Wax Disappearance Temperature Modeling with Machine Learning Techniques

By

Ainur Bayanova

Thesis submitted to the School of Mining and Geosciences of Nazarbayev University in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Petroleum Engineering

Nazarbayev University

April 2025

ORIGINALITY STATEMENT

I, Ainur Bayanova, hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at Nazarbayev University or any other educational institution, except where due acknowledgement is made in the thesis.

Any contribution made to the research by others, with whom I have worked at NU or elsewhere is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signed on **15.04.2025**

_____

ABSTRACT

Wax precipitation along with asphaltene and scale deposition present major problem for flow assurance in the oil production. Mitigation and prevention measures for wax precipitation requires understanding of wax forming conditions. Key parameter to define wax forming borderline is Wax Disappearance Temperature (WDT). The available methods to determine WDT include laboratory measurements, empirical correlations, thermodynamic modeling, and data-driven approaches. Laboratory measurements are costly and may logistically be challenging, thermodynamic approaches require detailed and accurate characterization of production fluid, and some may experience convergence issues. Data driven approaches provide a good alternative to earlier methods with almost no loss to accuracy of data. Use of machine learning (ML) techniques to determine WDT is reported in more recent literature, though limited studies are conducted. Use of more recent developments in ML for WDT determination are not well documented. Hence, an attempt was made to develop intelligent models using decision tree (DT) with boosting algorithms for WDT prediction: AdaBoost, Gradient Boosting Machines, XGBoost and CatBoost. Conventional Linear Regression (LR) and K-Nearest Neighbor (KNN) methods were also used for comparison with DT approaches. A detailed analysis of the input data from published WDT experimental studies in literature was performed that includes selection of input features, building dataset, validating data sources, and the input data analysis with statistical tools and graphical analysis. This research work resulted in building a database with 380 data points of experimental WDT. Overall, all DT boosting algorithms have performed better than the conventional LR and KNN techniques, with XGBoost and CatBoost being the top performers (R2 = 0.996 and RMSE = 0.791 and 0.7916, respectively). AdaBoost could be a model of choice if simplicity is preferred with negligible difference in performance (R2 = 0.9945 and RMSE = 0.9272) compared to the top performer. Model performance evaluation was based on both statistical assessment and graphical analysis such as parity plots and error distribution plots. Further trend analysis was assessed where all models indicate WDT increase with MW increase that validate models' generalization and predicting capacity. The developed XGBoost and CatBoost models are superior to the existing models reported in literature, both thermodynamic and data-driven methods. The developed XGBoost and CatBoost models offer higher accuracy and better generalization with possible implications in flow assurance schemes involving wax.

Keywords: flow assurance, wax disappearance temperature, machine learning (ML)

ACKNOWLEDGMENTS

First, I would like to thank my thesis supervisor, Dr. Ali Shafiei for his constant support and supervision and making this work happen in a timely manner.

I'm also thankful to Dr. Hisham Khaled Ben Mahmud and Dr. Mian Umer Shafiq for allocating their time to serve in thesis committee and conduct thorough review with valuable feedback.

I would also like to express my gratitude towards petroleum engineering faculty members for contributing their best to raise new generation of petroleum engineers. And I would like to acknowledge support of Maryam Mahmoudi, research group member, for introducing me to the fascinating world of machine learning.

Last, but not least, my appreciation to my family and friends for their emotional support, especially to my father, a strong believer of constant learning and academic world. And extending thanks to my classmates, Dias, Elmira, and Qaisar for making this journey a true joy!

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In this chapter, a brief background is provided on wax precipitation issue in the oil production for flow assurance in oil fields along with its common mitigation and prevention measures. The focus here is on prevention that requires understanding of wax forming process. The the key parameter to prevent wax precipitation is the wax appearance borderline. The preference to define wax formation by wax disappearance temperature (WDT) rather than appearance is discussed. The formulated problem statement aimed at preventing the problem is presented along with relevance of this work to oil industry. Then, research objectives, research methodology, and thesis organization and structure are also described.

1.1 Background

Production chemistry may present major problems for flow assurance, such as wax, asphaltenes, and scale deposition. Among these deposits asphaltenes are frequently mistaken for or classified alongside waxes. While, both wax and asphaltenes are of organic nature, precipitate in solid forms from hydrocarbon systems, their chemistry is very different, asphaltene structure being more complex. And while asphaltene particles would not melt, wax precipitation is a reversible process, and the solid particles melt. Thus, wax deposits are easier to mitigate but still present a wider challenge. Many problems initially associated with asphaltenes were eventually found to result from wax deposition (Becker, 2000).

