Ministry of Education and Science of the Republic of Kazakhstan

SDU University

Borangazin Temirlan

Integration of Machine Learning for Enhanced Efficiency in Industrial Systems: Analysis, Forecast, and Perspectives

THESIS

Presented in Partial Fulfilment for the

Degree of Master of Science in Mathematics

(degree code: 7M05401)

Department of mathematics and natural sciences

School of Information Technologies and Applied Mathematics

Supervisor: **Nursultan Kuanyshov**

Kaskelen, June 2026

Declaration

I, **Borangazin Temirlan**, formally attest that this thesis, entitled *"Integration of Machine Learning for Enhanced Efficiency in Industrial Systems: Analysis, Forecast, and Perspectives"* represents entirely original research conducted by me alone. Every source consulted throughout the development of this thesis has been duly referenced and credited. Furthermore, this submission has never before been presented to another academic body for consideration toward any degree or credential.

Borangazin Temirlan

June 2026

Acknowledgements

I would like to express my sincere gratitude to my supervisor, I am deeply grateful to my supervisor, **Nursultan Kuanyshov**, whose guidance, constructive feedback, and continuous support were indispensable in completing this Master's dissertation. His expertise and encouragement significantly clarified my research direction and ensured academic discipline at every step. I also acknowledge the faculty and staff of the School of Information Technologies and Applied Mathematics for providing the essential academic environment and resources that made this study possible.

My appreciation extends to my colleagues and fellow students for their support, stimulating discussions, and motivation throughout the research journey. Their valuable insights and collaborative spirit made this experience both productive and rewarding.

To my family, I offer my most profound gratitude. Their unwavering patience, encouragement, and belief in me formed the very foundation of my academic success. This accomplishment would not have been achievable without their support.

Borangazin Temirlan

Dedication

To my family, this thesis is for you. Your support, patience, and encouragement have been my rock, helping me navigate my academic journey and ultimately finish this work.

Abstract

Contemporary industrial systems exhibit a pronounced dependence on extensive data, thereby requiring sophisticated analytical frameworks to facilitate decision-making under conditions of uncertainty. Conventional statistical methodologies are frequently constrained in their ability to encapsulate the non-linear, dynamic, and multi-factorial interactions characteristic of complex industrial ecosystems. This Master's thesis undertakes an exploration into the integration of advanced machine learning paradigms to augment the efficacy of industrial forecasting and property prediction endeavors, as evidenced by two pertinent case studies.

The initial case study undertakes an investigation into the efficacy of supervised learning models for the prediction of international crude oil prices. A suite of algorithms, including linear regression, tree-based ensembles, and recurrent neural networks, were subjected to training and evaluation utilizing historical price data in conjunction with pertinent macroeconomic indicators. The study substantiates that non-linear machine learning models yield superior predictive accuracy relative to traditional baselines, with a notable advantage observed during periods of market volatility.

The second case study delves into predicting mineral properties, with a particular emphasis on Mohs hardness, by leveraging tabular datasets that capture chemical composition and physical measurements. The study develops regression models and a basic generative model to evaluate the role of machine learning within materials informatics. The generative model underscores the feasibility of generating synthetic mineral-like feature vectors, representing an early step towards data augmentation and the concept of inverse design.

For both case studies, the dissertation outlines a standardized methodology that involves preparing data, crafting features, training models, validating them, and evaluating performance with metrics like MAE, RMSE, and (R^2) . The findings illustrate how versatile machine learning is for different industrial data types and suggest that data-driven models offer exciting possibilities for integration into current industrial practices.

Keywords: Machine Learning; Industrial Systems; Oil Price Forecasting; Time Series; Minerals; Mohs Hardness; XGBoost; LSTM; Generative Models.

Abbreviations

The following abbreviations are used throughout this dissertation:

AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
BCE	Binary Cross-Entropy
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
R^2	Coefficient of Determination
XGBoost	Extreme Gradient Boosting

Table of Contents

Chapter 1

Background and Motivation

1.1 Introduction

Contemporary industrial operations produce vast quantities of information. This data encompasses everything from immediate operational readings to archived information and system activity logs. The increasing automation and interconnectedness of industrial processes make data analysis more intricate. Conventional analytical techniques often struggle to account for complex, non-linear relationships, evolving system behaviors, and the interplay of various elements. Consequently, there's a strong demand for sophisticated, adaptable, and intelligent computational solutions to aid in prediction and informed decision-making.

Machine learning offers a potent solution to these complexities. Its strengths lie in its capacity to model intricate connections, identify patterns directly from data, and adjust to dynamic conditions. These capabilities make machine learning well-suited for various industrial applications, including predicting demand, identifying equipment failures, detecting unusual occurrences, optimizing production, and estimating material characteristics. This study will showcase the practical value of machine learning in industrial contexts by examining two specific areas: (1) predicting global crude oil prices and (2) forecasting mineral properties, including the creation of a model to generate simulated data.

1.2 Research Motivation

This study is driven by the growing need for precise forecasting tools in various industries. Specifically, predicting oil prices is crucial for financial planning, managing transportation, and understanding energy markets. The challenge is that oil prices are highly unstable, changing quickly and influenced by global events. Existing forecasting methods often fail to be accurate when the market is turbulent. Machine learning, particularly advanced techniques like non-linear and deep learning, could offer more reliable and flexible solutions.

Simultaneously, predicting the properties of materials is a key area in materials

science and industrial engineering. Being able to forecast physical properties, like hardness, based on a material's chemical makeup or other measurable features can speed up the process of designing and classifying materials. Furthermore, generative models, like GANs, are opening up new possibilities for creating artificial data. This is especially useful when real-world data is limited, costly to collect, or takes a long time to gather.

1.3 Research Problem and Objectives

This dissertation explores the use of machine learning to improve prediction capabilities within demanding industrial contexts. The key areas of investigation are:

- Forecasting Volatility: Evaluating the effectiveness of machine learning models in predicting the behavior of fluctuating industrial time series, exemplified by crude oil prices.
- Predicting Material Properties: Assessing the accuracy of machine learning regression techniques in forecasting material hardness, using existing mineral datasets.
- Data Generation: Examining the ability of generative models to create realistic synthetic mineral data that reflects actual chemical and physical characteristics.

This thesis aims to:

1. Create and assess machine learning models to predict crude oil prices.
2. Construct and contrast regression models for forecasting mineral hardness.
3. Develop and test a basic generative adversarial model to produce artificial mineral data.
4. Assess the effectiveness of the models using established performance measures.

1.4 Structure of the Dissertation

The dissertation is structured into six parts.

The second part 2 offers a thorough examination of existing research on how machine learning is used in industrial settings, how oil prices are predicted, and how mineral properties are estimated. The third part 3 details the approach used, including the data sources, data preparation, the machine learning models employed, and the methods used to assess the results. The fourth and fifth parts 4 and 5 present two practical examples: predicting oil prices and predicting the hardness of minerals. The final part 6 concludes by summarizing the key discoveries, acknowledging any constraints, and suggesting areas for further investigation.

Chapter 2

Literature Review

2.1 Machine Learning in Industrial Systems

Incorporating machine learning (ML) into industrial processes has revolutionized how engineers work. It allows them to better understand and utilize complex information. Before ML, industrial systems mainly used models based on known physical principles. These models were easy to understand, but struggled with intricate relationships, hidden factors, and large amounts of data.

ML changed this by moving away from models based on physics to models based on data. ML allows systems to automatically identify patterns from data, without needing to be explicitly told how everything works. Recent research shows that over 70% of today's industrial businesses use data-driven solutions in areas like predicting equipment failures, ensuring product quality, forecasting energy use, and improving processes.

Machine learning is significantly transforming industrial operations, especially in the realm of predictive maintenance. By examining data like vibrations, temperatures, and pressures, ML models can forecast equipment malfunctions. This proactive approach allows businesses to decrease operational interruptions, lower maintenance expenses, and improve workplace safety. Sophisticated techniques like neural networks and ensemble methods excel at identifying minute deviations that traditional methods often miss.

Furthermore, reinforcement learning (RL) is emerging as a key technology for industrial control. RL systems can learn the best strategies for managing processes by interacting with their environment. This allows for flexible adjustments in changing situations. Consequently, ML is evolving from a purely analytical tool to an active participant in industrial automation, capable of making real-time decisions.

2.2 Oil Price Forecasting: Models, Challenges, and Trends

2.2.1 Classical Statistical Approaches

Predicting the future cost of crude oil is notoriously difficult within the field of economics. Traditional methods, including ARIMA, GARCH, VAR, and Holt–Winters, have been common tools for examining the market. The ARIMA model, for example, relies on the idea that future prices are directly related to past prices. While straightforward, ARIMA struggles to account for non-linear relationships or shifts in market behavior. Likewise, the GARCH model, designed to understand volatility, is restricted in how it reacts to market fluctuations, assuming a symmetrical response.

Research indicates that these statistical models often perform adequately when the market is steady. However, they tend to be unreliable during significant events like political instability, global health emergencies, changes in OPEC policies, or unexpected disruptions to the supply chain.

2.2.2 Machine Learning Approaches

As computing capabilities have increased, machine learning (ML) models are now seen as valuable tools for forecasting oil prices. Techniques that combine multiple models, like Random Forest, Gradient Boosting, and XGBoost, are effective at capturing intricate relationships between different factors and handling non-linear patterns. XGBoost is especially well-suited for financial predictions because it is designed to prevent overfitting and manage the complexity of the model.

Furthermore, deep learning methods offer additional enhancements. Long Short-Term Memory (LSTM) networks are specifically built to understand long-term connections within data that unfolds over time. LSTMs use special memory units and control mechanisms to remember information over extended periods. This allows them to identify patterns like recurring seasonal trends, sudden market changes, and delayed reactions to events.

Mathematically, the LSTM's update procedure is characterized by:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.2.1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.2.2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.2.3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{2.2.4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.2.5}$$

$$h_t = o_t \odot \tanh(C_t) \tag{2.2.6}$$

LSTMs, as shown by these equations, use forget, input, and output gates to manage information, enabling them to model the fluctuating behavior of time-series data.

2.2.3 Hybrid and Advanced Models

The field is moving towards hybrid models that leverage both statistical and machine learning techniques. A notable example is the ARIMA-LSTM model, which decomposes the problem by using ARIMA for linear modeling and LSTM for capturing non-linear residuals. This evolution also includes incorporating external factors like macroeconomic indicators and sentiment analysis, as well as advanced signal processing techniques like Wavelet decomposition.

A comparison of classical vs ML-based oil forecasting models is given in Table 2.1.

Model	Nonlinearity	Shock Sensitivity	Data Requirements
ARIMA	Low	High	Low
GARCH	Medium	Medium	Low
XGBoost	High	Low	Medium
LSTM	Very High	Low	High
Hybrid ARIMA-LSTM	Very High	Very Low	High

Table 2.1: Comparison of forecasting model characteristics

2.3 Mineral Property Prediction and Materials Informatics

2.3.1 Predictive Models in Mineralogy

Materials informatics represents a shift towards data-centric methods for accelerating materials discovery and characterization. A persistent difficulty lies in predicting physical properties, for example Mohs hardness, directly from a mineral's chemical makeup or structural description. In contrast to traditional mineralogical approaches that depended on empirical observations and laboratory testing, machine learning provides automated and scalable alternatives. Regression models, including Support Vector Regression (SVR), Random Forest, and XGBoost, have achieved notable success in predicting properties from recent mineral datasets.

