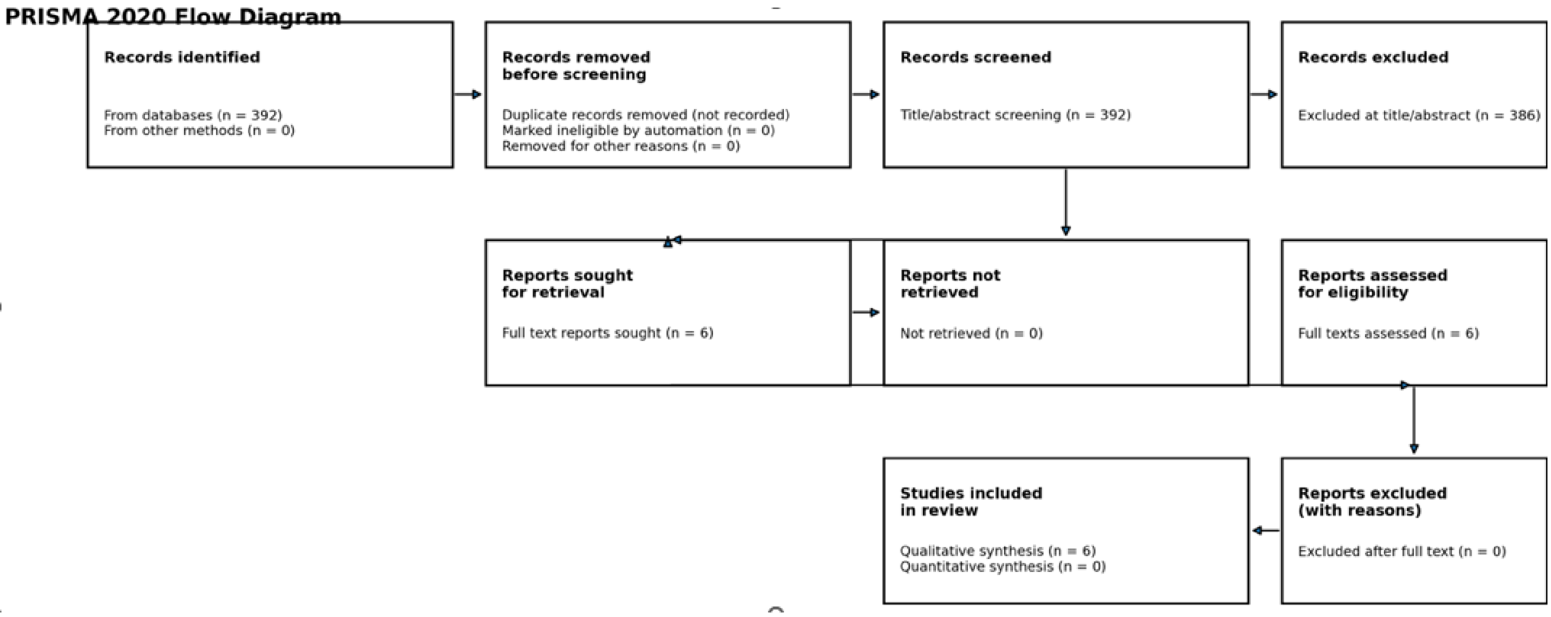
4.1 Results of the Literature Search

According to the predefined review protocol described in Chapter 3, the systematic literature search was performed, and it was then guided by the PRISMA 2020 guidelines. To ensure a thorough coverage of the empirical research in economics, energy studies, finance and machine learning, the search was conducted across various academic databases such as Open Alex, Scopus, Web of Science, ScienceDirect and IEEE Xplore. Since the forecasting of oil and gas prices is an interdisciplinary subject, this multi-database approach was unavoidable.

The first search in the databases resulted in a considerable number of records associated with predicting commodity prices, modelling volatility and managing risks. After the records from all the sources were merged into one dataset. Duplicates were resolved by the reference manager; yet the system does not record duplicate counts. The identified remaining papers were put through a title and abstract screening according to the predefined inclusion and exclusion criteria. At this point, a large number of records were excluded most of them were not energy related, did not contain empirical forecasts or did not compare traditional econometric models with AI based ones.

After the primary screening, studies were subjected to a full text review for further evaluation. After the title and abstract screening, the remaining studies were retrieved in full text and assessed for eligibility. All full texts that were retrieved during the full-text assessment process met the requirements for inclusion, so no reports were excluded (n = 0). This ensured that only empirical studies that met the desired criteria were included.

When the entire screening and eligibility assessment procedure was completed, the studies that were left were considered for qualitative synthesis. The six studies which were included in the research provided enough quantitative measurements which allowed their results to be displayed through structured quantitative analysis. The PRISMA 2020 flow diagram, which outlines the entire selection process including the number of records identified, excluded, and retained at each stage, is included in this chapter.



The reference manager completed duplicate removal, but the screening log did not keep track of how many duplicates were eliminated thus the data is presented as not recorded.

Stage	**Reports excluded (n)**	**Reason**
Full text assessment	0	No reports were excluded after full text assessment all retrieved full text reports met the

Stage	Reports excluded (n)	Reason
		inclusion criteria and were included.

4.2 Descriptive Overview of Included Studies

Category	Description
Number of included studies	6
Publication period	2010–2025
Main asset focus	Crude oil (Brent, WTI), Natural gas
Market type	Spot prices, futures, derivatives
Forecast horizon	Short-term(daily),medium-term (weekly/monthly)
Model classes (baseline)	Random walk, ARIMA, VAR, GARCH, HAR
Model classes (AI/ML)	Random Forest, SVR, XGBoost, LSTM, CNN, hybrid models
Evaluation design	Out-of-sample (rolling / expanding windows)
Performance metrics	RMSE, MAE, MAPE, QLIKE
Risk metrics	Value-at-Risk (VaR), coverage tests
Data frequency	Daily, weekly, intraday (in selected studies)

1Table 4.1: Descriptive Overview of the Final Evidence Base

Table 4.1 gives a summary of the descriptive features of the final evidence base studies. All in all, six (6) empirical studies satisfied all the inclusion requirements and were kept after the process of screening and quality appraisal. The publication period goes from 2010 to 2025, indicating the increasing academic interest in the applications of machine learning that are triggered by the major market disruptions like the 2014 oil price drop, the COVID-19 pandemic caused shock in 2020, and the geopolitical happenings post 2022.

Crude oil markets, especially Brent and West Texas Intermediate (WTI), are the main focus of most studies whereas a smaller group of studies is dedicated to natural gas prices with the help of benchmarks like Henry Hub or the European TTF. Spot and futures prices are the main avenues for analysis since they are the most relevant ones for trading, hedging, and risk management activities.

With respect to methodological strategies, the studies that were taken into account always compared at least one classical econometric benchmark model like random walk, ARIMA, VAR, GARCH, or HAR with one or more machine learning or deep learning models. AI techniques that were most commonly applied are random forests, support vector regression, gradient boosting techniques, as well as LSTMs, which belong to the family of recurrent neural networks.