Wax crystals are long-chain alkane hydrocarbons that solidify at temperatures below wax appearance temperature (WAT). A wax sample recovered during workover operation is shown in Figure 1.1. Wax deposition issues might be present both in oil systems and condensates with some condensate reservoirs containing wax levels exceeding 30%. The wax content in oil often increases as oil gravity increases. However, a higher wax content typically

has little impact on overall oil density. Unlike waxes, presence of asphaltenes decreases as API gravity increases, and they are almost never found in condensates (Bellarby, 2009).

Figure 1.1. Wax sample recovered with sucker rod during tubing replacement (Bellarby, 2009)

Wax deposits in well may impact well performance in following ways: If flowing fluid temperature along well drops below wax appearance temperature, then waxes will deposit and accumulate along the tubing wall reducing tubing inner diameter and leading to restricted production. Wax deposition can potentially block the tubing, entirely. However, it is more common for an equilibrium to be established, where flow (shear) and buildup of thermally insulating wax limits further accumulation. Low concentrations of high-molecular-weight waxes (with high melting points) tend to form hard deposits, while high concentrations of lower-molecular-weight waxes (with lower melting points) often lead to softer but more abundant deposits. Wax formation at installation depths of downhole devices such as safety valves should be avoided to prevent well safety issues (Bellarby, 2009).

When a well is shut-down, wax molecules within hydrocarbon structure will crystallize if temperature in tubing fluid drops below wax appearance temperature building a strong gel that will resist tubing fluid flow once the well is put back-on for production causing well re-start issues. Gel strength increases closer to the wellhead where temperature is lower. Yield stress will need to be modeled upfront to estimate pressure drop required to break the gel

strength and re-start the well. Pressure differential as high as 14,000 psi might be required in some cases, like sub-sea wells, which may not be feasible. In such cases, it may be necessary to displace the reservoir fluids below wax appearance depth after a shut-down. Additionally, insulation can help by slowing the rate of cooling (Singh et al., 2006).

In addition to above insulation and well fluid displacement techniques following mitigation measures could be considered: Hot oiling: commonly used in land wells, hot fluid is circulated down the well, preferably reverse circulation in case of tubing blockage with deposits. However, the efficiency is conditional with heat loss to surrounding formation and well construction. Deposits mechanical removal with well intervention: scraper and gauge run but these present fishing risks as well as physical completion equipment damage risks. Diluent and solvent injection either with continuous bullheading via annulus or batch treatments. Fluid selection is critical here, assessing cost vs efficiency, xylene and toluene are efficient in wax dissolution. However, a certain cost is associated and continuous bullheading will be costly and these fluids are toxic and may damage downhole equipment seals. Pour point depressant (PPD) and wax inhibitors are usually deployed in concentration of 100-1000 ppm may reduce wax pour point by 30°F to 50°F, commonly used in surface and subsea pipelines. PPDs reduce the likelihood of waxes co-precipitating into hard solids. However, once wax has formed then PPDs have little to no effect on its dissolution. This requires additional mitigation measures to be considered. Continuous injection of inhibitors and PPDs can be costly, and its use is limited to deep subsea projects and well shut-down and re-start operations. Electrical heat tracing is commonly used for fluid heating in surface facilities and downhole. Highly effective but a costly strategy (Biao & Lijian, 1995). All listed mitigation measures are either conditional (i.e., requires electricity) or do not warrant fail safe performance and are associated with operational risks and are costly. The most efficient way to mitigate wax deposition is to prevent wax

precipitation. For this, wax precipitation mechanism and parameters need to be well understood and assessed.

Accurate characterization of wax can be achieved if the distribution of alkanes and other hydrocarbons is well understood. Another important parameter to affect wax precipitation is system pressure. At low pressures, lower-molecular-weight hydrocarbons, which usually help to keep wax in solution, transition to the gas phase. As a result, the wax appearance temperature increases, by approximately 7–10°F for every 1000 psi drop in pressure below the bubble point. These two the input parameters, fluid composition and system pressure, help to assess wax precipitation borderline. Number of temperature values are used in some research works to define this borderline: Wax appearance temperature (WAT): the temperature at which wax becomes detectable for the first time. Cloud point: same as WAT, though it specifically refers to the point at which wax crystals begin to cloud the hydrocarbon solution. Pour point: the temperature at which fluid stops to flow. Yield stress or gel strength: measurement of pressure required to re-start fluid flow after it has become stationary (Bellarby, 2009).