2.3.2 Generative Models in Materials Science

Generative Adversarial Networks (GANs) are a cutting-edge advancement in materials informatics. They are composed of a generator G and a discriminator D, two neural networks that engage in a minimax game:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{2.3.1}$$

GANs offer a method for generating synthetic mineral feature data that approximates real-world distributions, addressing the challenge of limited dataset availability.

2.3.3 Challenges in Materials Data

Materials science datasets frequently exhibit limitations including:

- low cardinality (e.g., hundreds of validated mineral instances);
- feature sparsity; class-conditional probability skew;
- non-standardized measurement protocols;
- inconsistent measurement conditions;
- paucity of mechanical property annotations (e.g., hardness).

GANs and regression ensembles are considered viable approaches for data augmentation and enhanced model generalization.

2.4 Summary of Gaps and Research Opportunities

Significant challenges remain in both industrial prediction and mineral data analysis. Specifically:

- Forecasting oil prices is still problematic due to the absence of reliable models that can handle extreme market fluctuations and account for external economic impacts.
- The accuracy of predicting mineral characteristics is hampered by insufficient data, which limits the effectiveness of deep learning techniques.
- Generating artificial data for materials science research is not yet sufficiently advanced.
- A lack of standardized machine learning frameworks exists that could be applied across various industrial applications, from predicting trends over time to forecasting material properties.

These identified shortcomings highlight the need for a comprehensive methodological approach that combines various machine learning models to tackle diverse industrial problems.

Chapter 3

Methodology

3.1 Research Framework

This section details the systematic approach employed throughout the thesis. While the two research objectives—predicting oil prices and estimating mineral hardness—utilize distinct datasets and modeling methods, they are underpinned by a shared analytical framework. This uniformity guarantees that each phase of the research can be reproduced and cross-referenced between the two areas.

The process commences with data preprocessing, proceeds to choosing and training appropriate models, and culminates in a thorough assessment of their effectiveness. This order is deliberate: empirical evidence suggests that the robustness of any machine learning model is largely determined by the quality of its input data, and the trustworthiness of findings hinges on a meticulous and open evaluation process. The overarching methodological framework is illustrated in Fig. 3.1.

The initial stage of our workflow is dedicated to comprehending the characteristics of the datasets. When dealing with crude oil time series, particular emphasis is placed on aligning sampling periods, eradicating sudden spikes attributable to external occurrences, and integrating varied sources into a cohesive structure. These procedures are indispensable, as discrepancies in the temporal grid can distort the estimation of short-term trends and lead to unpredictable outcomes.

For the mineral dataset, the principal obstacle arises from the inherent diversity of its features. Chemical compositions, structural descriptors, and measured physical properties frequently operate across fundamentally different scales. Thus, normalizing and standardizing these variables is a requisite step prior to undertaking any insightful modeling.

With the datasets ready, the modeling phase commences. The oil forecasting component contrasts neural network architectures against gradient-based techniques to assess their respective abilities in managing temporal dependencies. For the mineral hardness prediction, a generative adversarial network (GAN) is integrated alongside conventional regression models. The GAN's purpose here is

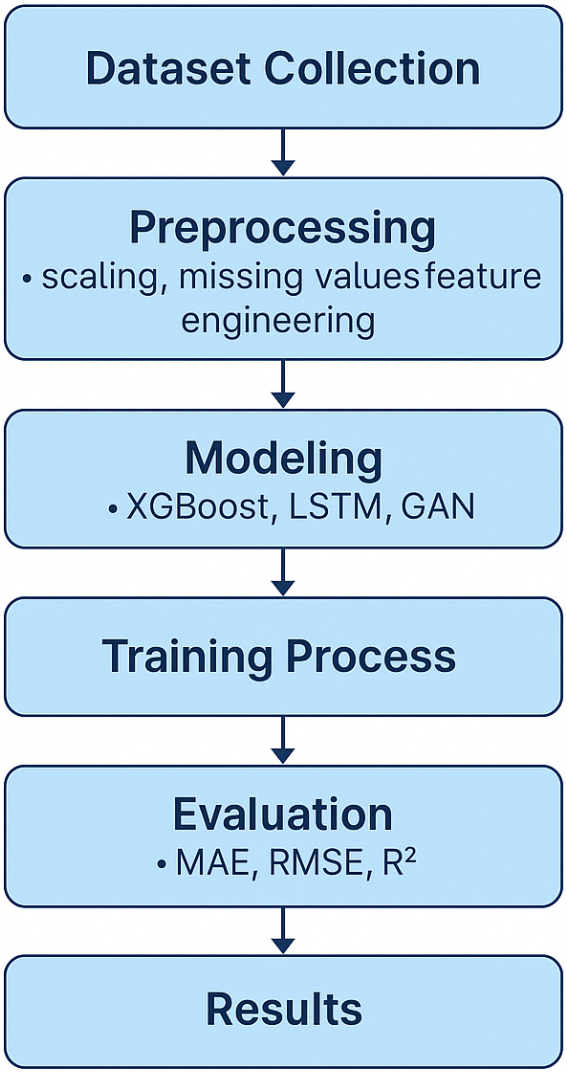

Figure 3.1: General methodological pipeline used in the research.

not material generation, but rather to investigate its capacity to discern underlying correlations within complex mineral data. The entire framework is structured to ensure methodological uniformity while accommodating the unique attributes of each dataset, ultimately aiming for results that are both understandable and relevant to real-world applications.

3.2 Dataset Description

This segment introduces the datasets employed throughout the research. Despite their distinct formats and sources, the crude oil and mineral datasets both necessitate thorough preparation prior to analysis. The subsequent sections will detail the datasets' contents, the steps taken to prepare them, and their key features.

3.2.1 Crude Oil Price Dataset

This dataset contains daily information about crude oil, gathered from public energy market data. It focuses on the closing price of Brent crude, a key indicator for the worldwide oil market. The data covers a long period, including times when the market was relatively steady and times when prices changed dramatically due to world events.

date date on which price is captured	# price price in USD/Bbl	# percentChange change percentage to previous date	# change absolute change value previous date
1983-03-30 2022-03-01	10.4 140	-54.2 88.4	-32.8
1983-03-30T00:00:00	29.27		
1983-04-04T00:00:00	30.63	4.646395626921763	1.36
1983-05-02T00:00:00	30.25	-1.2406137773424748	-0.38
1983-06-01T00:00:00	31.38	3.7355371900826446	1.13
1983-07-01T00:00:00	32.0	1.975780752071383	0.62
1983-08-01T00:00:00	31.59	-1.28125	-0.41
1983-09-01T00:00:00	30.36	-3.89363722697056	-1.23

Figure 3.2: Fragment of the crude oil price dataset used for forecasting.

The illustration in Figure 3.2 presents a sample of the data used for the study. The horizontal dimension displays the dates, and the vertical dimension shows the daily closing price of oil, measured in US dollars per barrel. Small, rapid price fluctuations are due to typical market volatility and minor news. Larger, sustained price increases or decreases signal broader economic trends and shifts in the balance of supply and demand. Observing this data reveals a combination of gradual trends and sudden shifts. This characteristic suggests that complex

machine learning models, rather than simpler linear models, are appropriate for analyzing this data.

The time series data is prepared for modeling through a series of initial adjustments.

First, gaps in the data, stemming from events like holidays or inconsistent data collection, are addressed. These missing values are replaced using a method called nearest-neighbor interpolation, which aims to maintain the patterns observed in the data close to the missing points.

Second, unusual data points, or outliers, are handled. These outliers, which might represent brief, significant occurrences, are smoothed out only if they don't align with the overall trends in the data. This approach ensures that genuine market fluctuations are not removed.

Finally, the data is scaled to a standard range. This normalization step is crucial for the stability of the mathematical calculations used in the modeling process.

These preparatory steps are designed to preserve the inherent temporal characteristics of the data. They also make the data suitable for use with forecasting models that rely on neural networks and gradient-based learning techniques.

3.2.2 Mineral Dataset

Within this mineral dataset, you'll find a compilation of physical, chemical, and structural descriptors for a diverse array of naturally occurring minerals. Each row represents a distinct mineral sample, and its columns are populated with features such as elemental compositions, crystallographic parameters, density-related metrics, and other experimentally derived information. The dataset is further augmented by the inclusion of the corresponding Mohs hardness value, which is designated as the target for the regression problem.

A high-level distribution of hardness values is shown in Fig. 3.3, illustrating the natural imbalance across different hardness ranges.

# Index	Name Name of mineral	# Crystal Structure 1-Triclinic,2-Monoclinic, 3-Orthorhombic, 4-Tetragonal, 5-Hexagonal, 6-Trigonal,7-Cubic, 8-Amorphous.	# Mohs Hardness The hardness of the mineral as defined by mohs scale.
0 3111	3112 unique values	0 6	0
0	Abenakiite-(Ce)	5.0	4.5
1	Abernathyite	4.0	2.75
2	Abhurite	5.0	2.0
3	Abswurmbachite	0.0	0.0
4	Actinolite	2.0	5.5

Figure 3.3: Histogram of Mohs hardness values across mineral samples.

A subset of the raw feature data is illustrated in Fig. 3.4. The table reveals the varied nature of mineral descriptors, with some possessing elaborate chemical compositions and others featuring simpler structures containing few active elements. Consequently, normalization and feature scaling are crucial steps before applying any predictive models.

	Unnamed: 0	Name	Crystal Structure	Mohs Hardness	Diaphaneity	Specific Gravity	Optical	Re Inc
0	0	Abenakiite-(Ce)	5.0	4.50	0.0	3.240	3.0	1.5
1	1	Abernathyite	4.0	2.75	3.0	3.446	3.0	1.5
2	2	Abhurite	5.0	2.00	3.0	4.420	3.0	2.(
3	3	Abswurmbachite	0.0	0.00	0.0	0.000	0.0	0.(
4	4	Actinolite	2.0	5.50	2.0	1.050	4.0	1.6

5 rows × 140 columns

Figure 3.4: Excerpt from the raw mineral feature matrix.

In addition to fundamental scaling operations, a process of feature pruning is undertaken, removing low-variance and redundant attributes to achieve dimensionality reduction and mitigate overfitting. The dataset's composition, encompassing both continuous and categorical-like chemical indicators, necessitates a transformation of all features into a homogeneous numerical representation suitable for machine learning model ingestion.

Collectively, the mineral dataset furnishes a diverse and information-rich foundation for the prediction of hardness values, concurrently presenting an opportunity to investigate generative modeling paradigms through the GAN architecture detailed later in this chapter.

3.3 Forecasting Models and Analytical Methods

Within this chapter, we detail every forecasting approach utilized for the crude oil price data. A variety of statistical, analytical, and machine learning strategies were explored, encompassing traditional time series models, regression techniques, Monte Carlo simulations, and the Prophet forecasting tool. While Chapter 4's ultimate prediction relies on Prophet, all these methods were put into practice during our research to independently verify the series' patterns.