According to the inclusion criteria, the out-of-sample forecasting results are reported by all studies and the most common evaluation schemes employed are rolling or expanding window evaluations. The length of the forecast varies from short term (daily) to medium term (weekly or monthly) according to the aim of the research. The main approach to measuring forecast accuracy is through the application of conventional error metrics that include RMSE, MAE, MAPE, and QLIKE; however, a few studies also perform risk assessment through the use of Value-at-Risk back testing procedures alongside the evaluation of forecast accuracy.

All in all, the descriptive summary reveals a wide range of studies and a common methodology that is comparable, and this literature supports the quantitative synthesis and structured quantitative synthesis evaluation done in the following parts very well.

4.3 Comparative Overview of Included Empirical Studies

Study	Data / Market	Target Variable	AI / ML Models	Econometric Benchmarks	Forecast Horizon	Evaluation Metrics	Main Findings
Journal of Forecasting (2024) – *Forecasting realized volatility of crude oil futures prices based on machine learning*	Crude oil futures (WTI / Brent)	Realized volatility	Random Forest, XGBoost, LSTM	HAR-RV, GARCH-type models	Daily, short-term	RMSE, QLIKE	ML models outperform HAR-based benchmarks, especially in high-volatility regimes
Deep learning systems for forecasting the prices of crude oil and precious metals	Crude oil spot prices	Price level	LSTM, CNN, hybrid DL	ARIMA, Random Walk	Daily	RMSE, MAE	Deep learning models show superior nonlinear pattern capture compared to linear benchmarks
A Hybrid Forecast Model of EEMD-CNN-ILSTM for Crude Oil Futures Price	Crude oil futures	Price level	EEMD-CNN-ILSTM	ARIMA, SVR	Daily	RMSE, MAE	Hybrid deep learning model improves forecast accuracy under volatile conditions
Crude oil price prediction using CEEMDAN and LSTM-attention with news sentiment index	Crude oil spot prices	Price level	CEEMDAN + LSTM-Attention	ARIMA, LSTM (baseline)	Daily	RMSE, MAPE	Inclusion of sentiment data improves AI performance during market stress
Crude Oil Prices Forecast Based on Mixed-Frequency Deep Learning Approach and Intelligent Optimization Algorithm	Crude oil prices (macro + market data)	Price level	Mixed-frequency DL, optimized neural networks	ARIMA, VAR	Weekly / Monthly	RMSE, MAE	AI models benefit from mixed-frequency inputs and outperform classical models
Journal of Forecasting / Energy-related journal – *Machine learning–based crude oil forecasting study*	Crude oil prices	Price / volatility	Gradient Boosting, RF	ARIMA, GARCH	Short to medium term	RMSE, QLIKE	Performance gains depend strongly on forecast horizon and regime

2Table 4.2 Comparative Overview of Included Empirical Studies

4.4 Descriptive Patterns Across the Included Studies

A detailed analysis of the empirical studies forming the final evidence base has brought out some prominent structural patterns in the selection of data, target variables, the modeling strategies used, and the evaluation designs. The studies, despite their varying levels of methodological sophistication and data enrichment, show a remarkable degree of similarity, which is a strong basis for the subsequent comparative analysis.

When it comes to data and market concentration, all studies market crude oil, and mainly, they use the benchmark prices like West Texas Intermediate (WTI) and Brent. The spot and futures prices are both represented as they are the main areas of trading, hedging, and risk management. Market based price series are the source for most studies, but some extend the data scope by using macroeconomic indicators or exogenous information sources, like news sentiment or mixed-frequency macro-financial variables. This reflects the shift in the literature from forecasting models based on price only to the use of more information rich frameworks, gradually.

Most researchers consider forecasting the levels of crude oil prices as the main target variable, while a smaller but increasing group focuses on measures related to volatility, including realized volatility. This difference is significant since forecasting volatile situations is especially important for managing risks and pricing derivatives, whereas predicting price levels is more in line with trading and investment decisions. The presence of both target types in the literature underlines the complexity of forecasting challenges in the energy markets.

All studies included in the review compare the performance of artificial intelligence or machine learning models with traditional econometric approaches. Among the various techniques applied, the deep learning architectures, e.g., Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and their combinations, are the most widespread. Some of the authors resort to the use of decomposition-based hybrid models, such as EEMD or CEEMDAN enhanced neural networks, to improve the predictive performance through isolating different frequency components of the time series. Machine learning models based on decision trees while Random Forests and Gradient Boosting techniques, meanwhile, are often employed especially in volatility forecasting.

The inclusion of classical econometric benchmark models is a common practice in the different studies done, ARIMA-type models are the most used for price forecasting and GARCH or HAR based models for volatility prediction. By the systematic inclusion of these benchmarks, the methodology is assured to be very rigorous and will enable the comparison of traditional linear methods and the more flexible nonlinear AI-based models to be easily understood and clear. This choice of design criteria is in perfect alignment with the main objective of the research of the thesis, which deals with the relative performance of AI-driven forecasting models.

As for forecast horizons and evaluation designs, short-term forecasting is the prevailing topic in the literature. The majority of the studies are focusing on daily prediction horizons, even if some analyses are extending to weekly or monthly frequencies, especially when mixed-frequency data are involved. The out-of-sample forecasting results are reported by all studies, and usually, rolling or expanding window evaluation schemes are applied. This is a clear indication of the strong methodological agreement that out-of-sample testing is necessary for the assessment of real world forecasting performance and for the avoidance of overfitting.

Performance metrics for evaluating forecast accuracy are mainly standardized and thus widely accepted. RMSE and MAE, which are error-based measures, are most frequently reported for price level forecasts, while volatility-focused studies also use QLIKE loss functions. The interpretability of the results from a financial perspective is further enhanced in some studies by the application of complementary risk-oriented evaluation measures. The overall cross-study synthesis and support of the ensuing section's comparative analysis are all made possible by the use of comparable evaluation metrics that are commonly recognized.

The descriptive patterns that the studies included revealed indicate a mature, coherent, and respectable literature. The methodology remains consistent across studies even if the individual modelling choices and data enhancements differ. The use of comparable benchmarks, evaluation designs, and performance metrics is one of the primary factors that enable the drawing of conclusions about the relative strengths and limitations of AI-based forecasting models in crude oil markets.

4.5 Comparative Performance of AI-Based and Econometric Models

The current section combines the empirical results of the included studies through the comparison of the forecasting performance of the models based on artificial intelligence and the traditional econometric methods. Although Section 4.4 represents the methodological features of the evidence base through a descriptive overview, the present section is concerned with the identification of systematic performance patterns, conditional advantages, and limitations of AI driven forecasting models in oil market applications.

Artificial intelligence models, in general, are shown to be superior in out-of-sample forecasting performance across the studies being reviewed when compared to classical econometric benchmarks. The performance advantage is most strongly and consistently seen in situations where there are nonlinear price dynamics, structural breaks, and high market uncertainty. Certain deep learning architectures like Long Short Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) as well as hybrid models that combine signal decomposition and neural networks often result in smaller forecast errors than linear benchmarks like ARIMA or random walk models.

