FOSSEE Semester Long Internship Report

On

Crude Oil Analysis Using ANN Model

Submitted by

Tanvi Sudhir Lalsare

3rd Year B. Tech Student, CSE (AI ML)

VIT Bhopal University

Under the Guidance of

Priyam Nayak

Research Scientist, FOSSEE project

February 24, 2026

Acknowledgments

I would like to express my sincere gratitude to the FOSSEE team at IIT Bombay for providing me with the opportunity to undertake this semester-long internship. This experience has been both intellectually enriching and professionally rewarding, allowing me to work on a meaningful project in the field of machine learning and data analysis.

During the course of this internship, I worked on developing predictive models for crude oil composition using various machine learning and chemometric techniques. This involved experimenting with multiple models, implementing proper validation strategies, and understanding the practical limitations of data-driven approaches. The experience greatly enhanced my problem-solving skills and deepened my understanding of real-world applications of machine learning.

I am especially grateful to **Priyam Nayak, Research Scientist at the FOSSEE project**, for his continuous guidance, support, and insightful suggestions throughout the internship. His mentorship played a crucial role in shaping my approach towards research and model development.

This internship has been a significant milestone in my academic and professional development, providing me with valuable experience and clarity for my future career path.

CONTENTS

1. Introduction

1.1 Background 5
1.2 Problem Statement 5-6
1.3 Objectives 6
1.4 Scope of Work 6-7
1.5 Organization of the Report 7-8

2. Literature Review

2.1 Overview of Crude Oil Characterization 9
2.2 Machine Learning in Petroleum Analysis. 9
2.3 Artificial Neural Network Approaches 9-10
2.4 Chemometric Methods (PLS Regression) 10
2.5 Research Gap 10-11

3. Dataset and Preprocessing

3.1 Dataset Description 12
3.2 Data Extraction from Crude Assay Files 13
3.3 Data Cleaning and Validation 14
3.4 Feature Selection 15
3.5 Data Normalization and Scaling 16

4. Model Development and Methodology

4.1 Overview of Modelling Approach 17
4.2 Model 1: ANN vs Random Forest(Baseline Model) 17-18
4.3 Model 2: Independent Random Forest (3-Target) 18-19
4.4 Model 3: ILR-Based Model(Compositional Modelling) 19-20
4.5 Model 4: Feature Engineered Random Forest 20
4.6 Model 5: Single-Target Paraffins Model 21

4.7 Model 6: Partial Least Squares (PLS) Regression 21-22
4.8 Model 7: Hybrid PLS-Based Model 22-23

5. **Model Evaluation and Validation**

5.1 Evaluation Metrics (R^2, MAPE) 24
5.2 K-Fold Cross Validation Strategy 24-25
5.3 Training vs Cross-Validation Analysis 26-27
5.4 Error Analysis 27-29

6. **Results and Discussion**

6.1 Comparative Analysis of Models 30-31
6.2 Performance of ANN vs PLS 31-32
6.3 Impact of Dataset Size 32-33
6.4 Observations and Insights 33-36

6.41 Effect of Multicollinearity 33
6.42 Compositional Constraint Considerations 33-34
6.43 Error Behavior Across Fractions 34
6.44 Training vs Validation Discrepancy 35
6.45 Hybrid Model Insights 35-36
6.46 Practical Implications 36

7. **Conclusion**

7.1 Summary of Work 37
7.2 Key Findings 37-38
7.3 Final Conclusion 38
7.4 Contribution of the Study 39

8. **Future Work** 40

9. **References** 41-42

CHAPTER 1

INTRODUCTION

1.1 Background

Crude oil is a complex mixture of hydrocarbons and other chemical compounds, and its composition varies significantly depending on the source. Understanding the chemical composition of crude oil is essential for refining processes, product optimization, and economic decision-making in the petroleum industry.

Among the various characterization techniques, the estimation of hydrocarbon fractions such as Aromatics, Naphthenes, and Paraffins is particularly important. These components influence key refining parameters such as cracking behavior, viscosity, combustion quality, and yield distribution.

Traditionally, determining these fractions requires laboratory-based techniques such as chromatography and distillation analysis, which are time-intensive, costly, and not always feasible for large-scale or real-time applications. With the increasing availability of crude assay data, there is a growing opportunity to apply machine learning and statistical modeling techniques to estimate these fractions using easily measurable input properties.

This study explores the use of machine learning and chemometric approaches to predict crude oil composition, focusing on the balance between model complexity and dataset limitations.

1.2 Problem Statement

The primary challenge addressed in this work is the prediction of Aromatics, Naphthenes, and Paraffins percentages using a limited dataset of crude oil properties.

Several key difficulties arise in this problem:

- The dataset is relatively small (approximately 50–65 samples), which limits the ability of complex models to generalize
- Input features such as density, sulfur content, and TBP distillation temperatures are highly correlated
- The outputs are compositional in nature and must satisfy a physical constraint (sum of fractions must equal 100%)
- Different fractions exhibit different levels of predictability, with paraffins being particularly difficult to model

These challenges require careful model selection, robust validation techniques, and consideration of physical constraints.