From all above, WAT is the main parameter commonly used to assess the wax appearance borderline. However, as Ji et al. (2004) argue, WAT is not a definitive equilibrium point as it depends heavily on measurement conditions, specifically the cooling rate. Faster cooling rates result in lower measured WAT values. For instance, WAT determined using visual microscopy is typically 10–20°C higher than values measured with laser-based solid detection systems, differential scanning calorimetry (DSC), or viscometry (Ji et al., 2004)

Wax Disappearance Temperature (WDT) represents a definitive solid-liquid equilibrium point unlike the Wax Appearance Temperature (WAT). Reliable experimental techniques such as equilibrium step heating, significantly enhance accuracy of WDT measurements. As a result,

measured WDT values align closely with standard values within a reasonable error range across different laboratories. There are different ways to define WDT: a) direct, laboratory measurements, b) theoretical, and empirical calculations, c) thermodynamic modeling and d) intelligent, data-driven methods (Bian et al., 2019).

1.2 Problem definition

Wax deposition is one of the major flow assurance issues encountered in the oil production and an efficient mitigation strategy would be to prevent it in production stream. The most efficient prevention technique is to keep production stream fluid temperature above the wax appearance borderline and main control parameter here is wax disappearance temperature (WDT). Most accurate method for WDT estimate is direct laboratory measurement. However, this could be costly and not practical in field applications. Hence, indirect ways are investigated to estimate WDT within acceptable range of accuracy: thermodynamic, empirical, and data-driven methods. Thermodynamic models require complex fluid characterization data and may result with convergence issues. Empirical models are limited with generalization and accuracy. Hence data-driven methods present an alternative to above approaches because of their high accuracy and low cost. Yet, very limited works are published where intelligent methods have been applied and much recent developments in ML offer an opportunity to develop more robust and reliable models with higher prediction for WDT estimation.

1.3 Relevance to the industry

Wax deposition is a common worldwide issue in oilfields that report high wax/paraffin content in the crude oil such as West Siberian Basin in Russia aggravated with cold climate, North Sea fields with challenge of colder waters, Alaska North Slope fields in the US with wax deposition issues mainly encountered in pipeline transportation, Lloydminster field in Canada, Campos basin in Brazil with additional challenge in subsea pipelines, Cabinda fields in Angola, Niger Delta fields in Nigeria with added challenge of deep-water, Daqing oil field in China, especially in cold seasons, Mangala oil field in India, Fateh oil field offshore Dubai with deposition

challenges in subsea pipelines. Wax deposition because of high paraffin content is also experienced in some oilfields in Kazakhstan such as Kumkol field, Akshabulak field (Boranbayeva et al, 2024), South Turgai oil field (Kozhabekov et al, 2019). While wax deposition is an expected problem in oilfields with high paraffinic oil, wax precipitation might also start to occur in oilfields that initially have not experienced wax issues or oilfields with low paraffinic content at a later stage of field development because of decline in reservoir pressure and temperature and reduced production rates that can cause laminar flow and increase in water cut.

As described earlier in this text, effective prevention measure for all these fields requires understanding of wax forming mechanism, specifically accurate definition of WDT. With limited availability and high cost of laboratory methods and complexity of work with thermodynamic models, data-driven methods represent the best alternative approach to estimate WDT. Additionally, robust yet accurate data-driven method to establish WDT would allow to extend application of wax prevention and mitigation measures to field practices without bottleneck in laboratories, thus preventing and mitigation flow assurance issues related to wax precipitation in real-time and ensuring flawless production operations in field.

1.4 Objectives of the thesis

The proposed research work aims to first to critically review the existing published works on laboratory, thermodynamic, empirical, and data-driven studies for WDT determination and prediction and then to develop and test intelligent models using ML-based techniques that bring additional value to the existing literature in terms of WDT prediction accuracy. Here are the main and specific objectives of the proposed research:

1.4.1 Main objectives

- To analyze wax precipitation mechanism to identify main fluid or environmental parameters that determine WDT.
- To build dataset from WDT data reported in the literature, assess and validate original data source for each reported data, perform statistical analysis of the dataset, exclude inconsistent and erroneous values, if any.
- To run decision tree methods with boosting algorithms such as AdaBoost, Gradient Boosting Machines, XGBoost and CatBoost to predict WDT on the compiled dataset.
- To assess performance of four DT methods with boosting algorithms against K-nearest neighbor and linear regression techniques.
- To compare results of DT methods against previous studies to assess if these recent developed intelligent models bring added value in terms of accuracy and practicality compared to existing reported intelligent models for prediction of WDT.