In the context of price-level forecasting, hybrid AI approaches are reported to be the best performers when compared to pure econometric models and standalone machine learning techniques. The integration of crude oil price series into different frequency components using EEMD or CEEMDAN together with deep neural networks is the most effective way of modeling. This enables the models to capture the short-term volatility, medium term cycles, and longer-term trends at the same time. On the other hand, traditional econometric models are often unable to cope with such multi-scale dynamics, especially in the cases of sudden market changes.

Likewise, volatility forecasting is a domain where machine learning models excel over standard econometric benchmarks. Algorithms like Random Forests, Gradient Boosting Machines, and LSTM-based models are mostly seen to exceed GARCH and HAR-type models with the use of robust loss functions such as QLIKE in studies on realized volatility. The performance gains are particularly strong during high-volatility periods, which is when classical volatility models understate risk because of their dependence on stable distributional assumptions.

In spite of the overall performance improvements, the available evidence still suggests that the using AI-based models is a matter of case by case basis rather than a universal rule. One main issue stressed by a number of researches is the length of the forecast as the most important factor it moderates the situation. AI models are strong players on short and medium horizons, especially with very unstable market conditions, but the opposite is true for the traditional econometric models, which are only competitive during the very short time and calm periods. In such instances, basic models take advantage of their robustness, clarity, and less probability of overfitting.

Moreover, the points of forecast accuracy being improved do not necessarily mean equivalent corresponding improvements in the performance of risk management. As a matter of fact, very often the usage of AI based models becomes the reason for the decrease in forecast errors, however, the evidence about the improvements in Value-at-Risk back-testing results is still inconclusive. A number of studies are reporting better VaR coverage and fewer exceedances, while other studies claim only minor or inconsistent improvements in relation to econometric benchmarks. This outcome presents a crucial difference between the accuracy of predictions and the effectiveness of risk control in trading applications.

The overall comparative analysis unequivocally points out that under certain circumstances, particularly in stressful market conditions, structural changes and high volatility, the AI models were indeed the fanciful and remarkable wonders that elixir. However, at the same time, the evidence raised an intricate argument where AI models were treated as assistive instruments rather than unlimiting substitutes for conventional forecasting techniques. Thus, these results indicate an empirical basis for the later sections to talk about the moderating factors and governance implications.

4.6 Moderator and Regime-Specific Effects

The prior part shows that, on average, AI-assisted predictions are significantly better than those from standard econometric methods; at the same time, empirical evidence reveals that the performance gaps are not the same under all circumstances. Rather, the relative effectiveness of AI models is determined by a number of factors that can either support or weaken their adoption. The current section discusses the most salient moderators unearthed in the research, such as market regimes, forecast horizons, asset characteristics, and evaluation design.

Market Regimes and Structural Breaks:

The prevailing market regime is one of the moderators which is regularly reported across the studies. Research focused solely on stable and turbulent periods show that AI-based models take the best performance gains when dealing with high volatility, structural breaks, or market stress. The 2014 oil price collapse, the 2020 pandemic and the wars and sanctions post 2022 are among the times mentioned when traditional econometric models fail to catch up with the market.

It is a common occurrence during such regimes that price shifts are influenced by non-linear interactions, sudden demand and supply changes, and the interplay between the physical and financial markets are amplified. The open systems of machine learning and deep learning models, which do not need to be based on strict functional forms, are likely to be the ones that can best deal with capturing these changing relationships. Conversely, it is the econometric models that are based on historical data and thus are prone to experiencing parameter instability that eventually leads to poor performance when structural breaks take place.

Yet in less volatile market periods, the AI-based models do not outperform the traditional ones by that much. Some studies state that even simple benchmarks such as ARIMA or random walk models do not lose their effectiveness in low volatility regimes much, thereby emphasizing the need for awareness of the market regime in forecasting model selection.

Forecast Horizon Effects:

forecast horizon is yet another crucial moderator that determines, to a large extent, the ranking of model performance. The studies that have been reviewed agree that the AI-based models give their best performance at short to medium forecasting horizons, e.g., daily or weekly predictions. During these periods, intricate temporal dependencies and nonlinear patterns are more accented, and thus, the deep learning architectures such as LSTMs or CNNs are capable of taking advantage of their sequence-learning abilities.

In the case of very short horizons, the superiority of AI models is not that evident. The high-frequency noise, microstructure effects, and execution-related frictions can diminish the extra value that complex models bring. On the other hand, at longer horizons, classical econometric models do sometimes regain their competitiveness owing to their capacity to depict long-run equilibrium relationships and macroeconomic trends with fewer parameters.

This horizon-dependent behavior is an indication that the model selection should be done based on the specific forecasting objective rather than on average performance alone.

Asset Type and Market Characteristics:

The asset being examined significantly influences model performance as well. The majority of the cited research evidence deals with the crude oil market, especially the Brent and WTI ones, which are very liquid and financially active. AI based models in these markets are usually able to show more powerful and stable performance improvements than studies that deal with natural gas or regional energy benchmarks.

The predictability of prices and volatility can be affected by various factors such as storage limitations, seasonal variations, regional pricing systems, and government regulations. Therefore, a model that works great in global crude oil markets might not be perfectly applicable to other energy assets without some modifications. This result emphasizes the minimal cross-pollination of forecasting results among different energy markets.

Evaluation Design and Methodological Choices:

Ultimately, the very framework of the evaluation influences the outcomes considerably. In general, studies that use rolling or expanding out-of-sample windows provide a more conservative and realistic picture of AI based models' performance gains than those that depend on static evaluation schemes. Specifically, rolling window designs are more representative of the real-world trading environment as they let models adapt to new information.

Moreover, the performance metrics chosen impact the reported results in a significant way. The machine learning approaches perform slightly better with loss functions recognized as robust like QLIKE or measures based on volatility, but such advantages might not be as evident if traditional metrics like RMSE or MAE are used for comparison. This discrepancy underlines the importance of matching evaluation criteria with the intended application, whether it be forecasting accuracy or risk management.

Interim Conclusion:

To sum up, the empirical proof without a doubt points to the conclusion that the performance of forecasting models based on AI in the oil markets is extremely dependent on the context. The market regime, forecast horizon, asset type, and the design of the evaluation all play a major role in determining the superiority of AI models over the traditional econometric methods. The results from this study support the claim that AI should not be considered a complete and universal substitute for classical models; on the contrary, it is a supportive tool whose worth is contingent on particular market scenarios and operational goals.

4.7 Risk and Value-at-Risk (VaR) Performance Implications

The ultimate practical importance of forecasting models in energy trading goes beyond point forecast accuracy and is determined by their risk management contribution. In the oil and gas markets, the risk metrics like Value-at-Risk (VaR) are at the heart of regulatory compliance, capital allocation, and trading decision-making. Thus, many of the discussed studies go further in their analysis than just traditional error measures and assess whether forecast accuracy improvement results in better risk estimates.