1.3 Objectives

The main objectives of this study are:

- To develop predictive models for estimating Aromatics, Naphthenes, and Paraffins from crude oil assay data
- To compare multiple modeling approaches, including Artificial Neural Networks (ANN), Random Forest, ILR-based methods, and Partial Least Squares (PLS) regression
- To ensure physically consistent predictions by enforcing compositional constraints
- To evaluate model performance using reliable metrics such as R^2 and MAPE
- To apply repeated K-fold cross-validation for unbiased performance estimation
- To identify the most suitable model for small-scale petroleum datasets

1.4 Scope of Work

This study focuses on:

- A dataset of crude oil assay samples containing approximately 13 input features
- Prediction of three output variables: Aromatics, Naphthenes, and Paraffins
- Implementation and comparison of multiple machine learning and chemometric models
- Use of statistical validation techniques to assess model reliability

The study does not include large-scale industrial deployment or real-time implementation but focuses on methodological evaluation and model reliability.

1.5 Organization of the Report

This report is structured systematically to guide the reader through the complete research workflow, from problem formulation to final conclusions.

Chapter 1 introduces the background, motivation, objectives, and scope of the study, providing a clear understanding of the problem being addressed.

Chapter 2 presents a detailed literature review, covering existing work in crude oil characterization, machine learning applications in petroleum engineering, and the use of Artificial Neural Networks and chemometric methods such as Partial Least Squares regression. It also identifies gaps in existing research that this study aims to address.

Chapter 3 describes the dataset used in the study and the preprocessing steps applied. This includes data extraction from crude assay files, handling missing values, feature selection, and normalization techniques.

Chapter 4 focuses on model development and methodology. It provides a detailed explanation of each model implemented, including ANN, Random Forest, ILR-based models, feature-engineered models, and PLS regression. The progression of models is presented to highlight improvements and limitations at each stage.

Chapter 5 discusses the evaluation framework used to assess model performance. It explains the metrics used (R^2 and MAPE), the implementation of repeated K-fold cross-validation, and the importance of distinguishing between training and validation performance.

Chapter 6 presents the results and discussion, comparing the performance of different models and analyzing their strengths and limitations. It also highlights key insights derived from the study.

Chapter 7 concludes the report by summarizing the findings and emphasizing the importance of model selection based on dataset characteristics.

Chapter 8 outlines potential future work, including improvements in dataset size, model architecture, and hybrid modeling approaches.

Finally, **Chapter 9** lists the references used in this study.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Crude Oil Characterization

Crude oil characterization is a fundamental aspect of petroleum engineering, involving the determination of chemical composition and physical properties that influence refining processes. The classification of crude oil into hydrocarbon fractions such as Aromatics, Naphthenes, and Paraffins is critical for understanding its behavior during processing and its economic value [1], [2].

Traditional analytical methods such as gas chromatography and true boiling point (TBP) distillation are widely used for determining these fractions. However, these methods are time-consuming, expensive, and require specialized laboratory infrastructure [3]. As a result, there has been increasing interest in predictive approaches that utilize readily available assay data to estimate crude composition.

2.2 Machine Learning in Petroleum Analysis

Machine learning techniques have been increasingly applied in petroleum engineering for property prediction, reservoir modeling, and process optimization [4], [5]. Models such as linear regression, decision trees, support vector machines, and ensemble methods have been used to predict crude oil properties based on input features like density, viscosity, and distillation characteristics [6], [7].

Despite their potential, the performance of these models is highly dependent on dataset size and quality. Small datasets often lead to overfitting, where models fail to generalize to new data [8]. Furthermore, many studies report performance based only on training or simple train-test splits, which may not accurately reflect real-world performance [9].

2.3 Artificial Neural Network Approaches

Artificial Neural Networks (ANNs) are widely used for modeling nonlinear relationships in complex systems. In the context of crude oil characterization, ANN models have been applied to predict SARA fractions and other properties [10], [11].

Alizadeh (2023) demonstrated the application of ANN for crude oil property prediction using a limited dataset, achieving moderate accuracy [12]. Similarly, other studies have shown that ANN models can capture nonlinear patterns effectively but are prone to overfitting when the dataset size is small [13].

A key limitation in many ANN-based studies is the lack of rigorous validation techniques such as cross-validation, leading to overestimation of model performance [14].

2.4 Chemometric Methods (PLS Regression)

Partial Least Squares (PLS) regression is a chemometric technique widely used for analyzing datasets with high multicollinearity and limited sample size [15], [16]. PLS projects input variables into a lower-dimensional space of latent variables, maximizing the covariance between inputs and outputs.

PLS has been successfully applied in chemical engineering and petroleum applications for property prediction, spectral analysis, and process modeling [17], [18]. Compared to ANN, PLS models are more stable and interpretable, particularly in small datasets where overfitting is a concern [19].

In recent studies, PLS has been shown to outperform complex machine learning models when the dataset is limited and highly correlated [20].

2.5 Research Gap

Despite advancements in machine learning applications for crude oil characterization, several limitations remain:

- Many studies rely on **small datasets without proper validation techniques** [9]
- Performance is often reported using **training data only**, leading to overestimated results [14]

- Limited consideration is given to **compositional constraints** (sum of fractions = 100%)
- Complex models such as ANN are used without addressing **data limitations and multicollinearity**

This study addresses these gaps by:

- Applying **repeated K-fold cross-validation** for robust evaluation
- Comparing multiple models under consistent conditions
- Incorporating **chemometric approaches (PLS)**
- Demonstrating the importance of **model simplicity in small datasets**

CHAPTER 3

DATASET AND PREPROCESSING

3.1 Dataset Description

The dataset used in this study consists of crude oil assay data collected from multiple sources, including structured datasets and raw assay files. A primary dataset, referred to as *ML_Set1_all_crudes.xlsx*, contains information for 53 crude oil samples along with their corresponding physicochemical properties and hydrocarbon composition.