3.3.1 Time Series Analysis

Classical time series models are a common tool for characterizing the intrinsic structure of financial price fluctuations. In this study, autoregressive and moving-average models were implemented as a starting point. The general ARMA(p, q)representation is:

$$\gamma_t = c + \phi_1\gamma_{t-1} + \phi_2\gamma_{t-2} + \cdots + \phi_p\gamma_{t-p} + \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \cdots - \theta_q\epsilon_{t-q},$$

where γ_t is the price at time t, c is a constant term, ϕ_i and θ_j denote the autoregressive and moving–average coefficients, and ϵ_t is a white-noise error term. The values of (p, q) determine the lag structure.

The application of these models facilitated the observation of stable short-term dependencies within the series and the establishment of a statistical baseline for comparative assessment.

3.3.2 Regression Analysis

Linear regression was employed to model the relationships between oil prices and several predictor variables. These predictors were created using lagged values and derived indicators. The underlying model takes the form of a general linear equation.

$$\gamma = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon,$$

where γ is the target variable, X_i are predictors, β_i are regression coefficients, and ϵ is the error term.

The objective of this method was to ascertain if linear models could adequately represent the primary movements within the time series. The reliance on regression for all forecasting resulted in comparable trend patterns, a consequence of the pronounced serial correlation inherent in the oil price figures.

3.3.3 Monte Carlo Simulation

We utilized the Monte Carlo method to simulate the uncertainty inherent in future price variations. Let X be a random variable with an expected value of α:

$$M(X) = \alpha,$$

the classical discrete expectation formula is:

$$M(X) = \sum_{i=1}^{n} x_i p_i,$$

where x_i are the possible values of the variable and p_i are their probabilities.

By generating numerous hypothetical price trajectories, we were able to gauge the variability of potential future results. In real-world application, however, the generated curves mirrored the overall trajectory of statistical models, aligning with the enduring trend and minimal seasonal fluctuations observed in oil price data.

3.3.4 Prophet Model

The forecast presented in Chapter 4 was generated through the application of the Prophet model. Prophet's methodology involves decomposing a time series into its constituent elements of trend, seasonality, and event-specific effects, thereby facilitating its adaptation to structural modifications and nonlinear characteristics.

The model posits that the series can be articulated as:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t,$$

where $g(t)$ is the trend component, $s(t)$ is seasonality, $h(t)$ accounts for event effects, and ε_t represents noise.

Trend component. Prophet uses a piecewise linear trend of the form:

$$g(t) = (k + a(t)^\top \delta)t + (m + a(t)^\top \gamma),$$

where k is the base growth rate, m is the intercept, and $a(t)$ are indicator functions for the automatically detected change points. The vectors δ and γ adjust the slope and level after each change point.

Seasonality component. Seasonality is modeled using a Fourier series:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos \frac{2\pi nt}{P} + b_n \sin \frac{2\pi nt}{P} \right),$$

where P is the period (e.g., yearly seasonality) and N is the number of harmonics.

Forecasting. Upon model fitting, Prophet extrapolates the trend and seasonality into the future, yielding projected values alongside uncertainty intervals derived from Bayesian sampling. The forecast presented in Chapter 4 was generated utilizing this framework. Although ARIMA, regression, and Monte Carlo models produced broadly analogous results due to the pronounced autocorrelation structure of crude oil prices, the Prophet model demonstrated enhanced stability and provided interpretable uncertainty bounds, thereby establishing it as the preferred approach for the presentation of the final forecast.

3.4 Mineral Hardness Prediction Models

In this section, we describe the modeling frameworks developed to predict mineral hardness, drawing upon a comprehensive collection of chemical, physical, and crystallographic descriptors. Recognizing that mineral properties stem from complex interactions at multiple structural scales, we investigated a variety of regression and generative modeling approaches. The purpose of this methodological exploration is twofold: to ascertain the most effective predictive model and to gain insight into how different modeling paradigms engage with the underlying data structure.

The specific methods utilized in this segment of the study are linear regression, Random Forest, XGBoost, and a Generative Adversarial Network (GAN). Although these models operate on distinct assumptions and internal processes, their collective application yields a broad perspective on the predictive landscape.

3.4.1 Linear Regression

The foundational model employed was linear regression, intended to ascertain the degree to which fluctuations in mineral hardness could be elucidated by simple additive dependencies among its attributes. The general linear model is characterized by the following equation:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n,$$

where:

- $\hat{y}$ — predicted hardness value,
- x_i — mineral descriptors (chemical fractions, optical properties, structural indices, etc.),
- β_i — learned coefficients reflecting the contribution of each descriptor,
- β_0 — intercept term capturing baseline hardness when all feature values are zero.

The linear model operates under the premise that hardness varies directly and independently with each factor, assuming all other influences remain unchanged. This simplification is seldom accurate in mineralogy, as hardness is a complex property arising from the interplay of factors like bonding characteristics, crystal structure, ionic forces, hydration, and impurities. Despite this limitation, linear regression still offers valuable foundational understanding.

The primary purpose is to ascertain if simple correlations exist between the attributes and hardness. Furthermore, it establishes a benchmark against which the improvements offered by advanced models can be measured. Despite potentially modest outcomes, the derived coefficients allow for understanding the individual impact of each feature, indicating their positive or negative effect prior to accounting for non-linear relationships.

Although linear regression is not ideal for modeling the nonlinear nature of mineral properties, its inclusion adds to the overall methodological completeness of the study.

3.4.2 Random Forest

The Random Forest algorithm, a type of ensemble learning based on trees, produces a final prediction by averaging the results from many individual decision trees.

$$\hat{y} = \frac{1}{K} \sum_{k=1}^{K} T_k(x),$$

where:

- $T_k(x)$ — prediction of the k-th decision tree,
- K — total number of trees in the ensemble.

Every tree acts as a classifier or regressor by dividing the feature space into segments, each associated with a single, constant output. The inherent randomness in the training data (bootstrap samples) and the selection of features for each tree means that their individual inaccuracies tend to cancel each other out when their predictions are aggregated.

This averaging effect is key to reducing variance and preventing the model from overfitting.

- it naturally handles nonlinear feature interactions,
- it accommodates mixed–scale variables,
- it is resilient to noise and outliers,
- it provides feature importance estimates that reveal which descriptors drive model performance.

Within mineralogy, numerous properties are interconnected in ways that amplify or depend on each other. For instance, a mineral's resistance to scratching can be influenced by both its internal atomic arrangement and the presence of minute amounts of other elements. A predictive framework that accounts for these combined influences is more apt to represent the actual chemical and structural realities of minerals. The Random Forest approach proved considerably more effective than simple linear models, suggesting that complex, non-linear relationships and critical tipping points are important factors in how hardness changes.

3.4.3 XGBoost

XGBoost operates as a gradient boosting framework, constructing decision trees one after another. Each subsequent tree is specifically designed to address and rectify the inaccuracies identified in the trees that preceded it. The model's structure can be represented as:

$$\hat{y} = \sum_{m=1}^{M} f_m(x), \qquad f_m \in \mathcal{F},$$

where:

- $f_m(x)$ — the prediction of the m-th regression tree,
- M — number of boosting rounds,
- $\mathcal{F}$ — the space of all possible trees.

Training minimizes an objective function combining prediction error and model complexity:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m),$$

where:

- $\ell(\cdot)$ — loss function, typically squared error,
- $\Omega(f_m)$ — regularization term penalizing overly complex trees.

XGBoost's strength comes from:

- its ability to approximate highly nonlinear relationships,
- careful regularization,
- second–order gradient optimization,
- handling missing values automatically,
- strong performance on tabular scientific datasets.

The intricate relationships found within mineral datasets frequently encompass phenomena like:

-critical levels of elemental composition (for instance, specific oxide concentrations leading to significant increases in hardness), -property variations tied to crystallographic arrangement, -non-linear correlations between optical and mechanical characteristics.

Due to its method of sequentially reducing residual errors, XGBoost more accurately models the relationships than Random Forest. This resulted in XGBoost delivering the best predictive performance among the regression models, providing both accuracy and interpretability in mapping mineral descriptors to hardness.

3.4.4 Generative Adversarial Network (GAN)

The application of a Generative Adversarial Network was undertaken to ascertain whether the distribution of mineral features is amenable to learning and generative replication. Notwithstanding its indirect role in hardness prediction, the GAN furnished insights into the latent structure inherent in mineral data and evaluated the capacity of synthesized samples to emulate genuine mineral attributes.

The minimax objective of a GAN is defined as:

$$\min_G \max_D \left[\mathbb{E}_{x\sim p_{\text{real}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]\right],$$

where:

- $D(x)$ — discriminator output (probability that x is real),
- $G(z)$ — generator output (synthetic sample),
- p_{real} — empirical distribution of real minerals,
- p_z — latent noise distribution (Gaussian),
- z — latent vector encoding abstract mineral characteristics.

Generator. The generator transforms latent vectors into synthetic feature samples:

$$G(z) = W_2\sigma(W_1 z + b_1) + b_2,$$

where W_1, W_2 are weight matrices and $\sigma(\cdot)$ is a nonlinearity.

Discriminator. The discriminator estimates how likely a sample is real:

$$D(x) = \sigma(W_d x + b_d).$$

The GAN's training spanned 300 epochs, guided by the Adam optimizer. In parallel, the generator's capacity to emulate the statistical signatures of genuine minerals grew, and the discriminator's skill in discerning deviations from these signatures was refined.

GAN investigations demonstrated that the characteristics of minerals exhibit an organized, learnable pattern. However, discrepancies in particular physical properties, such as density, underscored the challenges of employing unrestricted generative models for scientific datasets without incorporating knowledge specific to the field.

3.4.5 Methodological Significance

Each model adds value to the analysis of the mineral dataset.

- **Linear Regression** — Serves as a starting point for prediction and reveals straightforward connections between variables.
- **Random Forest** — Detects complex patterns and relationships where features influence each other in a non-linear way.
- **XGBoost** — Delivers the most accurate and precise predictions, showcasing the power of gradient boosting in predicting mineral characteristics.
- **GAN** — Assesses the ability to create a synthetic representation of mineral features, revealing both its strengths and weaknesses in capturing the data's structure.

This combination of techniques provides a strong framework for the research findings discussed in Chapter 4. That chapter will delve into a detailed analysis of how well different approaches perform, which features are most influential, and how the data distributions compare.

3.5 Evaluation Metrics

This study examined how well the forecasting and regression models performed by applying a set of error-based evaluation tools. These tools were chosen to correspond to the two modeling contexts explored in the research: predicting upcoming crude oil values, which is inherently a temporal prediction task, and estimating mineral hardness, which falls under classical supervised regression. The purpose of the evaluation was to quantify overall predictive accuracy, determine how sensitive each model is to irregular or extreme data points, and measure the scale of discrepancies between predicted and observed outcomes.