1.4.2 Specific objectives

- To locate most accurate technique among recently developed machine learning methods to model WDT.

1.5 Research methodology

First, the literature on WDT prediction is critically reviewed and analyzed to evaluate and compare the existing methods for WDT determination and to verify the need for further improvements. The laboratory data was gathered from these works to build a reliable dataset where the data sources are assessed and validated in term of their quality prior use. The the input data was thoroughly analyzed, and data pre-processing performed to identify outliers that distort the data modeling process. In the next stage, the target ML methods including DT techniques with optimizing algorithms were used to build the intelligent models. Here, techniques such as hyper-parameters fine-tuning were applied for the best

fit. Then, the model outputs and performances were compared with each other using conventional statistical approaches and graphical interfaces to select the top performers. The results were then validated with fundamental approach of trend analysis. In addition, a comparative analysis was performed with previous data-driven methods published in literature to re-validate if the target ML techniques bring added value for WDT prediction accuracy and robustness of the proposed models.

1.6 Thesis structure

This research work is presented in five sections. In Chapter 1 (Introduction), the background for wax precipitation issue with mitigation and preventive measures is set, that requires wax forming process characterization and the the key parameter – wax disappearance temperature or WAT definition, formulating the problem statement that initiated this research work with relevance to industry practices and setting clear deliverables expected this work. In Chapter 2 (Literature review), a detailed literature review on WDT determination is presented, assessing each of the methods: laboratory, thermodynamic, empirical, and data-driven. An interesting chronological development can be observed from study of historical publications on wax precipitation subject to help set fundamentals and define the trend in this subject. In Chapter 3 (Methodology), ML methods are introduced in general with focus on specific target ML techniques such as decision tree methods with boosting algorithms. In addition, a clear scope of work with detailed the input data analysis is set, thoroughly investigating each data source and cherry-picking the input data with further data analysis using statistical and graphical methods. In Chapter 4 (Results and Discussion) results of select DT methods: AdaBoost, Gradient Boosting Machines, XGBoost and CatBoost are presented. Results are also compared to conventional methods such as K-nearest neighbor and linear regression. Best performers are selected and are further compared to those ML works published in the literature on WDT determination to validate if new techniques bring added value to the subject. Certain observation calls out for additional discussion as listed at the end of Chapter 4. Lastly, in Chapter 5 (Conclusion and Recommendations) summary of conclusions that result from this research work are provided and further area to improve and adopt techniques in practice in field is proposed.

2 LITERATURE REVIEW

Four different methods have been outlined on WDT determination: laboratory measurements, thermodynamic modeling, theoretical calculations and intelligent, data-driven methods. Literature published on this subject was evaluated with regards to each of these methods as discussed in this chapter. First, different methods and devices for experimental measurements are discussed, with benefits and drawbacks of each method and if method is regulated with any industry standards. Experimental data published in literature is reviewed in scope of the measurement method used for study. Then, studies with thermodynamic modeling are reviewed in chronological order to illustrate the drawback of each model that was overcome by subsequent studies. Empirical models were then reviewed that compared various thermodynamic models with the experimental data obtained in the course of the work and with empirical correlation if any was developed. Lastly, data-driven methods that were published in literature were analyzed and models' performance compared using statistical metrics.

2.1 WDT Laboratory Measurements

2.1.1 WDT Laboratory Measurement Methods

A list of techniques available for Wax Disappearance Temperature (WDT) measurement with a brief description of the method, their benefits and drawback, and ability to accommodate high pressure testing, and if the measurement technique is regulated or referenced in American Society for Testing and Materials (ASTM) standards or not are summarized in Table 2.1. It should be noted here that Table 2-1 only includes the most commonly used methods.

Table 2.1. WDT measurement methods comparison.