In addition to this, a collection of raw assay files in compressed format (ZIP) was utilized. These files contain detailed crude-specific measurements such as density, sulfur content, nitrogen content, Conradson carbon residue (CCR), and distillation characteristics at various temperature cuts.

The target variables for this study are the three major hydrocarbon fractions:

- Aromatics (AromByWt %)
- Naphthenes (NaphthenesByWt %)
- Paraffins (ParaffinsByWt %)

These components are compositional in nature and satisfy the constraint:

Aromatics + Naphthenes + Paraffins = 100%

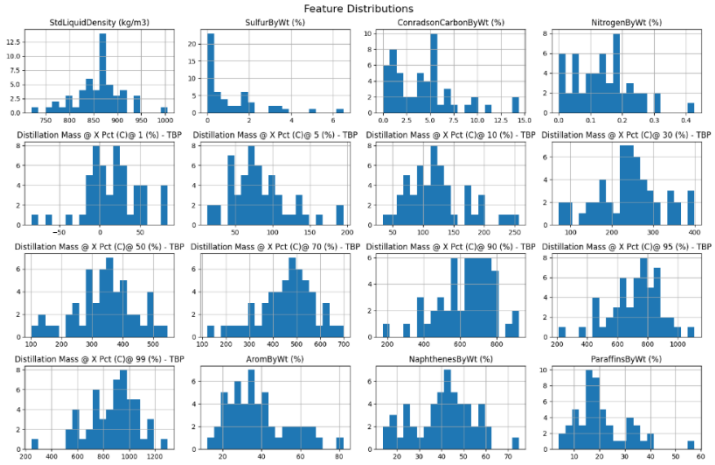

Figure 1 Distribution of Input Features

This compositional constraint plays a crucial role in model design and evaluation.

3.2 Data Extraction from Crude Assay Files

The raw assay data was provided in the form of multiple Excel files compressed into ZIP archives. Each file represents a single crude oil sample and contains multiple sheets corresponding to different physical and chemical measurements.

To prepare the dataset for modeling, the following steps were performed:

1. Extraction of ZIP Files
   The compressed assay files were extracted using Python's zipfile library.
2. Iterative File Processing
   Each Excel file was read using the pandas library, and relevant sheets were identified.
3. Feature Mapping
   Key features such as:
   - Standard Liquid Density
   - Sulfur Content
   - Nitrogen Content
   - Conradson Carbon Residue (CCR)
   - Distillation Temperature (T50)

were extracted and standardized across all files.

4. Dataset Consolidation
   The extracted data from all crude files was merged into a single structured dataset.
5. Validation Against ML Dataset
   The extracted dataset was cross-checked with the ML_Set dataset to ensure consistency in crude names and feature alignment.

The final extracted dataset consisted of 53 crude samples with 17 features, matching the reference dataset.

3.3 Data Cleaning and Validation

Data cleaning was a critical step due to inconsistencies in raw assay files. The following preprocessing steps were applied:

- **Handling Missing Values**
  Missing values were identified and replaced using appropriate strategies such as:
  - Zero imputation for negligible quantities
  - Mean substitution for continuous variables
- **Outlier Detection**
  Extreme values were examined using statistical summaries and were clipped to physically meaningful ranges.
- **Consistency Checks**
  Ensured that all features were in consistent units and formats.

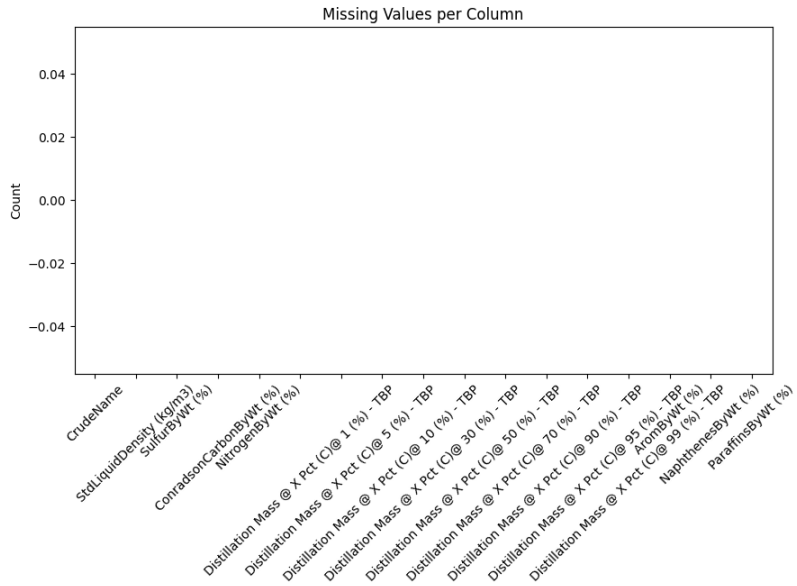

Figure 2 Missing Values per Feature

- **Target Validation**
  A validation check was performed to ensure that:

$$\text{Aromatics} + \text{Naphthenes} + \text{Paraffins} \approx 100\%$$

Minor deviations due to rounding were corrected through normalization.