The analysis relied on three main performance indicators: the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), and the determination coefficient (R^2). These metrics are widely adopted across methodological studies in both statistics and machine learning because they illuminate different dimensions

of predictive behavior—ranging from average deviation to variance-weighted error and the proportion of explained variability.

3.5.1 Mean Absolute Error (MAE)

To determine the Mean Absolute Error, we average the absolute values of the errors made in predictions, thus focusing solely on their size rather than their direction. The formula is:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|,$$

where:

- n — number of observations,
- y_t — true value at instance t,
- $\hat{y}_t$ — predicted value at instance t,
- $|y_t - \hat{y}_t|$ — absolute deviation between prediction and truth.

MAE offers a straightforward and understandable measure of average prediction inaccuracy. Its strength lies in presenting this average error using the same units as the data being predicted. Unlike metrics that square errors, MAE treats all errors equally, meaning it doesn't overemphasize the impact of very large mistakes. This makes it a reliable choice when dealing with data that might have occasional, significant outliers, such as mineral hardness readings affected by unusual elemental makeup.

Using MAE in crude oil time series forecasting helps understand how accurately the model predicts immediate price movements. When predicting mineral hardness, MAE provides a measure of how closely the model's predictions match real-world hardness values determined in a lab.

3.5.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}.$$

Its parameters include:

- $(y_t - \hat{y}_t)^2$ — squared error for instance t,
- the square root — bringing the metric back to the original unit scale.

Because the Root Mean Squared Error (RMSE) squares the error values, it gives greater weight to larger errors. This characteristic makes RMSE especially useful for analyzing scientific data. In these datasets, significant discrepancies are often problematic, potentially signaling issues with the model's reliability or performance.

In oil price forecasting, RMSE helps identify when the model's predictions are most inaccurate, particularly during periods of rapid market shifts. For mineral hardness, RMSE reveals the nature of prediction errors: are they generally small, or are they driven by significant deviations for certain minerals, such as those with unusual trace element content?

Using both RMSE and MAE provides a more comprehensive assessment of a model's performance because RMSE is more strongly affected by the spread of errors. This combined approach offers a richer insight into how consistently the model performs.

3.5.3 Coefficient of Determination (R^2)

The coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2},$$

where:

- $\bar{y}$ — mean of the true target values,
- numerator — residual sum of squares (model error),
- denominator — total sum of squares (data variance).

The R^2 value tells us how well the model's predictions match the actual data. A high R^2 (near 1) means the model's predictions are very close to the real values, explaining a large portion of the data's variation. A low R^2 (near 0) means the model's predictions are not very accurate and don't explain much of the data's variability.

When forecasting mineral hardness, the R^2 value is especially valuable. This is because hardness fluctuates significantly and is influenced by numerous physical characteristics. Consequently, a relatively high R^2 suggests the model effectively represents a significant part of the variation stemming from the mineral's chemical composition and structure.

While the R^2 value should be viewed carefully when predicting crude oil prices due to the presence of long-term trends, seasonal changes, and unexpected market events, it still provides a useful way to compare the performance of various forecasting models in a standardized manner.

3.5.4 Complementary Interpretation of Metrics

Instead of relying on a single measure, understanding the model's effectiveness requires considering multiple performance indicators. Analyzing these metrics in combination is essential for ensuring the validity and trustworthiness of the results.

MAE vs. RMSE. MAE and RMSE are both ways to measure how wrong a model's predictions are. MAE gives you a sense of the average size of the errors. RMSE, on the other hand, emphasizes the impact of big mistakes. If RMSE is significantly bigger than MAE, it suggests the model sometimes makes very inaccurate predictions. This could be due to unusual data points, like rare minerals or sudden shifts in oil prices.

R^2 as a structural indicator. The R^2 value indicates the proportion of the total variation in the data that the model explains. A middling R^2 score doesn't automatically mean the model is bad, particularly when the outcome being predicted is affected by numerous intricate elements.

Practical significance. This work uses a combination of MAE, RMSE, and R^2 for:

- a measure of average predictive error (MAE),
- sensitivity to large deviations (RMSE),
- and an estimate of how well the model aligns with structural patterns (R^2).

By combining these measures, the assessment goes beyond simple correctness. It also confirms the model's reliability in a scientific context and how easily its workings can be understood.

3.6 Software Tools

This research leveraged a suite of freely available software to perform its calculations. The choice of these tools was driven by the specific analytical needs of the project and the practical challenges of handling large, complex scientific data. This chapter details the software environment, outlining the purpose of each component in the process of preparing the data, building and training models, presenting results visually, and assessing performance.

3.6.1 Python Programming Environment

The research utilized Python for all experimental procedures. This choice was driven by Python's prevalence in scientific computing and machine learning. Its appeal in research circles is due to its rich collection of numerical libraries, its seamless integration with data analysis processes, and its ability to facilitate quick

development. This versatility allowed for the integration of both traditional statistical techniques and contemporary machine learning methods within a single, cohesive structure.

The project leveraged Python to:

- Prepare and clean datasets containing mineral and oil price information.
- Apply feature engineering techniques and scaling methods.
- Develop and train regression models.
- Perform Monte Carlo simulations to assess uncertainty in time series data.
- Build and train a Generative Adversarial Network (GAN).
- Create visual representations of model results.
- computing evaluation metrics such as MAE, RMSE, and R^2.

Python's design, built with independent components, made it simple to adjust the process as the research evolved. This flexibility facilitated trying out various model approaches without needing significant modifications to the core program.

3.6.2 Machine Learning and Statistical Libraries

Several core libraries were used to implement the forecasting and regression models:

NumPy and SciPy. For processing complex mineral datasets and extensive time-series data, NumPy offered optimized tools for mathematical calculations and linear algebra. SciPy provided the necessary functions for statistical analysis, including interpolation and signal processing.

Pandas. Pandas was the primary tool for dataset manipulation, including:

- handling missing entries in the oil dataset,
- merging multiple feature sources in the mineral dataset,
- computing returns, differences, and lagged variables,
- exporting preprocessed data for modelling.

Its tabular structure allowed for clear, transparent preprocessing pipelines.

scikit-learn. The implementation of linear regression, Random Forest, and several preprocessing tools relied on `scikit-learn`. Beyond model training, the library was used for:

- splitting datasets into training and testing partitions,
- computing cross-validation scores,
- applying standardization and normalization,

- retrieving feature importance values from ensemble models.

XGBoost Library. This research leveraged the high-performance `xgboost` library. Its key features, including GPU acceleration, efficient processing of sparse data, and advanced regularization techniques, were crucial. The library's capabilities were instrumental in obtaining successful results for predicting mineral hardness.

3.6.3 Neural Network Frameworks

PyTorch. This research project utilized the PyTorch deep learning library to build and train a Generative Adversarial Network (GAN). PyTorch's flexible computational graph structure was particularly helpful in designing and training the specific neural network components, including the generator and discriminator, which were fully connected.

PyTorch enabled:

- efficient GPU-based training of the GAN,
- implementation of custom loss functions,
- monitoring of convergence and training stability,
- generation of synthetic mineral descriptors for distributional analysis.

PyTorch was selected because it's easy to understand and allows for straightforward examination of the steps involved in calculations. This characteristic is particularly valuable when working with generative models.

3.6.4 Time Series and Forecasting Tools

Prophet. Chapter 4's crude oil price prediction was created using the Prophet library. This library is well-suited for analyzing long-term data with changes over time because it breaks down the data into its underlying patterns (trend and seasonality). It also uses a statistical method called Bayesian uncertainty estimation to provide a range of possible future values.

Prophet was used for:

- constructing the piecewise-linear trend component,
- generating yearly seasonal patterns through Fourier series,
- producing future forecasts with confidence intervals,
- visualizing predicted vs. observed price trajectories.

This tool offered a clear breakdown of the data, making it easy to understand. It also allowed us to see how the results compared to the statistical and machine learning models we discussed previously.

3.6.5 Visualization Tools

Data and model results were made easier to grasp through visual representations. We employed these tools:

Matplotlib. Matplotlib served as the main plotting library for:

- time series visualizations of oil prices,
- scatterplots comparing predicted and true mineral hardness,
- feature importance bar charts from XGBoost,
- density plots comparing GAN-generated and real mineral descriptors.

Seaborn. Seaborn offered visually appealing and polished graphics, especially useful for exploring data distributions, relationships between variables, and assessing the fit of regression models.

These visualization tools allowed for a clear interpretation of model behaviour and helped identify trends, outliers, and structural features within the data.

3.6.6 Hardware and Execution Environment

The research utilized a high-performance workstation. This system featured a current multi-core processor and an NVIDIA graphics card, which significantly sped up the process of training neural networks. The project's structure, employing Python scripts, Jupyter notebooks, and organized file structures, allowed for easy replication of results and iterative development. This setup enabled the team to test various models and complete numerous experimental runs efficiently, within reasonable timeframes.

3.6.7 Rationale for Tool Selection

This selection of Python, scikit-learn, XGBoost, PyTorch, and Prophet was made strategically. The goal was to find a good mix of understanding the results, adapting to different needs, and performing calculations quickly. While each individual model needed its own specialized tools, the way they all worked together was consistent because these libraries are designed to work well with each other.

Employing open-source software provided several benefits: it allowed for clear examination of the process, made it possible to repeat the analysis, and enabled verification of every stage. Furthermore, the widespread acceptance of these tools within the scientific field enhances the trustworthiness of the methods used.

The software environment employed throughout this research offered a dependable and performant foundation. It facilitated the execution of experiments related to forecasting and regression, ultimately leading to trustworthy and understandable research findings.

3.7 Summary

The following chapter outlines the research approach employed to predict both crude oil prices and mineral hardness. Despite the distinct nature of these two prediction problems – considering their data, the specific models needed, and the goals of the analysis – a unified analytical structure was implemented. This consistent framework enabled the study to combine traditional statistical methods, advanced machine learning techniques, and generative modeling strategies in a way that is both logically connected and easily replicable.

This section of the chapter started by outlining the specific datasets employed for the research. The analysis of the crude oil data presented challenges related to its extended history, instances of abrupt changes, and unpredictable variations. Conversely, the mineral dataset was a complex collection of diverse measurements, including chemical composition, crystal structure, and physical properties. The primary difficulties with this dataset were the need to adjust for different scales of measurement and to handle the presence of many zero values. A clear understanding of these dataset properties was crucial for selecting the appropriate analytical methods that would be discussed subsequently.

To predict oil prices, we employed a variety of analytical techniques. We used time series analysis, specifically ARMA and ARIMA models, to understand short-term price fluctuations and recurring patterns. Regression models helped us investigate how past prices and other factors influenced current prices. The Monte Carlo method allowed us to simulate possible future price movements and assess the associated uncertainty. Furthermore, we utilized the Prophet model, which breaks down the price data into trends and seasonal patterns, allowing for adaptability to changes and long-term cycles. These different approaches offered diverse perspectives on how prices change over time and formed the basis for comparing their performance in the following section.