Measurement method	Description	Advantages	Disadvantages	High Pressure (HP)	ASTM Standard	Reference
Differential Scanning Calorimetry (DSC)	DSC measures the heat flow associated with phase transitions as the sample is heated or cooled. The WDT is determined when the endothermic peak associated with wax crystal dissolution disappears.	high sensitivity, additional information on heat of transition	specialized equipment, time-consuming sample preparation	Std Low Pressure (LP). Custom HP available.	ASTM D4419	Monger-McClure et al. (1999) Zhao et al. (2015) Juyal et al. (2010) Alcazar-Vara and Buenrostro-Gonzalez (2013) Chen et al. (2014)
Cross-Polarized Light Microscopy (CPLM)	CPLM observes wax crystals under polarized light. As the sample is heated, wax crystals disappear at the WDT, which can be visually observed and recorded.	direct visualization of wax behavior, simple setup for qualitative assessment	less quantitative compared to other methods, can be subjective without proper automation	Std LP. Custom HP required.	*Not explicitly referenced in ASTM (D2500, D3117)*	Monger-McClure et al. (1999) Chen et al. (2014) Elsharkawy et al. (2000) Sarica et al. (2008)

Viscosity Measurements	The viscosity of a waxy fluid decreases significantly when wax crystals dissolve. The WDT can be identified as the temperature at which a noticeable change in viscosity is observed during heating.	standard viscosity measurement equipment, simple and cost-effective.	may not precisely pinpoint the WDT, affected by other fluid properties.	HP used	ASTM D2983	Elsharkawy et al. (2000) Han et al. (2013)
Nuclear Magnetic Resonance (NMR)	Low-field NMR detects changes in the mobility of hydrogen atoms in the fluid as wax crystals dissolve. The WDT is identified when the signal indicates the disappearance of solid wax.	high accuracy, non-invasive and rapid test	specialized NMR equipment, expensive setup	HP used	Not specifically mentioned in ASTM	Zhao et al. (2015)
Turbidimetry	Turbidimetry monitors the optical clarity of the fluid. As wax crystals dissolve, the fluid becomes less turbid. The WDT is determined when the fluid becomes completely clear.	simple and inexpensive, directly measures clarity	limited precision, depends on sample homogeneity	Std LP. Custom HP available.	ASTM D5773, ASTM D8420	Dantas Neto et al. (2009)
Rheology Analysis	Rheological properties (e.g., shear stress and strain) are measured as	provides insights into flow	equipment-intensive, requires	HP used	*Not explicitly*	Mehrotra & Bhat (2007)

	the sample is heated. The WDT is the temperature at which the behavior changes from solid-like to liquid-like.	behavior, useful for flow assurance studies	expertise for data interpretation		*referenced in ASTM*	Taraneh et al. (2008)
Electrical Conductivity Method	Electrical conductivity changes as wax crystals dissolve in conductive crude oils. The WDT is identified as the temperature at which the conductivity stabilizes.	quick and inexpensive, useful for conductive systems	limited to conductive fluids, affected by other impurities	Std LP. Custom HP required.	Not covered in ASTM	Chen et al. (2021)
Ultrasonic Velocimetry	Ultrasonic waves are passed through the sample, and changes in wave velocity are monitored. As wax disappears, the velocity stabilizes, indicating the WDT.	non-destructive, high sensitivity to phase changes	requires specialized ultrasonic equipment, may need calibration for different fluids	HP used	Not covered in ASTM	Chen et al. (2014) Wang et al. (2022)

Monger-McClure et al. (1999) compared DSC, CPM, filter plugging (FP), and Fourier transform infrared (FTIR) energy scattering techniques for cloud point measurements and concluded that if sample is in limited amount, then the DSC and CPM are recommended. If live conditions are needed then the FP and FTIR are recommended, best in terms of accuracy being the DSC followed by the CPM, and FP and FTIR provided acceptable results. They have also noted the impact of cooling rates, pressure, and water content on WDT measurements. ASTM visual technique are limited to transparent fluids. Zhao et al. (2015) have compared the

DSC, near–infrared (NIR) spectroscopy and NMR methods and concluded that DSC offers "high repeatability and accuracy" and also characterizes the wax forming process itself. Use of NIR is beneficial for "non-quiescent and thermal equilibrium conditions" where use of DSC is limited because of sub-cooling effects. NMR offers accurate WDT readings and quantifies the amount of solids and liquids. Juyal et al. (2010) have demonstrated a need for HP-DSC to account for effect of pressure and dissolved gas presence. Alcazar-Vara and Buenrostro-Gonzalez (2013) observed results from DSC agreeing with FTIR results; while rheometry was slightly off, either higher or lower for different samples, and WAT detection was challenging for densiometry method because of crude oil composition. Overall, they conclude that both rheometry and densiometry require specific amount of sample to wax in solution to be able to detect WAT. Chen et al. (2014) also have pointed out ASTM visual methods' limitation to transparent fluids and suggest CPM and DSC use for dark fluids. Based on their literature review they suggest DSC and rheometry being two of the most common methods used and carry on with introduction of ultrasonic device to focus on pressure and light components effect on WAT readings. Comparing results from DSC, rheometry, and ultrasonic device they conclude that the DSC offers the most accurate result followed by ultrasonic technique and lowest reading for rheology.