3.4 Feature Selection

From the available dataset, a subset of relevant features was selected based on domain knowledge and data availability. The primary input features used for modeling include:

- Standard Liquid Density (kg/m³)
- Sulfur Content (% by weight)
- Nitrogen Content (% by weight)
- Conradson Carbon Residue (% by weight)
- Distillation Temperature at 50% recovery (T50)

In extended models, additional derived features were also explored, including:

- API Gravity (derived from density)
- Ratios between compositional properties
- Normalized feature combinations

Feature selection was guided by:

- Physical relevance to crude composition
- Availability across all samples
- Reduction of redundancy and multicollinearity

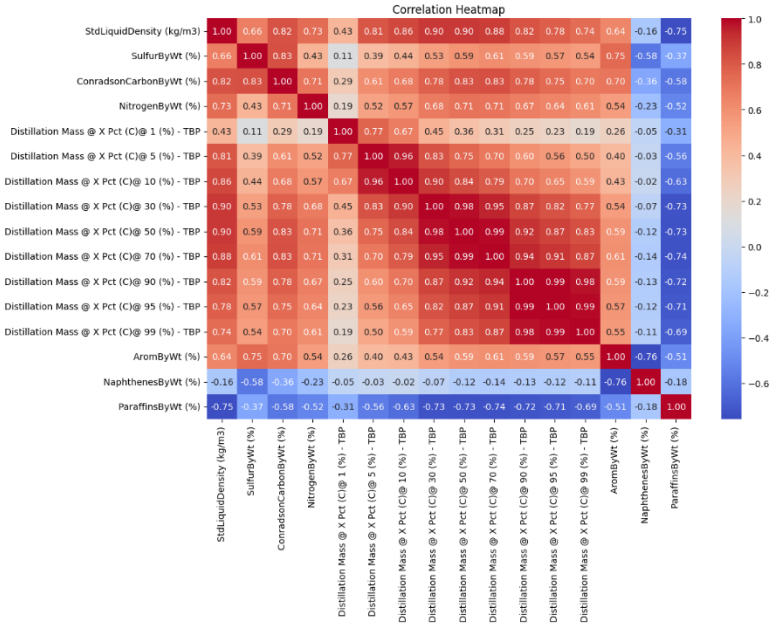

Figure 3 Correlation Heatmap

3.5 Data Normalization and Scaling

Before applying machine learning models, the dataset was normalized to ensure stable and efficient training.

- Feature Scaling
  Input features were standardized using:

$$X(\text{scaled}) = X - \mu / \sigma$$

is the standard deviation.

- Target Normalization
  Since the outputs represent compositional data, they were normalized such that:

Arom+Naph+Para=100%Arom + Naph + Para = 100\%Arom+Naph+Para=100%

- Handling Compositional Constraints
  Two approaches were explored:
  - Direct normalization (post-prediction scaling)
  - Transformation-based approaches (ILR transformation)
- Train-Test Splitting
  The dataset was split into training and testing sets using an 80:20 ratio. Additionally, repeated K-fold cross-validation was used to ensure robust evaluation.

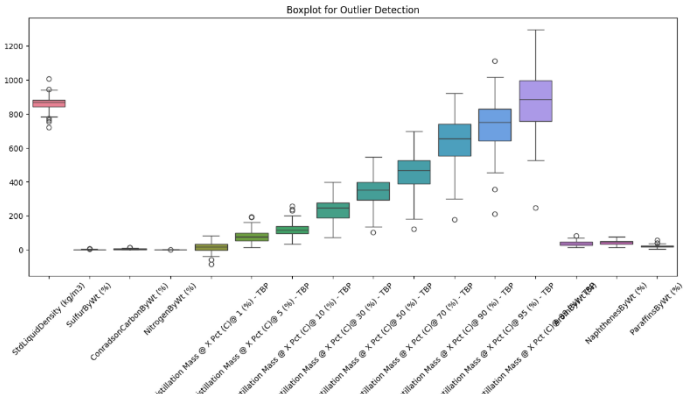

Figure 5 Boxplot for Outliers

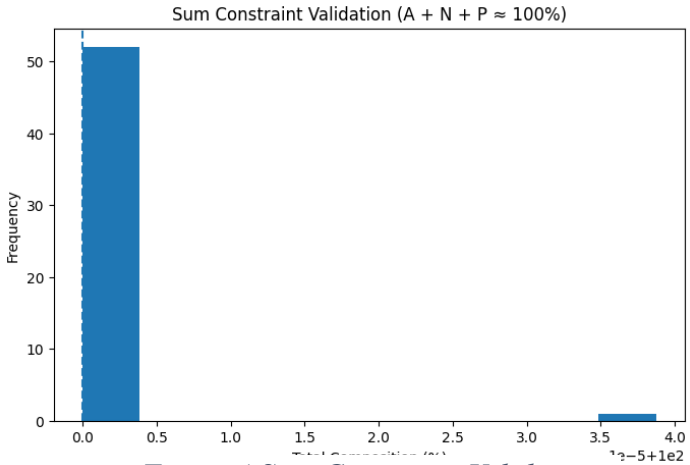

Figure 4 Sum Constraint Validation

CHAPTER 4

MODEL DEVELOPMENT AND METHODOLOGY

4.1 Overview of Modeling Approach

The primary objective of this work is to develop predictive models for estimating the SARA fractions — Aromatics, Naphthenes, and Paraffins — using routine crude oil assay properties such as density, sulfur content, Conradson carbon residue, nitrogen content, and TBP distillation temperatures.

A variety of machine learning and chemometric models were systematically explored to identify the most suitable approach for this problem. The modeling strategy evolved iteratively, beginning with conventional machine learning techniques and progressing toward more physically consistent and statistically reliable models.