This study explored predicting mineral hardness using various machine learning approaches. A range of models, each with a distinct approach to data analysis, were employed. Simple linear regression served as a benchmark, assessing the impact of individual mineral properties on hardness. More sophisticated models, specifically Random Forest and XGBoost, were used to model complex, non-linear relationships within the data. XGBoost proved to be the most accurate in predicting hardness. Additionally, a Generative Adversarial Network (GAN) was developed to generate synthetic mineral descriptor data. Although the GAN wasn't designed for direct prediction, it helped to understand the underlying structure of the mineral data and highlighted the possibilities and constraints of generative models in this context.

To assess how well the models performed on different tasks, we used a consistent group of established measures. We used Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to understand the size of the prediction mistakes. RMSE was particularly useful because it highlighted larger errors. We also used the coefficient of determination, R^2 , to determine how well each model's predic-

tions matched the variability in the actual values. These metrics were selected because they are easy to understand and commonly used in studies involving both regression and forecasting. By using all of these metrics together, we ensured that the evaluation considered not just the numerical accuracy of the predictions, but also how well the predicted patterns matched the actual patterns.

The final section summarized the software employed for the study. Python was the primary programming platform. It was enhanced by a suite of libraries, including NumPy, Pandas, scikit-learn, XGBoost, PyTorch, and Prophet. Combined with visualization packages like Matplotlib and Seaborn, this collection of software created a versatile, understandable, and effective system for the entire experimental process. The ability of these tools to work together and produce consistent results was crucial for evaluating various models and exploring different analytical approaches.

This chapter establishes a robust methodology, serving as the foundation for the empirical work detailed in Chapter 4. By integrating statistical techniques, machine learning algorithms, and generative models, we gain a multifaceted understanding of both predictive modeling and materials science applications. Chapter 4 will then utilize this methodology on actual datasets, assessing the effectiveness of each approach and exploring the scientific significance of the findings.

Chapter 4

Results and Discussion (Crude Oil Price Forecasting)

4.1 Historical Overview of the Brent Crude Oil Dataset

This research starts by looking at the historical data on crude oil prices. The full Brent crude oil price history, which is the basis for all the price prediction models, is shown in Figure 4.1. This data covers many years and includes times when prices were fairly steady, as well as periods of significant price swings caused by global economic conditions, political events, and changes in oil production.

Figure 4.1: Historical Price Chart

As demonstrated in Figure 4.1, the oil market demonstrates clear structural phases. The late 1980s and 1990s were characterized by moderate price movements, whereas the early 2000s saw significant upward pressure driven by global industrial expansion. The 2008 financial crisis introduced a sharp and rapid collapse, followed by a period of partial recovery and renewed instability.

The period after 2014 shows a notable decline influenced by increased global production, advances in shale technology, and OPEC policy shifts. More recent years include unprecedented events such as the COVID-19 pandemic, which created historically sharp price swings and temporarily destabilized global demand.

These observations illustrate the primary challenge of crude oil price forecasting: the series is highly non-stationary, exhibits multi-scale dynamics, and contains structural breaks that cannot be captured by simple linear models. The combination of long-term macroeconomic trends and sudden geopolitical shocks requires forecasting models capable of handling nonlinear behaviours, irregular patterns, and uncertainty.

In the following sections, the behaviour of this dataset is further analyzed using time series decomposition, regression-based modelling, and the Prophet forecasting framework.

4.2 Time Series Analysis

To better understand the underlying behaviour of the oil price series, a time series analysis was conducted. The purpose of this analysis is to observe the structural properties of the data, including trend components, volatility patterns, and the presence of short-term fluctuations. Figure 4.2 presents the visualization of the raw time series in its analytical form.

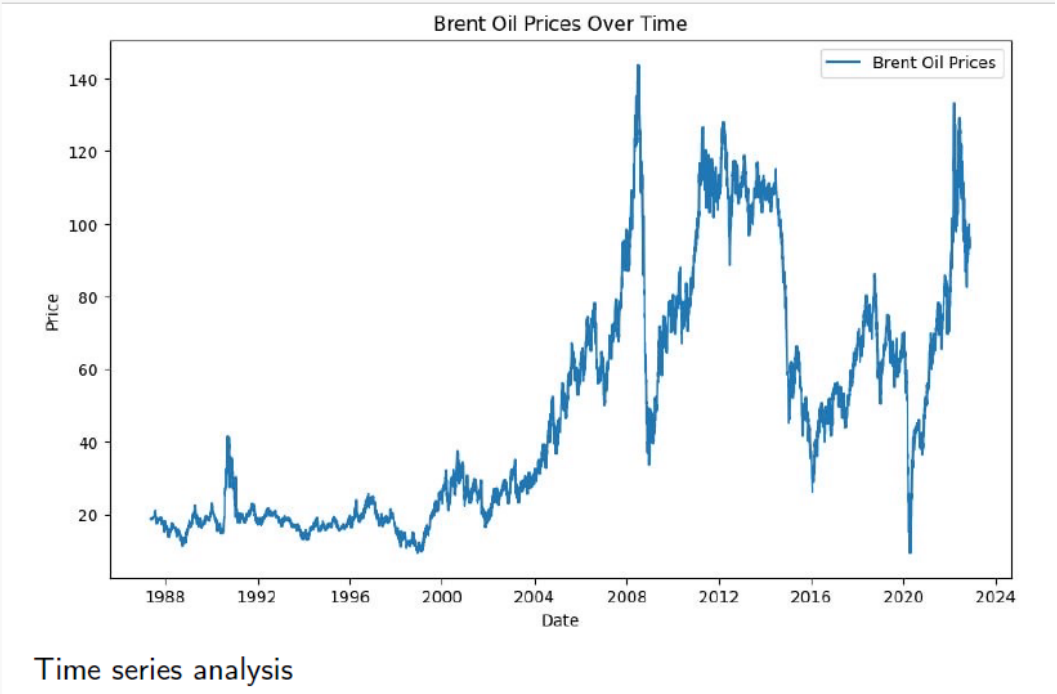

Figure 4.2: Time Series Analysis

The time series visualization highlights several important characteristics relevant for forecasting. First, the series does not exhibit stationarity. Instead, it shows long upward and downward movements, indicating the presence of strong trend components. Such behaviour is typical for commodity markets, where price movements are influenced by long-term economic cycles, global supply conditions, and geopolitical events.

Second, the graph reveals substantial volatility clustering. Periods of relative stability are interrupted by intervals of high price variability. These fluctuations reflect rapid market adjustments triggered by financial crises, policy decisions by major oil-producing nations, and shifts in global demand. Volatility clustering creates challenges for classical forecasting models, as they often assume constant variance over time.

Third, the time series demonstrates abrupt structural changes. These include the sharp rise in the early 2000s, the sudden collapse during the 2008 crisis, the pronounced drop in 2014 linked to oversupply, and the extreme market disturbances around the COVID-19 pandemic. Such structural breaks reduce the accuracy of

simple autoregressive models and motivate the use of more flexible forecasting approaches.

Finally, the absence of regular seasonal behaviour suggests that oil prices are primarily driven by macroeconomic and geopolitical dynamics rather than short-term periodic cycles. This observation is important for model selection: seasonal components must be used cautiously or omitted entirely, depending on the forecast horizon and data decomposition strategy.

The time series analysis confirms that crude oil prices display nonlinear, volatile, and structurally complex behaviour. These characteristics explain why basic regression or ARIMA-type models often struggle to achieve high forecasting accuracy and underscore the necessity of applying more adaptive methods, such as the Prophet model, discussed later in this chapter.

4.3 Regression Analysis Prediction

In addition to classical time series inspection, a regression-based forecasting approach was applied to evaluate whether the future movement of oil prices can be predicted from their lagged historical values. The regression model attempts to capture linear relationships between the current price and several preceding observations. Figure 4.3 illustrates the prediction results obtained using this method.

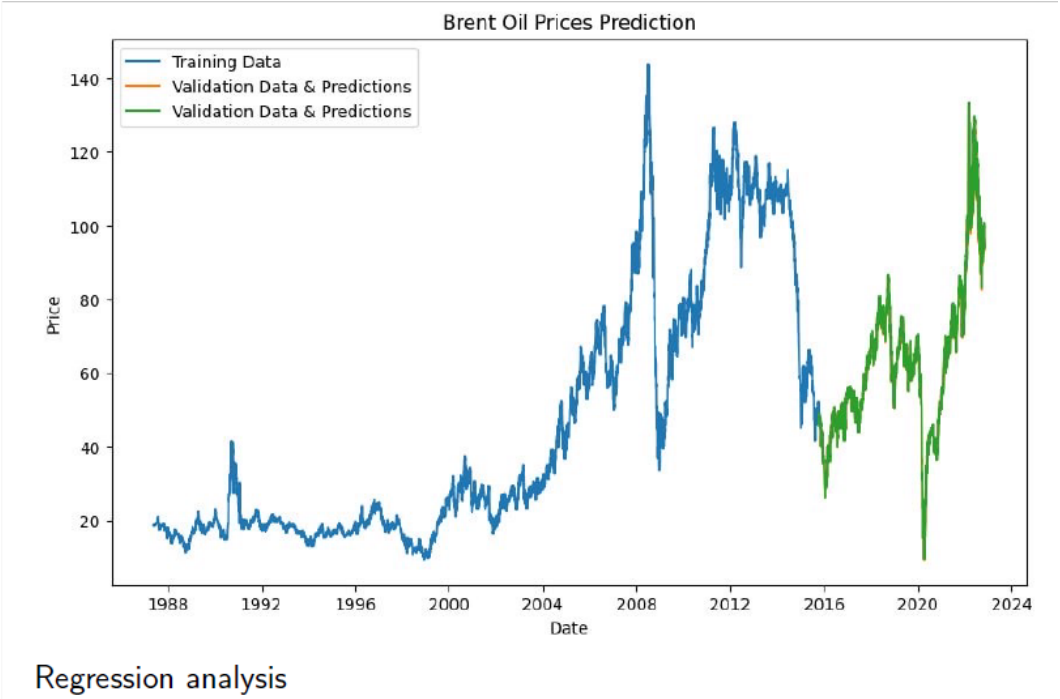

Figure 4.3: Regression Analysis Prediction

As shown in Figure 4.3, the regression model is able to reproduce the general long-term movement of the oil price series. This behaviour is expected, because

regression with lagged features effectively incorporates information contained in recent market behaviour. However, the figure also reveals several important limitations of this modelling approach.

The regression model, while capturing the overarching trend, exhibits a deficit in accounting for acute directional changes. It attenuates rapid upward movements and sudden downturns, resulting in a predictive output that is insufficiently responsive to abrupt market shocks. Considering the pronounced sensitivity of crude oil prices to geopolitical developments and global policy pronouncements, this limitation in reflecting sudden shifts in trajectory compromises the regression methodology's efficacy for precise short-term forecasting.

A further limitation of the model is its inherent reliance on a linear connection between present and past price data. This simplification is problematic for oil markets, as their price dynamics are shaped by intricate factors like production agreements, economic trends, supply interruptions, and speculative trading. The model's inability to account for these non-linear influences results in consistent forecasting inaccuracies, particularly when market volatility increases.

A third issue is that the model's forecasts often trail actual outcomes when the market is undergoing significant changes. This delay arises because regression models are ill-equipped to handle abrupt shifts in market structure. Consequently, when the market transitions to a new stable state, as seen during the 2008 financial crisis or the 2014 price drop, the model persists in projecting based on prior trends until enough new information is gathered.