To summarize, each of above methods has its own strengths and limitations, and the choice of method depends on factors such as sample composition, required accuracy, operating conditions, available equipment, and cost. For the most accurate results, DSC and NMR are often preferred in research and industry projects. However, simpler methods like turbidimetry or viscosity measurements may suffice for routine analyses. For high-pressure WDT measurements, NMR, rheology analysis, viscosity measurements, and ultrasonic velocimetry

are the most suitable, given their compatibility with high-pressure systems. Other methods provide custom made setup with HP cell such as DSC and CPM.

It should be noted here that, not all the mentioned Wax Disappearance Temperature (WDT) measurement methods are explicitly referenced in American Society for Testing and Materials (ASTM) standards for WDT measurement. ASTM standards are highly specific and typically provide a guidance for commonly accepted methods in industry. Currently, ASTM standards related to wax behavior primarily focus on wax appearance temperature (WAT) and pour point determination. However, not all WDT measurement methods are standardized. Key ASTM standards related to wax behavior characterization are as the following:

- ASTM D4419–90: Standard Test Method for Measurement of Transition Temperatures of Petroleum Waxes by Differential Scanning Calorimetry (DSC).
- ASTM D5771–21: Standard Test Method for Cloud Point of Petroleum Products and Liquid Fuels (Optical Detection Stepped Cooling Method).
- ASTM D5773–21: Standard Test Method for Cloud Point of Petroleum Products and Liquid Fuels (Constant Cooling Rate Method).
- ASTM D8420–21: Standard Test Method for Wax Appearance Temperature and Wax Disappearance Temperature of Petroleum Products and Liquid Fuels.
- ASTM D97–17b: Standard Test Method for Pour Point of Petroleum Products.

The most closely aligned methods with ASTM standards for wax behavior are DSC and turbidimetry. While other methods like CPLM, NMR, rheology, electrical conductivity, and ultrasonic velocimetry are either non-standardized or not used in specialized contexts. These advanced techniques are valuable in research and industry applications but are not yet formalized within ASTM standards for WDT measurement.

2.1.2 Published Laboratory Works

Most of the literature published on WDT refer to laboratory WDT measurements obtained in the studies outlined in Table 2.2. The measurement devices used in experiments and test pressures are listed in the table.

Table 2.2. Published laboratory WDT data and measurement methods.

Reference data source	Pressure	Measurement method
Robles et al. (1995)	Atmospheric	DSC (Perkin–Elmer DSC7), X-ray (Siemens D500 diffractometer)
Metivaud et al. (1999)	Atmospheric	DSC (Perkin–Elmer DSC7), X-ray (Siemens D500 diffractometer)
Dauphin et al. (1999)	Atmospheric	Optical device (Visual observation)
Daridon et al. (2002)	0 – 100 MPa	HP cross-polar microscope (Visual observation)
Ji et al. (2003)	Atmospheric	Not mentioned
Milhet et al. (2005)	0 – 100 MPa	HP cross-polar microscope (Visual observation), X-ray diffractor (Philips X'Pert)
Rizzo et al. (2007)	0 – 150 MPa	Special HP Optical device (Automated detection)
Mansourpoor et al. (2019)	Atmospheric	Viscometer (SVM300 Anton Paar), DSC (DSC823 by Mettler Toledo)
Shariatrad et al. (2022)	Atmospheric	Optical device (Visual observation)

Robles et al. (1995) and Metivaud et al. (1999) performed WDT measurements under atmospheric pressure using Perkin-Elmer DSC7 Differential Scanning Calorimetry (DSC) device and Siemens D500 diffractometer X-ray. Metivaud et al. (1999) reported further experiment details: samples were placed in DSC device at room temperature and later heated at a rate 2°K/min and transition temperature was recorded. X-ray scanning was performed at temperatures just below WDT to verify no solid crystal is present in the sample.

Dauphin et al. (1999) used laboratory apparatus that been used since 1992 by Daridon et al. (2002), which is basically a fully visible sapphire cell placed inside glass casing with heat-conducting transparent oil bath. The device allows recording both temperature and pressure. Samples are constantly stirred and heated at rate 0.5°K/hour and solid mass disappearance is observed and recorded.