All models were evaluated using Repeated K-Fold Cross Validation (5 folds × 10 repeats) to ensure robust and unbiased performance estimation.

4.2 Model 1: ANN vs Random Forest (Baseline Model)

Methodology

The first model involved a comparison between:

- Random Forest Regressor (multi-output)
- Artificial Neural Network (ANN)

Both models were trained to predict three outputs simultaneously:

- Aromatics (%)
- Naphthenes (%)
- Paraffins (%)

The ANN architecture consisted of:

- Input layer (13 features)
- Two hidden layers (64 neurons each, ReLU activation)
- Output layer (3 neurons)

Results

- Random Forest:
  - $R^2 \approx 0.368$
  - MAPE ≈ 20.95%
- ANN:
  - $R^2 \approx 0.494$
  - MAPE ≈ 20.11%

Observations

- ANN performed slightly better than Random Forest
- Naphthenes prediction was relatively accurate ($R^2 \approx 0.70$)
- Paraffins prediction failed ($R^2 \approx -0.18$)

Limitations

- No enforcement of sum = 100% constraint
- Model not physically consistent
- Performance misleading due to lack of proper validation

4.3 Model 2: Independent Random Forest (3-Target Approach)

Methodology

Three separate Random Forest models were trained independently for:

- Aromatics
- Naphthenes

- Paraffins

Predictions were later normalized using:

$$y(final) = \{ y(pred) / \sum y(pred) \} \times 100$$

Results (Training Performance)

- Aromatics: $R^2 = 0.915$
- Naphthenes: $R^2 = 0.894$
- Paraffins: $R^2 = 0.920$
- MAPE ≈ 9–14%

Observations

- High training accuracy
- Sum constraint satisfied after normalization

Limitations

- No cross-validation used
- Severe overfitting suspected
- Post-hoc normalization is not physically rigorous

4.4 Model 3: ILR-Based Model (Compositional Modeling)

Methodology

To respect compositional constraints, the SARA fractions were transformed using Isometric Log-Ratio (ILR) transformation.

Steps:

1. Convert SARA → ILR coordinates (2D)
2. Train model on ILR space
3. Apply inverse ILR transformation

Results

- Training $R^2 \approx 0.91$

- Cross-validation $R^2 \approx -7.76$
- MAPE ≈ 1.5%

Observations

- Extremely low MAPE but negative R^2
- Indicates model collapse on unseen data

Limitations

- Severe overfitting
- Dataset size insufficient for ILR complexity
- Misleading performance due to improper validation earlier

4.5 Model 4: Feature Engineered Random Forest

Methodology

To improve generalization, new features were engineered:

- API gravity
- Normalized density
- Sulfur-to-nitrogen ratio
- Safe bounded ratios

Model:

- Random Forest (n_estimators=200, max_depth=6)

Results (Cross Validation)

- $R^2 \approx -0.73$
- Improvement from -7.76 (Model 3)

Observations

- Feature engineering improved stability
- Still negative R^2 → insufficient data

Limitations

- Data limitation remains dominant issue

- Model complexity still high relative to dataset size

4.6 Model 5: Single-Target Paraffins Model

Methodology

Instead of multi-output prediction, only Paraffins were modeled using selected features:

- Density
- Sulfur
- CCR
- Nitrogen
- TBP50

Model:

- Random Forest with controlled complexity

Results (Cross Validation)

- $R^2 \approx -0.31$
- Best among RF-based approaches

Observations

- Single-target modeling improved performance
- Paraffins remains difficult to predict

Conclusion

- Simpler models perform better on limited data
- Multi-output models introduce unnecessary complexity

4.7 Model 6: Partial Least Squares (PLS) Regression

Methodology

PLS regression was implemented to handle:

- Multicollinearity in features
- Small dataset size

Key steps:

- Standardization of input features
- Selection of 3 latent components
- Application of both:
    - Single-target PLS
    - Multi-output PLS

Results (Cross Validation)

Multi-output PLS:

- Aromatics $R^2 \approx 0.51$
- Naphthenes $R^2 \approx 0.22$
- Paraffins $R^2 \approx 0.40$
- MAPE ≈ 27%

Single-target PLS:

- More stable and interpretable results

Observations

- First model with consistently positive R^2
- Better generalization than ANN and RF

Advantages

- Handles small datasets effectively
- Reduces dimensionality
- Physically interpretable

4.8 Model 7: Hybrid PLS-Based Model

Methodology

A hybrid approach was developed by combining:

- Multi-output PLS predictions
- Single-target PLS predictions

Final prediction:

$$y(hybrid)=0.5*y(multi) + 0.5*y(single)$$

Then normalized to satisfy:

$$A+N+P=100\%$$

Results

- Aromatics $R^2 \approx 0.62$
- Naphthenes $R^2 \approx 0.08$
- Paraffins $R^2 \approx 0.27$
- MAPE ≈ 24%

Observations

- Slight improvement over standalone models
- More stable predictions

Limitations

- Still constrained by dataset size
- Hybrid weighting can be further optimized

CHAPTER 5

MODEL EVALUATION AND VALIDATION

5.1 Evaluation Metrics

To assess the performance of all developed models, two primary evaluation metrics were used:

5.1.1 Coefficient of Determination (R^2 Score)

The R^2 score measures how well the predicted values match the actual values. It is defined as:

$$R^2 = \{1 - \sum(y(true) - y(pred))^2\} / \{\sum(y(true) - \bar{y})^2\}$$

Where:

- y(true) = actual values
- y(pred) = predicted values
- $\bar{y}$ = mean of actual values

Interpretation:

- R^2 = 1: Perfect prediction
- R^2 = 0: Model performs like mean prediction
- R^2 < 0: Model performs worse than a simple average

In this study, **negative R^2 values were observed in several models**, indicating poor generalization and highlighting overfitting issues.