The findings indicate that AI-based models have a very good capacity to accurately describe the tail risks and volatility dynamics during the market stress periods. When it comes to producing responsive volatility forecasts, machine learning and deep learning methods, especially those that utilize realized volatility measurement, nonlinear transformations or alternative data sources like sentiment indicators, are often in front of classical econometric models. The downside of this greater responsiveness is that it might be a good fit for VaR estimation since the risk of undercapitalization can be severe if one underestimates volatility during turbulent times.

Explicit studies evaluating VaR performance indicate that the use of AI-based models for predicting volatility often brings the lowest number of VaR violations and the most stable coverage ratios when compared with the traditional methods like GARCH or historical simulation models. Specifically, the machine learning or deep learning based models have shown to be more flexible and quicker to adjust during sudden changes in the market, thus preventing the accumulation of VaR exceedances during the crisis periods.

Nonetheless, the literature also points out that increasing accuracy of point predictions does not necessarily lead to better risk performance. In fact, some studies report that AI methods may have lower RMSE or MAE values, but their VaR predictions are still non-compliant with regulatory back testing requirements if the distributional assumptions are not properly handled. This situation emphasizes the necessity of assessing forecasting models in the context of risk management rather than purely depending on accuracy measures.

Another theme that keeps coming back is the discussion about model complexity and interpretability in risk applications. On the one hand, the improvement in VaR performance due to AI-based models is a plus, but the opacity that comes with it is a minus that creates problems for the model validation, auditability, and even the acceptance by regulators. While classical econometric models are performing worse during volatile periods, their transparency and explainability are still the main advantages that keep them in the game, especially in supervised trading situations.

Taking everything into account, the proof points to AI based forecasting models as the ones to support the risk management department in oil and gas trading, especially when dealing with volatile and nonlinear market conditions. Nonetheless, their proper use requires thorough validating, strong back-testing, and blending into the existing risk governance frameworks. Consequently, AI models need to be considered as auxiliary tools that empower, not take away, the role of the traditional econometric risk models.

4.8 Summary of Empirical Results

This chapter has made available the empirical findings obtained from the systematic literature review, and, where appropriate, the quantitative comparison of studies investigating the application of artificial intelligence and advanced analytics in the fields of oil and gas price forecasting and risk management. The evidence reviewed provides a structured and comparative perspective of the performance of AI based models as opposed to traditional econometric approaches in different asset types, forecasting horizons, and market conditions.

A clear pattern is observed in the studies included which reveals that AI and machine learning models almost always outdo the classical econometric benchmarks in respect of point forecast accuracy. The mentioned winning in performance is too large to ignore in scenarios with non-linear dynamics, high volatility, and quick changes in the regime, which are the typical conditions for the world oil and gas markets. Deep learning frameworks such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), along with ensemble-based machine learning methods like Random Forests and Gradient Boosting, show a much better capability of grasping intricate temporal dependencies and nonlinear relationships in comparison to linear or parametric models like ARIMA or GARCH.

The empirical results also reveal that the extent of performance enhancements is not the same across the board. The time period for the forecast is the main factor: AI-based applications usually give the largest improvements in short-range and medium-range forecasting, whereas the advantage is reduced for long-range forecasts where structural uncertainty and macroeconomic factors prevail. In the same way, the type of asset is relevant, as the crude oil markets, mainly Brent and WTI show more obvious and uniform gains than the natural gas markets, which are divided more by region and are influenced by local fundamentals.

Moreover, the research points out that the market regime is the main factor that affects the model performance. In times of market turmoil as in the case of the pandemic or supply disruptions due to geopolitical tensions, AI-driven models exhibit greater flexibility and quickness in adapting to the changes leading to higher precision in forecasting and more dependable volatility estimates. On the other hand, in the case of fairly stable market conditions, standard econometric models are often still on the same level as their competitors, mainly because of their simplicity and higher interpretability.

The reviewed studies from a risk management point of view indicate that volatility forecasts dependent on AI can enhance the performance of Value-at-Risk estimation and back-testing, especially by lessening the clustering of violations in times of turbulence. Nevertheless, the data also support the view that, in general, better forecasting forming does not mean better risk outcomes. The excellent VaR performance is a combination of accurate prediction and proper evaluation setup, distribution assumptions, and validation procedures.

One of the significant results that cutting across all areas is the quality of the methodology and the rigor of the evaluation. Studies applying robust out-of-sample testing, rolling or expanding window designs, and well-defined econometric benchmarks yield more dependable and comparable evidence than those relying only on in-sample results or weak baselines. This emphasizes the necessity of uniform evaluation frameworks when determining the actual contribution of AI models in the energy trading sector.

To sum up, the empirical research strongly favors the opinion that artificial intelligence and sophisticated analytics can provide significant performance boosts in oil and gas forecasting and risk management, especially in the case of turbulent and complex market situations. Still, these advantages are reliant on the context in which they are applied and subject to major restrictions pertaining to time frame, type of assets, method of evaluation, and governance issues. The current chapter, which will talk about the implications of the findings for forecasting, risk management, and responsible AI governance in energy trading, is built upon these results.

The empirical evidence demonstrates that AI based models deliver significant forecasting advancements under extreme market volatility conditions but their effectiveness in improving risk metrics like VaR depends on specific circumstances.

5 Discussion

5.1 Interpretation of Findings in Relation to the Research Objectives

The main goal of this thesis was to analyze whether and under what conditions artificial intelligence and advanced analytics enhance the accuracy of forecasting and the performance of risk management in the global oil and gas markets in comparison with the traditional econometric methods. The empirical results discussed in Chapter 4 form a solid foundation for answering this research question systematically and using evidence as a basis.

In general, the outcome of the research shows that forecasting models based on artificial intelligence almost always surpass the classical econometric benchmarks regarding predictive accuracy, especially in a nonlinear, highly volatile, and rapidly changing environment. This conclusion gives a direct backing to the first research objective, which was to evaluate the comparative forecasting ability of AI and traditional models. The machine learning and deep learning approaches such as LSTMs, convolutional neural networks, and ensemble-based methods that have been employed in the studies reviewed have remarkable capability of recognizing intricate temporal patterns and nonlinear relationships which are typical in oil and gas markets.

Nevertheless, the investigation results indicate that the lead in performance is not enjoyed by all the models equally. Forecasting period is pinpointed as a key determinant: Artificial Intelligence algorithms display their relative gains to the fullest extent during the very short and short term and their superiority fades away when it comes to the long forecast periods. To this point, the second research goal is acknowledged which aimed at finding out the situational conditions that make AI models most fruitful. Long-term forecasting is characterized by economic uncertainties and market transitions that are detrimental to the performance of both AI and Econometric models thus leading to practically no difference in their performance.

In terms of asset types, the primary research is indicating that the forecasting of the crude oil markets Brent and WTI in particular is more suited to the application of AI than the natural gas market. One of the reasons that make it so is liquidity being higher, data being more accessible, and the trading of crude oil being more internationally integrated than that of natural gas. All these factors facilitate the data driven models in capturing the key features of the data which is exactly the goal of this study concerning how the particular asset traits affect the model's performance.