Daridon et al. (2002) studied the effects of pressure on wax precipitation by measuring wax melting temperatures in the range of pressures up to 100 MPa (in 20 MPa step increase). Samples are placed in cell below WAT, then slowly heated to record a rough WDT value. Then, this cycle is repeated where the cell is cooled to temperature 2°K below the estimated WDT and increased in 0.1–0.2°K steps every 5–10 minutes to record more exact WDT values. Pressure is recorded to verify the equilibrium point. System pressure is bled off to reference pressure at the end of each temperature cycle. This method offers results repeatable within 0.2°K.

Ji et al. (2003) strongly opposes use of WAT for wax characterization as WAT does not represent solid-liquid equilibrium (SLE) point and is affected by testing process such as cooling rate where faster cooling rates may result in lower WAT values. Ji et al. (2003) also suggest that measurement method affects WAT, referencing to Ronningsen et al. (1991), who reported WAT values higher by 10–20°C when detected using visual microscopy compared to DSC, laser-based solids detection systems, and viscometry. Hence, use of WDT is strongly advised as it represents the SLE point. Though measurement technique may also affect WDT test accuracy, and equilibrium step heating is recommended for convergent WDT test results. To further support WDT use versus WAT, Ronningsen et al. (1991) observation is cited where WDT was reported 28°C higher than WAT when using visual microscopy. Ji et al. (2003) have developed a new thermodynamic model and validated the model using laboratory data for

binary systems from Robles et al. (1996) and ternary systems from Metivaud et al. (1999). Additionally they have generated own laboratory data for ternary system C_6+C_{16}+C_{17} and binary systems C_6+C_{16}, C_6+C_{17}, C_{16}+C_{18}, C_{16}+C_{20}, C_{15}+C_{19} and three multi-component mixtures. Both literature and measured data are in atmospheric pressure. Unfortunately, no further details on experiments were provided.

Milhet et al. (2005) used exactly the same laboratory setup as reported by Daridon et al. (2002), a high-pressure cross-polar microscope, applying same procedure as outlined earlier: heating the sample in 0.1–0.2°K every every 5–10 minutes and studied WDT in the range of pressures up to 100 MPa in an increment of 20 MPa. Additionally, Milhet et al. (2005) compared their laboratory data to previous published results. For C_{14}, C_{15}, C_{16} pure components, readings were compared with data reported in 8 different publications for various pressures up to 100 MPa and deviation was found within 1.2°K. For C_{14}+C_{16} binary system at atmospheric pressure Milhet et al. (2005) data is compared with those obtained by Parczewska (1998), who used kinetic and dilatometric measurements with results matching and to Rajabalee (1995) test results from DSC and slight discrepancy was reported, Milhet et al. (2005) data reading higher. Milhet et al. (2005) further claim visual microscopy data being more accurate referring to works by Ji et al. (2003) and Ronningsen et al. (1991). For C_{14}+C_{15} binary system data was compared with 3 different sources and similar a discrepancy is noted.

Rizzo et al. (2007) designed a new device extending features of high-pressure sapphire cell used in cross-polarizing microscope with optical detection system to reduce human error. Change in reflected, retracted light intensity helps to define phase transition. Device temperature is regulated within 0.1°K and pressure measurement tolerance is within 0.2 MPa. WDT readings obtained for synthetic mixtures and diesel were compared with studies conducted earlier and published in literature that use HP CPLM.

Mansourpoor et al. (2019) used earlier referenced laboratory data from Metivaud et al. (1999), Daridon et al. (2002), Ji et al. (2003), and Milhet et al. (2005). Additionally, Mansourpoor et al. (2019) have analyzed 9 Iranian oil samples using both viscometry and DSC methods. It is worth to note that Mansourpoor et al. (2019) have not directly measured WDT but instead measured WAT and estimated WDT by assuming WDT higher than WAT by 3°C based on studies performed by Ronningsen et al. (1991) (+/- 3°C) and Nitin and Anil (2004) (3.2+/-0.6°C).

Shariatrad et al. (2022) reported WDT measurement of 30 mixtures of C_{11}+C_{14}+C_{16}+C_{18} quaternary system. In their experiments they have noted an increase in WDT with an increase in weight fractions of heavier components in the mixture. Measurement device constitutes of stainless-steel cell with glass window for observation, built-in stirrer for equilibrium establishment with turbulence, system cooling and heating is provided with ethanol bath into which the test cell is subjected. Thermometer provides tolerance of 0.1°K. For test procedure equilibrium step heating is adhered. Shariatrad et al. (2022) suggested visual observation is within the acceptable accuracy range referring to previous works conducted on this comparison such as Parsa et al. (2015) claiming visual observed WDT being in line with CPM within 0.2°K. As for laboratory procedure, sample is placed inside cell and system temperature is reduced to 10°K below the estimated WAT temperature. Once the wax crystals appear, the system is subjected to step heating with heating rate of 0.1°K/hour.