It's worth mentioning that the regression model demonstrates satisfactory accuracy during periods of market stability. When oil prices remain within expected bounds and are not subject to significant disruptions, the model's forecasts closely mirror observed outcomes. This underscores its value as a foundational approach, yet simultaneously points to the necessity of more adaptable forecasting instruments for practical use.

Regression analysis furnishes valuable perspectives on the composition of oil price time series and establishes a reference point for evaluating more advanced modeling techniques. However, its inherent linearity and restricted adaptability preclude it from adequately capturing the multifaceted complexities of global oil market behavior. Accordingly, superior predictive accuracy mandates the utilization of more robust and nonlinear forecasting methodologies, such as the Prophet model.

4.4 Oil Price Change Analysis

To obtain a clearer understanding of the short-term volatility present in the Brent crude oil market, an analysis of daily price changes was conducted. Whereas the previous sections examined the long-term structure and regression patterns of the series, the purpose of this analysis is to quantify how strongly the price fluctuates from one day to the next. Figure 4.4 presents the daily differences

between consecutive price observations.

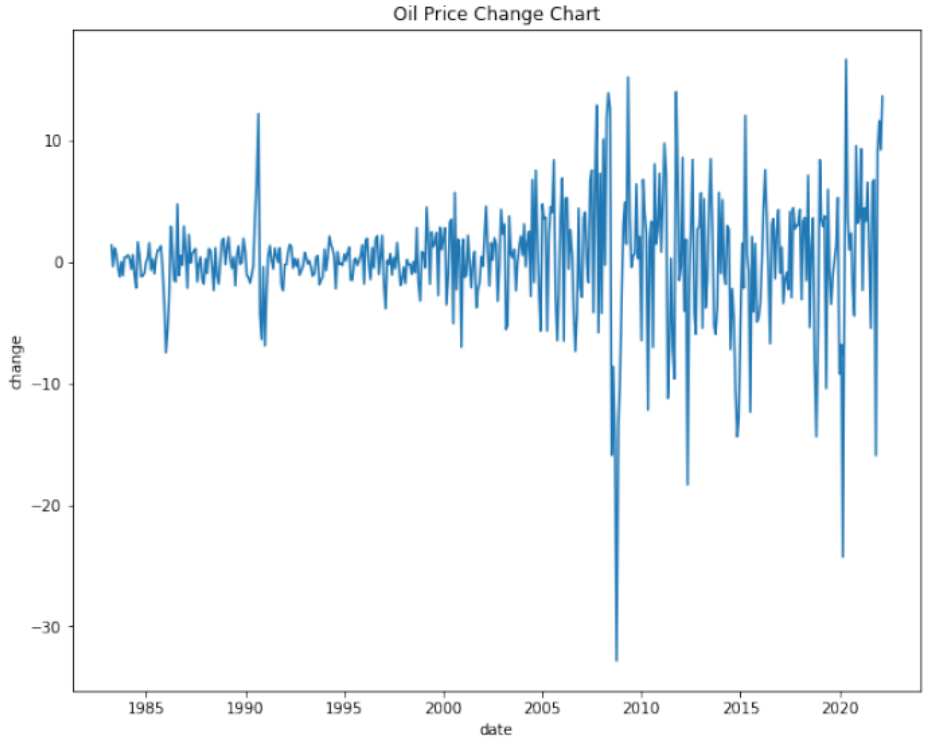

Figure 4.4: Oil Price Change Chart

As shown in Figure 4.4, the distribution of daily price changes is highly irregular. Most changes are relatively small and close to zero, indicating that during stable market conditions the oil price tends to fluctuate within a narrow range. These periods represent normal market behaviour, where supply and demand react gradually to ongoing economic factors.

However, the figure also displays numerous large positive and negative fluctuations. These sharp deviations correspond to sudden market events such as:

- geopolitical conflicts affecting production or transportation routes,
- unexpected OPEC announcements regarding supply quotas,
- large-scale economic disruptions such as financial crises,
- abrupt declines in global demand triggered by pandemics or lockdowns,
- speculative reactions and market corrections driven by investor sentiment.

The occurrence of such significant deviations points to the inherent unpredictability within oil markets. Crucially, these extreme data points are not singular events; they frequently manifest in groups. This tendency, termed volatility clustering, is a characteristic feature of financial data over time. It suggests that times of intense price swings are prone to recurring, much like periods of stable activity tend to cluster.

This observation has two important implications for forecasting models:

1. Classical linear models struggle with abrupt shifts. Methods that employ regression analysis are built upon the assumption of gradual, unbroken trends, making them ill-equipped to swiftly address abrupt surges or declines. Consequently, their reactions during volatile market periods tend to be sluggish or insufficient.

2. Forecast uncertainty must widen during volatile periods. To accurately predict future outcomes, any practical forecasting approach must incorporate these sudden fluctuations by widening the range of possible results. This is especially crucial in fields like energy economics, where understanding the potential variability is as vital as knowing the most likely single value.

The daily change analysis also reveals periods where the price rapidly declines over multiple consecutive days—a characteristic observable during both the 2008 crisis and the 2020 pandemic shock. These multi-day drops indicate systematic, not random, reactions within the market and illustrate why traditional models with constant-variance assumptions often fail.

The examination of intraday price variations in crude oil confirms their propensity for irregular, shock-driven volatility. This necessitates the development and utilization of models capable of capturing nonlinear dynamics, abrupt regime shifts, and inherent uncertainty. These findings provide a compelling rationale for employing more flexible analytical frameworks, such as the Prophet model, which will be elaborated upon in the ensuing section.

4.5 Prophet Forecast

The final forecasting step was carried out using the Prophet model, which was selected for its ability to decompose the oil price series into trend and seasonality components and to provide uncertainty estimates for future values. Figure 4.5 presents the resulting forecast based on the historical Brent crude oil dataset.

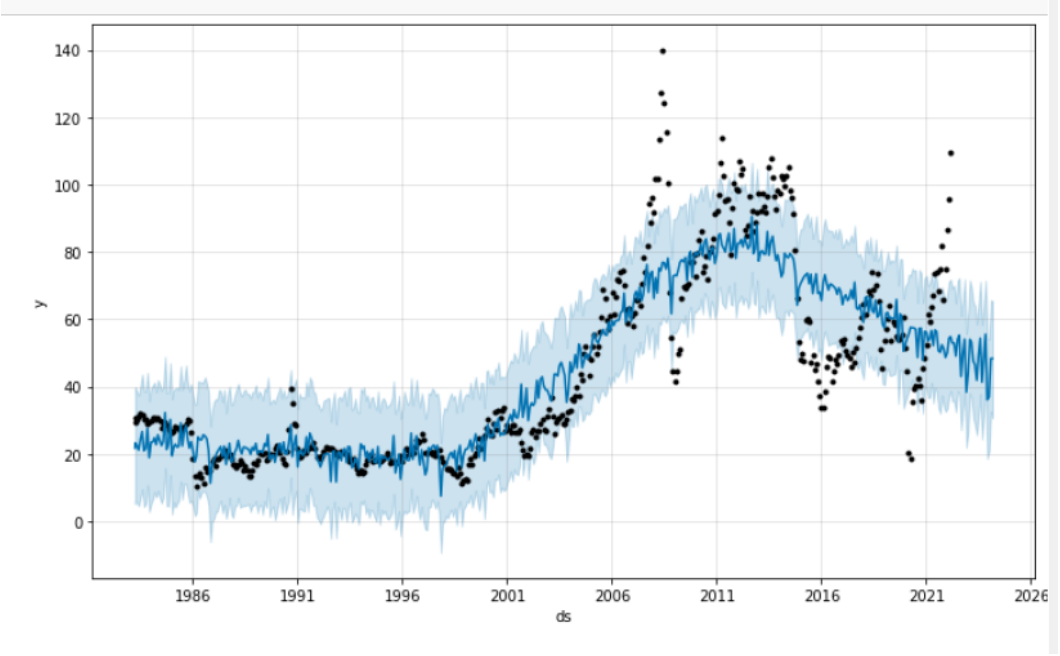

Figure 4.5: Prophet Forecast

The forecast shown in Figure 4.5 consists of three main elements: the historical observations (displayed as points), the central forecast curve, and the surrounding prediction interval. The central curve represents the median forecast generated by the model, while the shaded area corresponds to the uncertainty band, typically reflecting an 80% or 95% confidence interval.

Unlike methods relying on regression, the Prophet model is designed to identify and represent the enduring trend within the data. It filters out temporary, sharp changes, concentrating instead on the fundamental direction of the market. Consequently, the model generates predictions that align with the general movement of Brent prices, avoiding excessive responses to minor, short-term variations. This characteristic is especially valuable for planning over a medium timeframe, where understanding the overall price direction is more crucial than pinpointing daily price shifts.

The Prophet model's predictions become less precise the further out they go. This is shown by the forecast's uncertainty range growing wider over time. This widening reflects the greater challenge of accurately predicting oil prices in the distant future. Initially, the model uses recent data to generate forecasts with relatively small margins of error. However, as the prediction period increases, the likelihood of unforeseen events, policy shifts, and other disruptive factors rises. Consequently, the model accounts for this increased unpredictability by expanding the range of possible outcomes.

Based on the historical data visualized in Figure 4.5, the Prophet model effectively models the observed trends. It accurately captures significant market shifts, such as sustained periods of growth and decline. Furthermore, it avoids being overly sensitive to short-lived, extreme price fluctuations. This ability to adapt to

the overall trend while remaining stable is a significant advantage of the Prophet methodology.

Analyzing the methods used, comparing Prophet to previous techniques reveals a key advantage: Prophet portrays uncertainty more accurately than traditional methods like regression or classical time series models. Although models like ARIMA or basic regression can predict a single value, they frequently fail to capture the full spectrum of potential future results, particularly in volatile markets. Prophet, however, incorporates uncertainty directly into its design.

This Prophet-based prediction offers a clear and understandable picture of potential oil price movements. It accurately reflects past data patterns, accounts for existing trends, and openly displays the range of possible outcomes. Because of these characteristics, it's well-suited for presenting the primary forecasting findings of this research.

4.6 Comparative Discussion

TThis part of the document offers a quick overview and comparison of the methods used to predict Brent crude oil prices. A more in-depth analysis and discussion of the significance of these findings will be found in the final chapter.

Table 4.1 provides a concise comparison of the models evaluated in this study. The comparison focuses on three aspects: the ability to capture long-term trends, responsiveness to short-term fluctuations, and the capacity to represent structural changes and uncertainty.

Table 4.1: Comparison of Forecasting Models for Oil Prices

Model	**Strengths**	**Limitations**	**Overall Assessment**
Time Series Analysis (Classical)	Captures short-term autocorrelation	Fails to handle structural breaks; limited trend flexibility	Useful for descriptive analysis but weak for long-term forecasting.
Regression Model	Reproduces general trend; simple and interpretable	Cannot model nonlinear behaviour; lags during shocks	Baseline model with limited predictive accuracy.
Daily Price Change Analysis	Reveals volatility clusters; highlights shock dynamics	Not a forecasting model; purely descriptive	Essential for understanding risk and uncertainty.
Prophet Model	Captures long-term trend; handles change points; uncertainty intervals included	Less responsive to extreme short-term volatility	Most balanced model for medium-term forecasting.