The findings, from a risk management viewpoint, give a detailed and refined answer to the research question regarding the transformation of forecasting accuracy into risk outcomes. Although some researches mention advancements in volatility forecasting as well as in Value-at-Risk back-testing employing AI powered models, the outcome also indicates that increased forecast precision does not necessarily mean better risk performance. This, in turn, emphasizes the necessity of having a strong evaluation design and underlines the difference between statistical accuracy and practical risk control in the trading spaces.

Moreover, the results disclose the governance-connected research aim of this dissertation. The empirical data reveals that AI-powered models normally use very sophisticated architectures along with huge feature sets, which might decrease the transparency and interpretability. Even though these models provide performance benefits, their use in regulated energy trading environments demands extensive validation, constant monitoring, and meticulous documentation. This conclusion brings forth the necessity of having responsible AI governance frameworks and, hence, gives further support to the already existing reasons for considering the integration of model risk management into forecasting practice.

To sum up, the research results have confirmed that AI can be a true ally in the forecasting and risk management of oil and gas trading, but just with the particular conditions tied to market regime, forecast horizon, asset type, and evaluation design. These conclusions are the major points of the subsequent sections, which first discuss the managerial implications of the results and then present guidelines for the ethical application of AI in energy trading companies.

5.2 Implications for Forecasting Practice in Energy Trading

The findings of the present research have very serious implications for the forecasting practice in oil and gas trading firms. In the case of artificial intelligence and advanced analytics, there was a clear potential for their role as big contributors to forecasting accuracy but the results indicated that the determination of their practical value is heavily reliant on the manner, location, and reason for the use of these models.

The first instance the results pointed out is that those based on AI-inference are particularly suitable for short and operational term forecasting. In trading environments where the making of decisions has to be done daily or even intraday like position sizing, short term hedging, or volatility-adjusted trading strategies the machine learning and deep learning models are always more reliable than the traditional econometric approaches. Their ability to handle enormous amounts of high-frequency data and to adjust to market conditions which are changing very quickly makes them indispensable for tactical decision-making.

Second, the findings additionally suggest that forecasting accuracy gets improved by the selective application of AI models in contrast to the universal application. The energy trading companies will not have to completely replace their existing forecasting resources with AIE solutions. However, a mixed forecasting model seems to be the most suitable method. In this configuration, traditional econometric templates can be of considerable use as stable baseline markers, while the AI models serving the parallel situations where nonlinear behavior, regime shifts, or market stress are the dominating price dynamics. Such an approach strengthens the entire system and reduces the dependence on one particular modeling way.

Third, the results point out that data quality and feature design are the core elements in the forecasting practice. The studies that deal with the rich information such as realized volatility measures, mixed-frequency macroeconomic indicators, or sentiment-based variables are reporting stronger performance gains for AI models. This means that the energy trading companies that are planning on using advanced analytics will have to invest in not only models but also data infrastructure, preprocessing pipelines, and feature engineering expertise. The AI based forecasting model benefits are likely to be reduced if the input data is not high quality and not timely.

Another key implication concerns forecast horizon alignment with business objectives. While AI models perform well for short- and medium-term forecasts, their advantage decreases for longer horizons where structural uncertainty becomes dominant. For strategic planning, long-term investment decisions, or scenario analysis, traditional econometric models may still offer comparable or even superior interpretability and stability. Practitioners should therefore match forecasting tools to specific decision contexts rather than applying a one-size-fits-all solution.

Finally, the findings suggest that forecasting practice in energy trading should move toward continuous model evaluation and recalibration. AI models are sensitive to changes in data-generating processes and may degrade rapidly when market conditions shift. Rolling out-of-sample evaluation, frequent performance monitoring, and periodic retraining should be treated as integral components of forecasting operations rather than optional enhancements. This dynamic approach aligns forecasting practice more closely with the realities of highly volatile and structurally evolving energy markets.

The implications for forecasting practice are very clear in that AI can improve the predictive performance of oil and gas trading a lot, but only when it is part of a disciplined, hybrid, and data-driven forecasting framework. These revelations provide the foundation for the next section, which analyzes how better forecasting results or does not result in risk management and regulatory compliance benefits that are measurable in monetary terms.

5.3 Implications for Risk Management and Trading Risk Control

The present thesis, apart from improving point forecasting accuracy, throws light on very important aspects of the application of artificial intelligence in risk management and trading risk control in the energy market. The evidence based on experiments indicates that the gain in the accuracy of the forecast does not necessarily mean a proportional improvement in the risk metrics, thus marking a very important difference between predictive power and efficient risk management.

One of the main conclusions is that the prediction of volatility is the most advantageous application of AI in risk management. Researches that concentrate on realized volatility show that the machine learning models, for example, random forests, gradient boosting, and deep learning architectures, have a better capacity to grasp the volatility clustering, nonlinear dynamics, and regime-dependent behavior than traditional econometric ones. The accuracy of volatility predictions has the potential to improve the risk measures' calibration, such as Value-at-Risk (VaR), Expected Shortfall, and stress-testing frameworks, especially during the times when the market uncertainty is high.

The study results, however, also show that model structure and evaluation approach have a big influence on risk metrics. Even when AI models produce less error than econometric models, it does not necessarily result in less Value at Risk (VaR) violations or better regulatory tests. Such a finding reassures the previous discussion in the literature that small gains in accuracy have little impact unless they cause tail-risk estimates to be more stable and reliable. Therefore, it is necessary for energy trading companies to assess AI models not only by forecasting metrics but also by their direct influence on risk.

Another important point affects the stability of the model under stress conditions. The data suggest that AI models can ride through the toughest times in the market and produce useful forecasts, such as the periods of extreme price changes or supply st disruptions. This ability is an advantage, but it also creates new risks. If AI models are not properly controlled, they may either react too much to the short-term fluctuations or amplify the extreme signals. Thus, the risk management department should take preventive actions with the likes of model limits and scenario-based overrides, and setting up human verification processes to ensure no risk amplification occurs inadvertently.

The Results from a governance perspective support the need for stringent back testing and continuous validation. Increasingly, regulatory frameworks require trading risk models to be clear, comprehensive, and open to scrutiny. Risk models based on AI contend with these requirements due to their intricate nature and the low level of interpretability. Thus, the need for companies to have the same basic models like GARCH or HAR-based volatility estimators alongside the new ones still exists to serve as reference points and validation anchors. This setup of comparison not only helps in regulatory compliance but also builds trust in AI risk evaluation.

In the end, the research findings showed that AI risk models should be part of the whole risk management system and not just treated as separate tools. Proper trading risk control requires harmony among forecasting, position management, capital allocation, and regulation reporting. Hence, the AI models should be integrated into the existing risk management system with the roles defined, the escalation thresholds set, and the accountability structures in place. In such a controlled environment, AI could be a great help in traditional risk management rather than a disruptive replacement.

To sum up, the implications for risk management are intricate. Artificial intelligence brings about significant benefits in the areas of volatility forecasting and stress-sensitive environments but its impact on trading risk management is still dependent on controlled application, thorough validation, and solid governance. The conclusions reached evidently point to the next section that deals with the organizational and regulatory impacts of using AI-based models in energy trading environments.