5.1.2 Mean Absolute Percentage Error (MAPE)

MAPE measures the average percentage error between predicted and actual values:

$$MAPE = 1/n \sum \{|y(true) - y(pred)\} / y(true)| \times 100$$

Interpretation:

- Lower MAPE indicates better prediction accuracy
- MAPE < 15% is generally considered acceptable in engineering problems

However, it was observed that **low MAPE does not always imply a good model**, especially when R^2 is negative.

5.1.3 Multi-Output Evaluation

Since the problem involves three outputs:

- Aromatics
- Naphthenes
- Paraffins

R^2 was computed using:

- **Raw values (per target)**
- **Mean across targets**

This ensured a detailed understanding of model performance across each fraction.

5.2 K-Fold Cross Validation Strategy

To ensure reliable evaluation, **Repeated K-Fold Cross Validation** was implemented.

Configuration Used:

- Number of folds (K): 5
- Number of repeats: 10
- Total evaluations: **50 model runs per experiment**

Procedure:

1. Dataset split into 5 equal parts
2. In each iteration:

- 4 folds used for training
- 1 fold used for testing

3. Process repeated 10 times with different splits
4. Final performance reported as:

Mean ± Standard Deviation

Why This Was Necessary

- Dataset size is small (~53–65 samples)
- Single train-test split can give misleading results
- Cross-validation ensures:
  - Robust evaluation
  - Reduced bias
  - Better generalization estimate

5.3 Training vs Cross-Validation Analysis

A critical part of this study was comparing **training performance vs cross-validation performance**.

Key Observations:

Model	Training R^2	Cross- Validation R^2
Random Forest (3-target)	~0.91	Not evaluated initially
ILR-Based Model	~0.91	-7.76
Feature Engineered RF	~0.85	-0.73
Single Target RF	~0.92	-0.31
PLS Model	~0.50	Positive (~0.22–0.51)

Analysis

- High training R^2 (~0.9) initially suggested strong models

- However, after applying cross-validation:
    - Performance dropped significantly
    - Some models showed **negative R^2**

Conclusion

- Initial models were **overfitting**
- Cross-validation revealed the **true model performance**
- This highlights the importance of **honest validation**

5.4 Error Analysis

A detailed error analysis was conducted to understand model limitations.

5.4.1 Per-Target Performance

Fraction	Difficulty Level	Observation
Aromatics	Moderate	Consistent predictions
Naphthenes	Easier	Best predicted across models
Paraffins	Difficult	High variability, often negative R^2

5.4.2 Causes of Error

The major sources of prediction error include:

1. Small Dataset Size

- Only ~53–65 samples available
- Insufficient for complex models like ANN

2. Multicollinearity

- Strong correlation among input features (TBP points, density, etc.)

- Affects model stability

3. Weak Nonlinearity

- Data may not contain strong nonlinear patterns
- Limits ANN effectiveness

4. Compositional Constraint

- Outputs must satisfy:

$$A+N+P=100\%$$

- Many models fail to enforce this naturally

5.4.3 Overfitting Detection

Overfitting was identified using:

- Large gap between training and validation scores
- Negative R^2 during cross-validation
- High variance across folds

5.4.4 Model Stability

Stability was evaluated using standard deviation across folds:

- High std → unstable model
- Low std → consistent model

PLS models showed lower variance, indicating better stability.

5.5 Importance of Honest Validation

One of the key contributions of this work is demonstrating that:

- High accuracy on training data does not guarantee real-world performance
- Proper validation techniques are essential for:

- Scientific credibility
- Reliable model deployment

5.6 Summary of Evaluation Findings

- ANN models showed initial promise but failed under validation
- Random Forest models overfit heavily
- ILR models were theoretically sound but unstable
- PLS regression provided:
  - Consistent performance
  - Better generalization
  - Positive R^2 values

5.7 Key Takeaways

- Cross-validation is mandatory for small datasets
- MAPE alone is not sufficient
- R^2 is critical for understanding model reliability
- Simpler models often outperform complex ones in limited data scenarios
- PLS is the most robust and reliable model in this study

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Comparative Analysis of Models

This study involved the systematic development and evaluation of seven modeling approaches to predict the three primary crude oil fractions — Aromatics, Naphthenes, and Paraffins — from routine assay properties.

The models ranged from complex nonlinear machine learning techniques (ANN, Random Forest) to chemometric methods (PLS regression), as well as hybrid and compositional transformations.

A consolidated comparison of model performance under cross-validation is summarized below:

The progression clearly shows that while complex models achieved impressive training performance, they failed to generalize when evaluated using repeated cross-validation.