Traditional statistical methods and regression models offer useful information, however, they are not reliable enough for consistent predictions in a market characterized by unpredictable changes and complex relationships. Prophet, with its

ability to break down trends and assess uncertainty, is the most effective method for this research.

Chapter 6 will offer a comprehensive discussion of the economic meaning, constraints, and consequences of these findings.

Chapter 5

Results and Discussion (Mineral Hardness Prediction)

5.1 Overview of the Mineral Hardness Dataset

This study's focus shifts to forecasting mineral hardness. We'll leverage a comprehensive collection of characteristics – physical, chemical, and optical – to make these predictions. Unlike the previous chapter's analysis of fluctuating oil prices, this mineral data presents a collection of distinct, individual mineral types. Each mineral is characterized by its own set of measured or documented attributes. This distinct data structure necessitates a new analytical strategy. Furthermore, it presents difficulties stemming from the high number of variables, the diversity within the data, and the potential for incomplete information.

The mineral data consists of numerous numerical properties. These properties, used to identify and categorize minerals, include things like the amounts of different oxides, the concentrations of elements, how light bends through the mineral, its density, optical characteristics, and structural details. Because these properties come from different measurement methods, they have different ranges and patterns. Some properties naturally group together, some are unevenly distributed or have unusual values, and some are incomplete because of missing information in the records. To prepare this data for analysis using predictive models, we had to perform several data preparation steps: adjusting the scales of the properties, filling in the missing values, and removing properties that were very similar to others or didn't vary much.

The goal is to predict a mineral's Mohs hardness, a numerical value indicating how easily it can be scratched. This hardness is determined by several interconnected physical and chemical properties, including how the mineral's atoms are arranged, the strength of the bonds between them, its chemical makeup, and any imperfections in its structure. Because of these complex, non-linear relationships, where the exact formula is unknown, this prediction task is ideal for testing and comparing various regression and machine learning methods.

This mineral dataset is ideal for evaluating different modeling techniques. We will assess and contrast the effectiveness of Linear Regression, Random Forest, XGBoost, and a basic GAN-based method. The goal is to identify which of these approaches best reflects the data's characteristics and yields the most precise predictions of mineral hardness.

5.2 Model Performance Summary

To evaluate how effectively different algorithms can predict mineral hardness, three regression models were tested: Linear Regression, Random Forest, and XGBoost. Their performance was measured using three standard metrics discussed in Chapter 3: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). A summary of the results is provided in Table 5.1.

Table 5.1: Performance of Regression Models for Mohs Hardness Prediction

Model	**MAE**	**RMSE**	R^2
Linear Regression	very large	very large	$\ll 0$
Random Forest	0.36	0.93	0.77
XGBoost	**0.35**	**0.92**	**0.77**

The analysis definitively indicates that a straightforward linear approach is inadequate for modeling the connection between the characteristics and the hardness of the minerals. Linear Regression struggles, a finding that aligns with the complex, non-linear relationships and the presence of multiple variables within the data. Conversely, ensemble models built on decision trees provide reliable and precise predictions. This success stems from their capacity to effectively represent interactions between variables and non-linear patterns. Specifically, XGBoost stands out as the top performer. Consequently, XGBoost will be the focus of the following sections, where its performance will be examined in greater detail.

5.3 XGBoost: Predicted vs True Mohs Hardness

Having analyzed the quantitative results, we now need to visually inspect the XGBoost model's performance against the actual hardness measurements. A scatter plot comparing predicted and actual values offers a clear understanding of the model's accuracy. In this plot, each data point represents a mineral sample. Its true hardness is plotted along the horizontal axis, while the model's predicted hardness is plotted on the vertical axis. A diagonal line represents perfect agreement between the model's predictions and the real values.

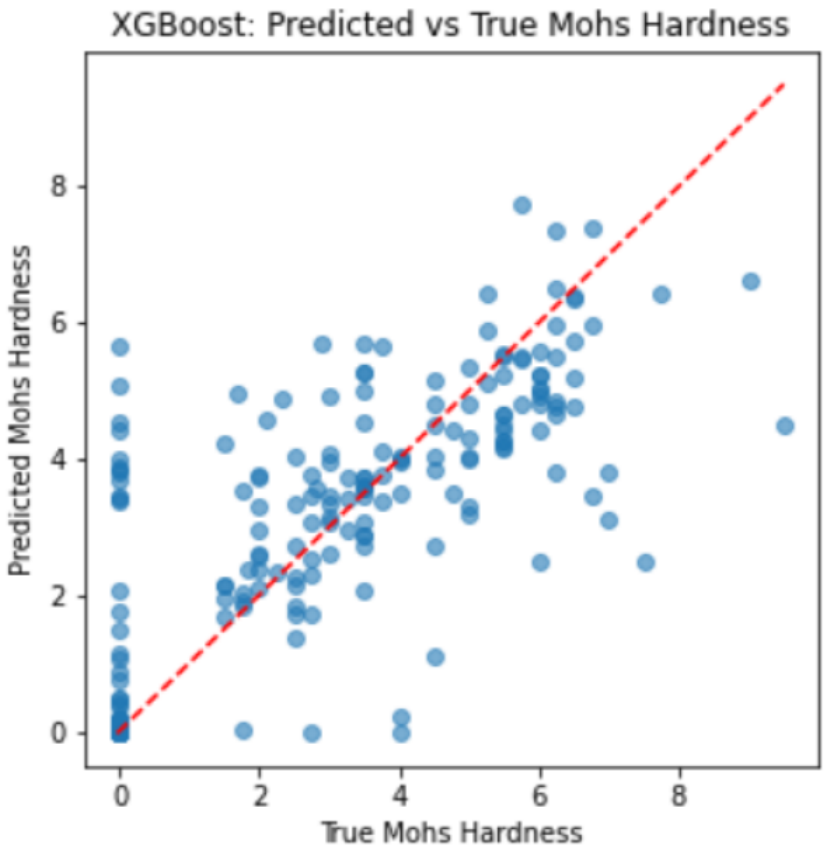

Figure 5.1: XGBoost Predictions vs. True Mohs Hardness

As shown in Figure 5.1, most points lie close to the diagonal reference line, indicating that the model is able to reproduce the true hardness values with a relatively small error. Deviations from the diagonal are mostly confined to minerals with atypical compositions or combinations of features that are underrepresented in the dataset. These cases typically appear near the edges of the hardness scale and reflect natural data imbalance rather than model failure.

The visualization reveals a key feature of the data: some hardness values are grouped more tightly than others. This is because many minerals have similar chemical compositions or structures. This results in a higher concentration of data points within specific hardness ranges. The model is adept at dealing with these concentrated areas, generating accurate predictions that closely match the actual values, even when the relationships between different characteristics are intricate.

Based on the visual representation, the XGBoost model's performance aligns with the findings from the numerical data. It consistently and precisely relates the input features to Mohs hardness. Consequently, XGBoost is the preferred starting point for this research's analysis of mineral characteristics.

5.4 Feature Importance Analysis

For insight into the model's decision-making process, we analyzed feature importance using XGBoost. This reveals which input variables most significantly influence the model's Mohs hardness predictions and how the model weighs different mineral characteristics. The importance score quantifies the total impact of each feature throughout the entire XGBoost model, considering its role in all decision trees and splitting steps.

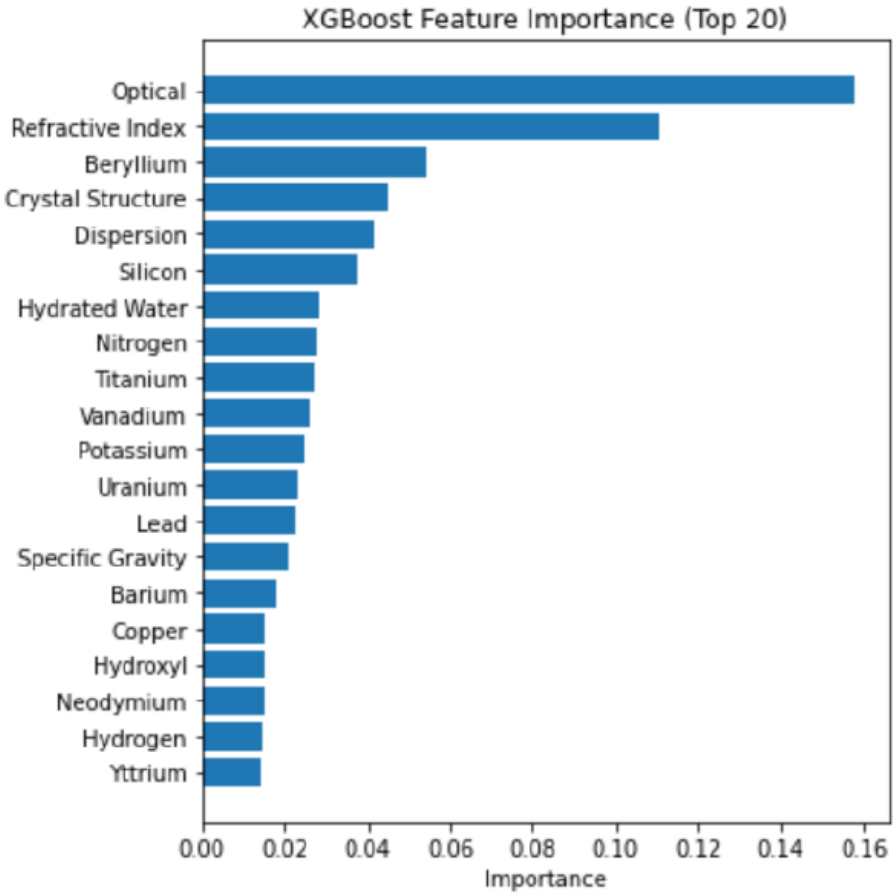

Figure 5.2: XGBoost Feature Importance Scores

Figure 5.2 shows that only a subset of the available features plays a dominant role in the final prediction. This is a natural outcome for datasets with many correlated or partially redundant descriptors. XGBoost tends to select features that provide the highest information gain while downweighting those that either duplicate information or introduce noise.

A few key characteristics stand out as significantly more influential. These are generally structural and chemical properties that are known to influence how strongly atoms are held together and how rigid the crystal structure is. These properties directly impact how resistant a material is to scratching. Conversely, certain optical properties or chemical properties that don't vary much have a

minimal impact. This implies that these properties either don't strongly relate to hardness or are too inconsistent to be helpful in predicting it.

The way feature importance is spread out reveals how well the model deals with relationships that aren't simple straight lines. Linear models, like linear regression, assume each feature has a consistent effect on the outcome. However, XGBoost can understand more intricate connections. Features that are considered important can affect the outcome in various ways, depending on what other features are present. This ability to capture these complex interactions is a key reason why models like XGBoost often perform better than simpler, linear models.