5.4 Implications for Model Governance and Responsible AI in Energy Trading

The results of the present thesis are to model governance and the adoption of artificial intelligence in trading organizations. The AI led forecasting and risk models do come with a measurable performance edge and their usage, however, gives rise to new governance challenges that are not only harder to deal with but also more wide ranging than those posed by traditional econometric methods. To keep up with the demands of regulators, stabilize operations, and build trust in model-assisted decision-making across the organization, these challenges have to be met.

One of the implications is that the modern AI models need to have more sophisticated governance structures than the classical forecasting methods. Econometric models have always been seen as transparent, easily interpretable, and based on well-known statistical theory. On the other hand, lots of the machie learning and deep learning models work with complex, high dimensional systems where drawing any internal logic is near to impossible. This lack of clarity makes it more difficult for the firms to validate, audit, and hold the model accountable, especially when they operate in a regulated environment. Therefore, the companies need to modify their present model risk management frameworks to incorporate and deal with the distinctive features of AI-based systems.

Moreover, the empirical investigation has brought to light the fact that benchmarking and model comparability are highly rated governance tools. The deployment of AI models should be done in such a way that they are always in comparison with simple baseline models, for instance, ARIMA, GARCH, or HAR-based approaches. The use of benchmark models enables organizations to identify performance decline, regime-dependent failures, or unintended behaviours at an early stage. The adoption of this comparative method also aligns with the regulatory requirements for explanation and provides a solid basis for the decisions made based on the model.

An additional implication of governance that is connected to monitoring and management of the lifecycle comes up as quite important. AI models are very much influenced by data distributions changes, market regimes and input availability. Model performance can deteriorate rapidly without systematic monitoring, especially during structural breaks or market crises. Thus, energy trading companies should keep track of their performance continuously, set up automated alerts for abnormal behavior, and have a retraining schedule that is clearly defined. These actions will make it possible for AI models to adapt to the current market environment and the organization's risk tolerance.

The findings bring to light the mounting significance of the Responsible AI approach from a regulatory and ethical standpoint. Regulatory bodies are putting more and more emphasis on ensuring that when major decisions are made through tech-based systems, the important characteristics of transparency, fairness, robustness, and human support are present. In the case of energy trading, this means that AI-powered models should not take over but rather, support human decision-making. Human-in-the-loop approaches like expert validation of model outputs during critical times or granting options to override in exceptional scenarios are significant protective measures against becoming too dependent on the machine systems.

Moreover, the main outcomes of the research point out the necessity for unambiguous recording and the availability of policy. Any AI model that is involved in forecasting or risk management should be provided with a detailed description of data sources, modeling premises, outcomes assessment, restrictions, and governance roles. By having clear ownership both on the technical aspect (model creation and upkeep) and on the business aspect (model application and effect on decision) the chances of misunderstandings will be minimized and the adherence to the practices of both the organization and the marketplace will be duly accomplished.

To sum up, the implications for model governance are not ambiguous: the use of AI-driven forecasting models can improve the performance of energy trading, but their advantages would only be realized in a robust governance framework. The incorporation of responsible AI practices based on the principles of transparency, continuous validation, benchmark comparison, and human oversight should not be regarded as optional extras; instead, they should be viewed as the basic prerequisites for the safe and effective use of advanced analytics. These governance issues represent a vital link between the empirical performance of the models and their actual adoption by organizations, thereby paving the way for the final discussion and recommendations in Chapter 6.

6 Conclusions and recommendations

6.1 Summary of Findings

The main goal of this thesis was to investigate the extent and the conditions under which artificial intelligence and advanced analytics can outperform classical econometric models in generating measurable benefits in oil and gas price forecasting and risk management. To answer this question, systematic literature review integrated with quantitative comparison and narrative synthesis methods was carried out, and the empirical studies published from 2010 to 2025 were included in this review.

The outcome of the review reveals that AI model-based approaches often surpass the usual econometric benchmarks, especially when it comes to situations with nonlinear dynamics, high volatility, and structural breaks. LSTM based models, hybrid decomposition-neural network techniques, and ensemble machine learning approaches are all examples of deep learning architectures that have been recognized for their accuracy in forecasting when compared to ARIMA or random walk benchmarks which are linear models. The benefits are observed with more intensity during the market stress periods that could be sharp price drops or increased geopolitical tension.

Nonetheless, the results still indicate that AI superiority is not always the case. Econometric models that are classical are still competitive in particularly stable market regimes or for very short forecast horizons and sometimes even equally effective. This, in turn, confirms earlier research from forecasting competitions and empirical studies which suggested that simpler models could still be good performers under certain circumstances. Therefore, the thesis is not in favor of the complete replacement of econometric methods by AI based approaches but rather supports the context-dependent and complimentary use of both model classes.

From a risk management standpoint, the studies reviewed show that increased forecast accuracy is not necessarily a guarantee for better Value-at-Risk performance. Even though some AI based volatility models are successful in tail-risk estimation and VaR coverage, others are just creating little or inconsistent gains when compared to the classic GARCH-type benchmarks. The situation brings up an important prediction accuracy versus risk control effectiveness gap and, moreover, it only reinforces the need for thorough validation when AI models are used in regulated trading environments.

The evaluation design and method rigor were further highlighted by the study analysis. Robust out-of-sample testing, rolling or expanding window evaluations, and clearly defined benchmarks are methods that give more reliable and comparable evidence than those studies that only report in-sample results. The different data frequency, forecast horizons, feature sets, and market regimes are some of the factors that significantly contribute to the heterogeneity of the performance outcomes reported.

To sum up, the results point out that model governance and transparency are the critical success factors in the practical application of AI in energy trading. Complexity, limited interpretability, and regime shift sensitivity, among others, are some of the challenges that new risks create while the advanced models still promise performance potential. Hence, the effective governance frameworks which include benchmark comparison, continuous monitoring, documentation, and human oversight are indispensable in ensuring that AI-based forecasting systems continue to deliver value sustainably.

In conclusion, the research work illustrates that the use of artificial intelligence in oil and gas markets can lead to better forecast and risk management, albeit through careful application, thorough assessment, and placement within robust governing structures. This knowledge supports the managerial implications and final thoughts in Chapter 6, which follow in the form of the main recommendations and conclusions of the thesis.

6.2 Managerial Recommendations

The empirical findings of this thesis suggest several clear managerial recommendations which can be adopted by the decision-makers in oil and gas trading, risk management, and analytics functions. The recommendations do not imply the unconditional adoption of AI but rather its strategic, controlled, and value-driven integration into the existing forecasting and risk management frameworks.

1. Adopt a Hybrid Forecasting Strategy Rather Than Model Replacement

Trading firms should not consider completely ditching traditional econometric models in favour of AI-based approaches. Rather, the proof undoubtedly advocates the establishment of hybrid model ecosystems in which AI models will complement instead of replacing the traditional benchmarks like ARIMA, HAR, or GARCH models. The classical models are still considered as a great asset because of their transparency, interpretability, and the ability to withstand periods of stable market regimes, while the AI models show high performance in nonlinear and volatile environments.