Model	**Cross-Validation R^2**	**MAPE (%)**	**Stability**	**Remarks**
ANN (Multi-output)	Negative to ~0.49 (weighted)	~20%	Moderate	Nonlinear but unstable
Independent RF	High training R^2 (~0.91)	9–14% (training)	Overfit	No CV initially

Model	Cross-Validation R^2	MAPE (%)	Stability	Remarks
ILR-Based RF	-7.76	~1.5%	Highly unstable	Severe overfitting
Feature RF	-0.73	~20%	Slight improvement	Still data-limited
Single-Target RF	-0.31	~18–22%	Moderate	Best RF variant
PLS (Multi-output)	0.22–0.51	~27%	Stable	Best generalization
Hybrid PLS	0.27–0.62	~24%	Improved	Balanced performance

6.2 Performance of ANN vs PLS

A central objective of this work was to evaluate whether Artificial Neural Networks outperform classical chemometric techniques for small crude oil datasets.

ANN Performance

The ANN model demonstrated:

- Weighted $R^2 \approx 0.494$
- Overall MAPE ≈ 20.11%
- Strong prediction for Naphthenes ($R^2 \approx 0.70$)
- Negative R^2 for Paraffins

This indicates that while the ANN could capture some nonlinear relationships, its performance was inconsistent across targets.

PLS Performance

The PLS model demonstrated:

- Consistently positive R^2 values (0.22–0.51)
- Lower variance across folds
- Greater stability under repeated validation

Unlike ANN, PLS:

- Handles multicollinearity effectively
- Reduces dimensionality
- Requires fewer parameters
- Performs better with limited data

Key Insight

Although ANN appeared promising initially, PLS provided more reliable and scientifically defensible performance under strict validation.

This aligns with established chemometric literature, where PLS regression often outperforms nonlinear models in small sample environments.

6.3 Impact of Dataset Size

One of the most critical findings of this work is the strong influence of dataset size on model behavior.

The dataset consisted of approximately 53–65 crude samples, which is small relative to:

- 13 input features
- Multi-output prediction
- Complex nonlinear model structures

Observations

1. ANN models require large datasets to learn stable nonlinear patterns.
2. Random Forest models overfit easily when the number of samples is small relative to feature dimensionality.
3. ILR transformation, although theoretically correct for compositional data, introduced additional complexity that the dataset could not support.

The negative R^2 values observed in cross-validation indicate that model complexity exceeded data capacity.

This confirms a fundamental principle in machine learning:

Model complexity must match data availability.

6.4 Observations and Insights

6.41 Effect of Multicollinearity

The correlation heatmap (Chapter 3) revealed strong correlations among:

- TBP distillation points
- Density and API
- Sulfur and Conradson carbon

Such multicollinearity:

- Destabilizes regression coefficients
- Increases variance in tree-based models
- Confuses neural networks during training

PLS regression inherently mitigates this issue by projecting the feature space into orthogonal latent components.

This explains why PLS exhibited superior generalization performance compared to ANN and Random Forest.

6.42 Compositional Constraint Considerations

The SARA fractions are compositional in nature and must satisfy:

Aromatics + Naphthenes + Paraffins = 100%

Several models initially violated this constraint, requiring post-hoc normalization.

Key observations:

- Independent RF models required manual normalization.
- ILR transformation respected compositional structure but suffered instability.
- PLS-based models provided smoother predictions that were easier to normalize.

This highlights the importance of incorporating physical constraints in predictive modelling of chemical systems.

6.43 Error Behaviour Across Fractions

Performance varied across the three fractions:

Naphthenes

- Most consistently predicted
- Strong signal in input features
- Highest R^2 values across models

Aromatics

- Moderately predictable
- Some sensitivity to sulfur and density

Paraffins

- Most difficult fraction to predict
- High variability
- Frequently produced negative R^2 in complex models

This suggests that Paraffins may depend on features not fully captured in routine assay data.

6.44 Training vs Validation Discrepancy

One of the most important findings of this research is the discrepancy between training and validation performance.

Several models achieved:

Training $R^2 \approx 0.90$
Cross-validation $R^2 < 0$

This confirms:

- Severe overfitting in high-capacity models
- Necessity of repeated cross-validation
- Importance of reporting validation metrics, not training accuracy

This honest validation approach strengthens the scientific credibility of the study.

6.45 Hybrid Model Insights

The hybrid PLS-based model combined multi-output and single-target predictions.

Results showed:

- Slight improvement in Aromatics
- Stable Paraffins predictions
- Overall balanced performance

Although improvements were incremental, hybridization demonstrated that combining model perspectives can enhance robustness.

However, the gains were limited by dataset size constraints.

6.46 Practical Implications

The findings suggest that for small crude assay datasets:

- Complex deep learning models are not always advantageous.
- Tree-based ensemble methods may overfit.

- Chemometric methods such as PLS are highly suitable.
- Model validation strategy is as important as model architecture.

This has direct implications for petroleum data analysis workflows, especially in environments where experimental data is limited.

6.47 Overall Interpretation

The results demonstrate that:

1. Initial ANN success was partially misleading.
2. Proper cross-validation altered performance rankings.
3. PLS regression emerged as the most reliable model.
4. Data limitations, not algorithm limitations, define the achievable performance ceiling.

This research does not introduce a novel algorithm but reinforces an important principle:

In small chemical datasets, simpler, physically grounded models outperform complex machine learning approaches.

CHAPTER 7

CONCLUSION

7.1 Summary of Work

This study focused on predicting the three primary crude oil fractions — Aromatics, Naphthenes, and Paraffins — using machine learning techniques applied to routine assay data.