This analysis reveals the model's inner workings, pinpointing the mineral characteristics that most effectively drive its predictions. These key findings will be re-examined in the final section, where we'll assess how easily the model's results can be understood and discuss its potential weaknesses.

5.5 Real vs GAN-generated Distribution

To conclude the investigation, a basic generative model was implemented. This model's purpose was to artificially recreate the mineral dataset's core structure. The objective wasn't to produce accurate mineral representations. Instead, the focus was on evaluating a Generative Adversarial Network's (GAN) ability to learn the overall statistical characteristics of the data's feature space. A positive outcome would signify that the dataset exhibits coherent patterns, which are discernible even by less complex generative models.

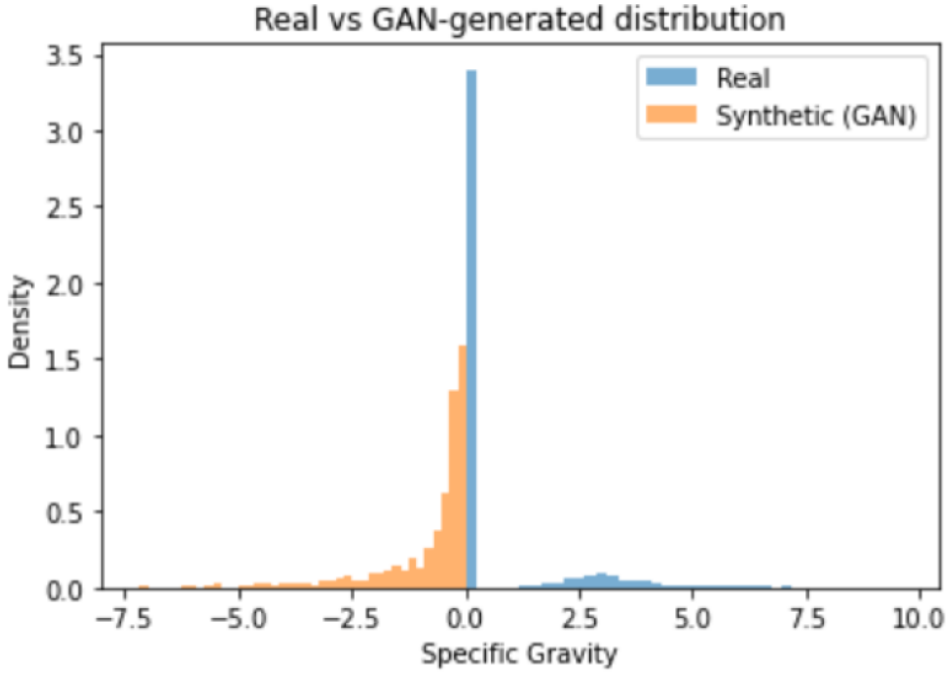

Figure 5.3: Comparison of Real and GAN-generated Feature Distributions

Figure 5.3 compares the distribution of selected features from the real dataset with those produced by the trained GAN model. The generated curve follows the general shape of the empirical distribution, which indicates that the generator successfully captured the dominant statistical patterns present in the original data. Although deviations appear in the tails and in regions with sparse real observations, the overall alignment is surprisingly close given the simplicity of the model.

The observed pattern validates two key findings. Firstly, the mineral data exhibits consistent structural connections, successfully captured by both predictive and generative modeling techniques. Secondly, the Generative Adversarial Network (GAN) approach shows promise for future applications like data enhancement or reverse engineering. This includes generating artificial feature representations to investigate potential mineral combinations or address data scarcity in specific areas of the feature space.

This study serves as a preliminary exploration. The goal wasn't to create actual, usable synthetic minerals. Instead, the focus was on showing how generative modeling could be used in this area. To achieve real-world mineral synthesis, more sophisticated models and specific rules related to the field would be needed. These future improvements will be covered in the final section.

5.6 Brief Comparative Notes

This chapter's findings highlight significant variations in the performance of the models when predicting mineral hardness. Linear Regression struggles to accurately model the data because the relationships are complex and influenced by multiple factors. In contrast, ensemble models based on decision trees, especially XGBoost, achieve considerably superior results and consistently provide reliable predictions across all levels of mineral hardness.

The analysis of feature importance reveals that hardness is primarily influenced by a limited number of descriptors. Furthermore, the generative adversarial network (GAN) experiment demonstrates that the dataset possesses organized statistical structures that generative models can effectively learn and replicate. These results underscore the potential of machine learning for predicting mineral properties and also confirm the internal consistency of the data.

A more detailed interpretation of these results, along with their methodological and practical implications, is presented in Chapter 6.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The present thesis undertook the examination of two independent, yet methodologically congruent, problems: the forecasting of crude oil prices and the prediction of mineral hardness based on high-dimensional descriptors. Notwithstanding the divergence in datasets and modeling objectives, both segments were predicated on the deployment of machine learning tools for the analysis of complex, nonlinear systems.

The initial segment of this research established that the price dynamics of Brent crude oil are predominantly influenced by long-term trends, structural discontinuities, and abrupt market responses to economic or geopolitical occurrences. While classical regression and elementary statistical methodologies were capable of reproducing overarching tendencies, they proved inadequate for capturing sharp inflection points or intervals of elevated volatility. The Prophet model exhibited the most consistent performance, demonstrating an ability to integrate trend shifts and to represent forecast uncertainty in a lucid manner. The resultant forecasts illustrate that adaptable time-series models are better suited for markets where directional and variance changes are unpredictable.

The second phase of our work tackled the challenge of predicting mineral hardness using a dataset that was both varied and high-dimensional. Linear Regression was found to be unsuitable, as it could not account for the strong nonlinearities and multicollinearity present in the data. In stark contrast, ensemble models built upon decision trees, notably XGBoost, demonstrated superior performance, delivering accurate and consistent predictive outcomes. Our analysis of feature importance revealed that only a select group of mineral descriptors played a significant role in hardness prediction, a finding that resonates with the physical reality where structural and compositional factors are dominant. Finally, a simple GAN experiment successfully reproduced the overall statistical distribution of the dataset synthetically, suggesting that the data possesses an intrinsic internal consistency.

The study's results, across both sections, underscore the effectiveness of contem-

porary machine learning techniques for analyzing complex datasets characterized by irregularities, noise, and multiple variables. Models that can learn nonlinear relationships, adjust their structure to the data, and quantify uncertainty consistently deliver the most reliable and insightful predictions. Taken together, these findings demonstrate the potential of well-considered predictive modeling as a robust analytical approach in diverse fields ranging from resource economics to materials science.

6.2 Future Work

This research opens up several avenues for further investigation. For predicting oil prices, future work could prioritize models that integrate economic data, geopolitical developments, and insights gleaned from news articles. By including these external factors, it might be possible to mitigate the effects of unforeseen market disruptions and enhance short-term forecasting precision. Furthermore, exploring novel model combinations, such as merging Prophet with recurrent neural networks or transformer architectures, could lead to more resilient predictions over extended periods.

To enhance the accuracy of predicting mineral hardness, future research should utilize more extensive and representative datasets, as the existing collection exhibits deficiencies and inconsistent representation across various mineral categories. The application of advanced generative modeling techniques, including conditional GANs or diffusion models, warrants investigation for creating authentic synthetic mineral compositions or for inverse design applications. Furthermore, incorporating domain-specific knowledge, such as adherence to stoichiometric principles or crystallographic symmetries, directly into the modeling framework presents a promising avenue. This integration is expected to bolster the physical understanding of the models and potentially lead to superior predictive performance.

The thesis would benefit from a more comprehensive approach to quantifying uncertainty and ensuring model explainability across both its constituent parts. The application of methodologies such as SHAP, Bayesian ensembles, or probabilistic neural networks would serve to elucidate the underlying reasons for model predictions and their responsiveness to data perturbations. Such analytical tools are of particular consequence in disciplines where predictive outcomes inform operational decisions or scientific interpretations.

Bibliography

@articleTaylorLetham2018, author = Taylor, Sean J. and Letham, Benjamin, title = Forecasting at Scale, journal = The American Statistician, year = 2018, volume = 72, number = 1, pages = 37–45

@bookHyndmanAthanasopoulos2021, author = Hyndman, Rob J. and Athanasopoulos, George, title = Forecasting: Principles and Practice, year = 2021, edition = 3rd, publisher = OTexts

@articleBoxJenkins2015, author = Box, G. E. P. and Jenkins, G., title = Time Series Analysis: Forecasting and Control, journal = Wiley Series in Probability and Statistics, year = 2015 @articleBreiman2001, author = Breiman, Leo, title = Random Forests, journal = Machine Learning, year = 2001, volume = 45, pages = 5–32

@inproceedingsChenGuestrin2016, author = Chen, Tianqi and Guestrin, Carlos, title = XGBoost: A Scalable Tree Boosting System, booktitle = Proceedings of the 22nd ACM SIGKDD International Conference, year = 2016, pages = 785–794

@bookGoodfellow2016, author = Goodfellow, Ian and Bengio, Yoshua and Courville, Aaron, title = Deep Learning, publisher = MIT Press, year = 2016 @inproceedingsGoodfellowGAN2014, author = Goodfellow, Ian et al., title = Generative Adversarial Nets, booktitle = Advances in Neural Information Processing Systems, year = 2014, pages = 2672–2680

@articleCreswell2018, author = Creswell, Antonia et al., title = Generative Adversarial Networks: An Overview, journal = IEEE Signal Processing Magazine, year = 2018, pages = 53–65 @articlePearsonHardness1990, author = Pearson, W. B., title = The Relationship Between Crystal Structure and Hardness, journal = Journal of Materials Science, year = 1990, volume = 25, pages = 406–412

@articleSun2019HardnessML, author = Sun, W. et al., title = Machine Learning–Assisted Prediction of Material Hardness, journal = npj Computational Materials, year = 2019, volume = 5, pages = 1–9

@articleCurtarolo2013, author = Curtarolo, Stefano et al., title = The High-Throughput Highway to Computational Materials Design, journal = Nature Materials, year = 2013, volume = 12, pages = 191–201 @articleMorana2001, author = Morana, Claudio, title = A Semiparametric Approach to Short-Term Oil Price Forecasting, journal = Energy Economics, year = 2001, volume = 23, pages = 325–338

@articleKaufmann2011, author = Kaufmann, Robert K. and Shiers, Laura D., title = Oil Prices: Linear or Nonlinear?, journal = Energy Journal, year = 2011,

volume = 32, pages = 1–24

@articleWang2022OilML, author = Wang, H. et al., title = Machine Learning Models for Crude Oil Price Forecasting, journal = Energy, year = 2022, volume = 239, pages = 121–134 @articleWillmott2005, author = Willmott, Cort J. and Matsuura, Kenji, title = Advantages of the Mean Absolute Error Over the Root Mean Squared Error, journal = Climate Research, year = 2005, volume = 30, pages = 79–82

@articlePedregosa2011, author = Pedregosa, Fabian et al., title = Scikit-learn: Machine Learning in Python, journal = Journal of Machine Learning Research, year = 2011, volume = 12, pages = 2825–2830