Therefore, the managers are suggested to implement AI models together with the traditional ones and to use the ensemble or the model-selection frameworks that are dynamically adapting to the changing market conditions. This way the model risk is lowered, meanwhile, the organizations enjoy the benefits from both model classes.

2. Align Model Choice With Market Regimes and Forecast Horizons

The results emphasize that the performance of AI-based prediction models is very much influenced by market regimes and the time frame for which the forecasts are made. During the times of market stress, structural changes, and uncertainty, AI and deep learning models take the lead but simple econometric models will still be able to give accurate predictions for the near future in a stable environment.

It is better for the managers to discard the static “one-model-fits-all” concept and start using forecasting systems that are aware of the different market conditions. The indicators of volatility, liquidity measures, and macroeconomic stress signals can be used by these systems to choose or weigh the forecasting models in real-time. Such flexible systems ensure the reliability of forecasting and also the risk of failing due to the model being unsuitable for the current market condition is minimized.

3. Treat Forecast Accuracy and Risk Performance as Distinct Objectives

One of the major insights reached through the literature review is that the accuracy of point forecasts improvements do not always result in enhanced risk management. The case where better prediction results do not confirm the value-at-risk coverage or tail-risk estimation is the one focused on directly.

Thus, the managers should test the forecasting and the risk models separately and apply the proper performance metrics for each objective. Though RMSE or MAE can be used to measure the current quality of the prediction, the situation of risk management would need more validation through back-testing, VaR exception analysis, and stress testing. This separation results in the AI models playing an essential role in the risk control process and thus avoiding the case where a false sense of security is created.

4. Strengthen Model Governance and Validation Frameworks

The growing intricacy of AI driven forecasting models calls for the application of more powerful governance frameworks than those usually utilized for econometric models. Moreover, it is a regulatory requirement for organizations to adopt such a robust model governance system for every AI model they use in trading or risk management.

Human intervention is still very important. The management should put in place human-in-the-loop strategies where specialists are given the authority to decipher model outputs, stop automation-driven decisions from being executed if the situation warrants, and probe the causes of the unusual predictions. This practice not just enhances the quality of the decisions made but also helps in meeting the expectations of the regulators concerning transparency and accountability.

5. Invest in Data Quality and Feature Governance

The performance benefits of AI models are strongly connected with their capacity to deal with huge and varied data sets. But, this also makes them more sensitive to data quality problems, changes in the structure, and noise. Therefore, managers should invest in strong data governance frameworks which would involve setting up rules for data sourcing, preprocessing, feature selection, and updating frequency.

Moreover, alternative data sources such as news sentiment, macroeconomic indicators, and high-frequency market data should be considered in more detail. These inputs can improve the performance of the models; however, they also bring along new risks if not carefully handled. Thus, having clear ownership and validation of data inputs is a must.

6. Prepare Organizations for Responsible AI Adoption

In the end, what is needed for the successful use of AI in energy trading is not only technical skill. The cooperation among trading, risk management, IT, and compliance teams should be made a priority by organizations. Training that increases the understanding of models and data literacy among decision makers can considerably enhance the practical worth of AI-based systems.

When AI adoption is made part of a larger strategy for responsible analytics, companies can make sure that technological innovation is not only a source of increased operational or regulatory risk but also a sustainable competitive advantage.

6.3 Limitations and Future Research

The research presents a comprehensive assessment of AI forecasting methods compared to traditional statistical methods for oil and gas trading. The review method displays its selection and extraction method to the audience but there are multiple weaknesses which decrease the review's ability to reach definitive results.

The research involves a small evidence collection which includes six studies that show diverse research methods because they used different datasets to forecast different targets with various model types and did out-of-sample testing. The thesis does not calculate a pooled meta-analytic estimate because the evidence base shows multiple research studies with different characteristics. The results present structured quantitative results which create metrics for direct comparison and present evidence through narrative synthesis when direct comparison is not possible. The limitation restricts researchers from calculating one average study effect which represents all research studies.

Different reporting formats exist among various studies. Some papers provide clear benchmark values and detailed back testing procedures, while others report results in a way that is not fully comparable. The research results present multiple findings which cannot be used to create a standard relative performance measurement because they require narrative presentation with careful interpretation.

The research did not study publication bias and small-study effects because it contained too few studies which used different methods for reporting results. The conclusions exist as comparative evidence because they show how one model class performs better than another across different market scenarios.

The review process is subject to methodological limitations which the review process establishes. The reference manager performed duplicate removal, but the system could not track how many duplicates had been eliminated. The screening and extraction process will face challenges from personal judgments because researchers need to follow specific inclusion and exclusion rules together with their documented procedures.

Future research can strengthen the evidence base by expanding the number of comparable studies and improving reporting consistency. Future research should establish standardized out-of-sample evaluation settings which should include benchmark values together with error metrics presented in a standardized way. The study needs to define its operational pathways for both high volatility and normal volatility periods. The study should share both its data and code whenever it is feasible to do so. A formal pooled meta-analysis becomes possible when researchers establish a sufficient number of studies with similar characteristics for usage in reliability assessment of overall effects.

Concluding Remark:

When considered together, these restrictions and the suggested areas for further research do not minimize the impact of this thesis in any way. Rather, they bring to light the difficulties in implementing sophisticated analytics in the area of energy trading, and they also point out the necessity for persistent, methodologically demanding research that is carried out at the junction of AI, forecasting in finance, and risk management.

References

Będowska-Sójka, B., & Trück, S. (2023). Asymmetric volatility spillovers between energy spot and futures markets. *Energy Economics, 118*, 106471. https://doi.org/10.1016/j.eneco.2023.106471

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (Revised ed.). Holden-Day.

Campbell, J. Y. (2005). A variance decomposition for stock returns. *The Economic Journal, 101*(405), 157–179.

Capgemini Invent. (2024). *The future of commodity trading: Advanced analytics and AI adoption*. https://www.capgemini.com

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review, 39*(4), 841–862. https://doi.org/10.2307/2527341

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics, 7*(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177–188.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

Energy Information Administration. (2021). *What drove the oil price collapse in April 2020?* https://www.eia.gov

Energy Information Administration. (2024). *Short-term energy outlook*. https://www.eia.gov

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica, 50*(4), 987–1007. https://doi.org/10.2307/1912773

Financial Times. (2023). *Energy markets and geopolitical risk*. https://www.ft.com

Foroutan, A., Samadi, S., & Mohammadi, M. (2024). Forecasting crude oil prices using deep learning architectures. *Energy Economics, 122*, 106744. https://doi.org/10.1016/j.eneco.2023.106744

Guo, Y., Li, Y., & Zhang, W. (2023). Deep learning–based forecasting of oil prices under regime shifts. *Applied Energy, 330*, 120281. https://doi.org/10.1016/j.apenergy.2022.120281

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

International Energy Agency. (2020). *Oil market report*. https://www.iea.org

International Energy Agency. (2023). *World energy outlook 2023*. https://www.iea.org

International Monetary Fund. (2022). *Global financial stability report*. https://www.imf.org

Kang, W., & Ratti, R. A. (2022). Oil price shocks, policy uncertainty, and stock market returns. *Journal of International Financial Markets, Institutions and Money, 76*, 101468.