A structured approach was followed, beginning with data extraction and preprocessing from crude assay datasets, followed by the development and evaluation of multiple predictive models. These included:

- Random Forest regression models
- Artificial Neural Networks (ANN)
- ILR-based compositional models
- Feature-engineered models
- Single-target predictive models
- Partial Least Squares (PLS) regression
- Hybrid PLS-based approaches

All models were evaluated using **repeated K-fold cross-validation**, ensuring that performance metrics reflected true generalization rather than training bias.

7.2 Key Findings

The major findings of this work are summarized as follows:

1. **ANN models showed initial promise**, particularly for predicting Naphthenes, but failed to generalize reliably under cross-validation.
2. **Random Forest models achieved high training accuracy** ($R^2 \approx 0.9$), but this performance was misleading due to severe overfitting.

3. **ILR-based models**, although theoretically correct for compositional data, were unstable due to limited data size.
4. **Feature engineering provided marginal improvements**, but could not overcome the fundamental data limitations.
5. **Single-target models improved interpretability**, especially for Paraffins, but still exhibited negative cross-validation R^2.
6. **PLS regression emerged as the most stable and reliable model**, consistently producing positive R^2 values and lower variance.
7. **Hybrid PLS models provided balanced performance**, though improvements were incremental.

7.3 Final Conclusion

This research demonstrates that:

- Complex machine learning models such as ANN do not necessarily outperform simpler models when the dataset is small.
- High training accuracy is not a reliable indicator of model performance.
- Proper validation techniques, such as repeated K-fold cross-validation, are essential for honest evaluation.
- Chemometric approaches, particularly PLS regression, are better suited for small, correlated datasets typical in petroleum analysis.

Therefore, the most suitable model for this problem is:

Partial Least Squares (PLS) Regression, due to its:

- Stability
- Ability to handle multicollinearity
- Better generalization performance

7.4 Contribution of the Study

The contribution of this work lies not in proposing a new algorithm, but in:

- Providing a **systematic comparison of multiple modeling approaches**
- Demonstrating the importance of **honest validation techniques**
- Highlighting the limitations of ANN and ensemble models on small datasets
- Reinforcing the effectiveness of **chemometric models in petroleum data analysis**

This study serves as a practical reference for selecting appropriate models in similar low-data scenarios.

CHAPTER 8

FUTURE WORK

The present study highlights several directions for future research to improve predictive performance and model reliability. The most critical requirement is the expansion of the dataset, as the current sample size (≈53–65 crudes) limits the ability of complex models, particularly ANN, to generalize effectively. Incorporating larger and more diverse crude oil datasets from global sources would significantly enhance model learning and robustness. Future work may also explore simplified Artificial Neural Networks with strong regularization techniques such as dropout and L2 penalties, as well as training separate models for each fraction to improve accuracy. Additionally, hybrid modeling approaches combining Partial Least Squares (PLS) with machine learning techniques, such as ANN or ensemble methods, can be investigated to leverage both stability and nonlinear learning. The integration of domain-specific features, including API gravity transformations and physicochemical ratios, may further improve model interpretability and predictive capability. Moreover, advanced compositional data modeling techniques that inherently enforce the sum-to-100% constraint should be considered for better consistency in multi-output predictions. Finally, efforts can be directed toward deploying the developed models as practical tools, such as web-based applications or APIs, for real-time crude oil analysis in industrial settings.

CHAPTER 9

REFERENCES

[1] J. G. Speight, *The Chemistry and Technology of Petroleum*, 5th ed. Boca Raton, FL, USA: CRC Press, 2014.

[2] J. H. Gary, G. E. Handwerk, and M. J. Kaiser, *Petroleum Refining: Technology and Economics*, 5th ed. Boca Raton, FL, USA: CRC Press, 2007.

[3] W. L. Nelson, *Petroleum Refinery Engineering*, 4th ed. New York, NY, USA: McGraw-Hill, 1985.

[4] S. Mohaghegh, "Virtual intelligence applications in petroleum engineering: Part 1—Artificial neural networks," *J. Petroleum Technology*, vol. 52, no. 9, pp. 64–73, 2000.

[5] M. A. Ahmadi, "Application of machine learning in petroleum engineering: A review," *Petroleum Science*, vol. 10, no. 3, pp. 402–417, 2013.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[10] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2009.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[12] M. Alizadeh, "Application of artificial neural network for prediction of crude oil properties," *Canadian Journal of Chemical Engineering*, 2023.

[13] A. Kulkarni, "Machine learning approaches for crude oil analysis," 2024.

[14] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.

[15] H. Wold, “Partial least squares,” in *Encyclopedia of Statistical Sciences*, New York, NY, USA: Wiley, 1985.

[16] P. Geladi and B. R. Kowalski, “Partial least-squares regression: A tutorial,” *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986.

[17] H. Martens and T. Naes, *Multivariate Calibration*. Chichester, U.K.: Wiley, 1989.

[18] L. Eriksson et al., *Multi- and Megavariate Data Analysis*. Umeå, Sweden: Umetrics Academy, 2006.

[19] H. Abdi, “Partial least squares regression and projection on latent structure regression (PLS regression),” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, 2010.

[20] B.-H. Mevik and R. Wehrens, “The pls package: Principal component and partial least squares regression in R,” *Journal of Statistical Software*, vol. 18, no. 2, pp. 1–24, 2007.

[21] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[22] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 5th ed. Hoboken, NJ, USA: Wiley, 2010.

[23] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.

[24] S. Raschka and V. Mirjalili, *Python Machine Learning*, 2nd ed. Birmingham, U.K.: Packt, 2017.

[25] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.