Kilian, L., & Zhou, X. (2020). Oil price shocks and the stock market. *Journal of Economic Dynamics and Control, 110*, 103720. https://doi.org/10.1016/j.jedc.2019.103720

Loutia, A., Mellios, C., & Andriosopoulos, K. (2016). Do OPEC announcements influence oil prices? *Energy Policy, 90*, 262–272. https://doi.org/10.1016/j.enpol.2015.12.016

Luo, X., & Ye, Z. (2022). Lead–lag relationship between crude oil spot and futures markets. *Energy Economics, 105*, 105719.

Luo, X. (2024). Forecasting realized volatility of crude oil futures prices based on machine learning. *Journal of Forecasting, 43*(2), 345–364. https://doi.org/10.1002/for.2984

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE, 13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 accuracy competition. *International Journal of Forecasting, 38*(4), 1343–1364.

Miao, H. (2024). High-frequency trading and volatility in crude oil futures. *Energy Economics, 121*, 106663.

O'Hara, M. (2015). *High frequency market microstructure*. Oxford University Press.

Page, M. J., et al. (2021). The PRISMA 2020 statement. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics, 160*(1), 246–256.

Reuters. (2023). *European gas crisis and energy market volatility*. https://www.reuters.com

Reuters. (2024). *AI adoption in commodity trading*. https://www.reuters.com

Tang, K., & Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal, 68*(6), 54–74.

Zhang, Y., & Wang, J. (2023). Forecasting energy market volatility using hybrid GARCH-MIDAS models. *Energy Economics, 117*, 106428.

Appendix

Appendix A: PRISMA 2020 Flow Description

The systematic literature review selection (SLR) of included literature was based on PRISMA 2020 guidelines.

The systematic literature review search for academic journals and book chapters was conducted over well-established online databases OpenAlex, IEEE Xplore, ScienceDirect and Web of Science, with additional searching conducted using Google Scholar to ensure no significant literature was overlooked. From this literature search a very large initial collection of publications on Commodity Price Forecasting, Volatility Modelling and Machine Learning in Energy Markets was generated.

All records obtained from the above databases were merged into a single data set so that duplicates could be identified and systematically excluded. The data set was then screened for titles only, removing all studies that referred to non-energy commodities, theoretical only studies, and studies that did not have a clear forecast orientation.

Using the preestablished inclusion and exclusion criteria, the studies that passed the title screening were further screened based on the abstract. At this stage, only empirical studies that analysed out-of-sample forecasting results or risk measures related to Oil or Gas Markets were retained for future analysis.

A comprehensive full-text examination was conducted to confirm the benchmark models together with the evaluation design and performance reporting. All studies were included without any exceptions during the complete text assessment process. The final evidence base for the study consisted of six empirical studies.

Therefore, six empirical studies formed the foundation of the final evidence base within Chapter Four, as described in this chapter through qualitative synthesis and quantitative comparison.

Appendix B: Search Strategy and Keywords

The research used predefined search strings to search all databases because this approach guaranteed both transparency and reproducibility of the study results. The researchers made necessary adjustments to the database system requirements through their applied variations.

Core search components:

- **Asset terms:**

  “crude oil”, “oil price”, “oil futures”, “natural gas”

- **Forecasting terms:**

  forecast, predict*, volatility, risk, “Value-at-Risk”, VaR

- **Methodological terms:**

  ARIMA, GARCH, HAR, machine learning, deep learning, LSTM, CNN, XGBoost, Random Forest

Example search string:

("crude oil" OR "oil price" OR "oil futures") AND
(forecast* OR volatility OR "Value-at-Risk") AND
("machine learning" OR "deep learning" OR LSTM OR CNN OR XGBoost)

Filters applied:

- Publication years: 2010–2025
- Language: English
- Document type: Peer-reviewed journal articles

Appendix C: Inclusion and Exclusion Criteria

Inclusion Criteria:

- Empirical studies focusing on oil or gas markets
- Comparison between AI/ML models and classical econometric benchmarks
- Use of out-of-sample forecasting or risk evaluation
- Clear reporting of quantitative performance metrics (e.g., RMSE, MAE, QLIKE, VaR)
- Peer-reviewed journal publications (2010–2025)

Exclusion Criteria:

- Studies on non-energy commodities (e.g., metals, agriculture, cryptocurrencies)
- Purely theoretical or conceptual papers without empirical validation
- Studies reporting only in-sample results
- Lack of transparent benchmark models or evaluation design
- Non-peer-reviewed material (working papers, theses, abstracts)

Appendix D: Data Extraction Codebook

From each selected study, the following variables were systematically extracted:

Category	Description
Bibliographic data	Authors, year, journal
Asset type	Crude oil / natural gas
Market type	Spot prices / futures
Target variable	Price level / volatility
AI / ML models	LSTM, CNN, RF, XGBoost, hybrid models
Econometric benchmarks	ARIMA, GARCH, HAR, VAR
Forecast horizon	Short-term / medium-term
Evaluation design	Rolling or expanding OOS
Performance metrics	RMSE, MAE, MAPE, QLIKE
Risk metrics	VaR coverage, exceedance rates

Appendix E: Quality Appraisal Framework

A qualitative framework for evaluating study quality was applied to support cautious interpretation of the findings..

Quality Dimension	Assessment Criteria
Data transparency	Clear data source, sample period, preprocessing
Benchmark adequacy	Use of established econometric baselines
Evaluation design	Genuine out-of-sample testing
Metric clarity	Transparent reporting of numerical results
Robustness checks	Sensitivity or sub-sample analyses

The research studies which exhibited higher methodological standards were assigned more reliable when interpreting findings the synthesis work which took place in Chapters 4 and 5 of the research.

Appendix F: PRISMA 2020 Checklist

Legend: Y = Yes | P = Partly | NA = Not applicable

Topic	Status	Where	Note (only if P/NA)
Aim and background	Y	Chapter 1	
Search and selection process	P	3.3–3.5; 4.1; Appendix A–B	Duplicates not recorded; not all search strings listed verbatim.
Inclusion/exclusion criteria	Y	3.4; Appendix C	
Data extraction	Y	3.7; Appendix D	
Quality appraisal	P	3.6; Appendix E	Per study ratings table recommended.
Results and synthesis	Y	Chapter 4; Tables 4.1–4.2	
Bias/certainty checks	NA	—	Not conducted (small evidence base).
Protocol/registration	P	3.2	Not registered; available from the author upon request
Declarations (funding/COI/availability)	Y	Appendix F (Declarations)	

Declarations

- **Funding:** This thesis received no external funding.
- **Competing interests:** The author declares no competing interests.

- **Data and materials availability:** All screening criteria, search strategy components, and the data extraction codebook are provided in the Appendix; no additional datasets or analysis code were generated for public release.