2.1.3 WDT Laboratory Measurement Conclusive Remarks

A critical review of the laboratory works reported in the literature shows that despite most of earlier studies being based on WAT measurements for wax characterization, WDT is indeed a preferred parameter, as WAT does not represent solid-liquid equilibrium point and is strongly affected by cooling rate and was reported to deviate from true SLE point as much as 28°C. As

such for WDT modeling WAT studies should not be used. Mansourpoor et al. (2019) recommended that the "estimated WDT" from WAT should be used with caution. The authors have also highlighted testing procedure importance for accurate WDT measurement, equilibrium step heating is recommended, and most accurate studies have controlling heat rate at 0.1–0.2°K/5–10 min.

High pressure capability of testing apparatus is critical for proper wax formation characterization to analyze WDT dependency on system pressure. Three out of nine studies reviewed above offer WDT data in high pressures up to 150 MPa. Not all devices are readily available with pressuring capacity. Hence those that provide one should prefer for test procedure design. Juyal et al, 2011 noted that pressure might play a detrimental role in reservoir fluids with high gas/oil ratio (GOR). WAT and consequently WDT would increase with higher GOR. They have tested WAT at different time lag, longer time allowing more gas to dissolve and noted, longer they waited closer WAT data was recorded to model WAT. Which raises another important point that experiments need to be performed at the conditions imitating operating conditions.

While reviewing various WDT measurement methods in the literature, Rizzo et al, 2007 suggest viscosity, acoustic, and FTIR spectroscopy have a limited capacity in detecting last crystal disappearance for WDT because of very limited number of solids contained in the system just below WDT, size of the solid crystals below 0.1 μm and high opacity of the system. From measurement methods listed in Table 2.1, the following methods were used in the reviewed published works: Differential Scanning Calorimetry (DSC), Cross-Polarized Light Microscopy (CPLM), CPLM with high pressure cell and Viscometry. Among the listed methods, DSC and CPLM were found most popular. For DSC vs CPLM comparison as Rizzo et al have noted that opinions differ: some researchers argue DSC method provides better

results than other methods (Kok et al, 1996); while others claim DSC providing less accurate results cross-polarizing microscope (Milhet et al., 2005 and Coutinho et al., 2005). Both CPLM and DSC can be custom built with high pressure cell option.

2.2 WDT Thermodynamic Modeling

Back in 1980's there was a limited number of research works published on wax modeling (Won, 1989). Very few laboratory data were reported including cloud point temperatures and *n*-paraffin concentrations by Nikolaeva et al. (1976) without molecular weight data and cloud point measurements and molecular weights by Hansen et al. (1988) with limited information on wax content. On theoretical calculations front, works were either limited to certain conditions such as narrow temperature range below CP or did not account for multi-component wax phases. In the text below, some early fundamental works on hydrocarbon waxes applied in liquid-solid and vapor-liquid-solid phases are listed. It should be noted here that not all the thermodynamic models are analyzed in scope of this work but rather those that are most relevant and are considered as seminal research works in this area.

Won (1986) published a seminar paper that is often referenced by most of the papers as fundamental study to guide with thermodynamic calculation for wax build-up phenomena. There was a previous work published by the same author but limited to liquid-solid equilibria modeling of hydrocarbon mixtures based on ideal solution approach (Won, 1985). The scope of work in the subject work was extended to include all three phases vapor-liquid-solid equilibria for wax in paraffinic hydrocarbon mixtures. Won (1986) has applied modified regular solution method to model liquid-solid phase equilibria; while, using SRK-EOS for vapor-liquid phase equilibria.

Equilibria is achieved in fluid for component *i* when fugacity of the component in solid, liquid, and gas phase are equal. Solid-liquid equilibrium coefficient, K_i, a ratio of mole fraction of component *i* in solid over to liquid phase, is defined as function of activity coefficient, melting temperature, heat capacity, and volume change of fusion. Latter two terms in exponential function were neglected. In study performed by Won in 1985, the ratio of activity

coefficients was considered 1, but this would result in C_5-C_{10} solubilities in solid solution to be over-estimated. Won (1986) has expressed activity coefficient ratio by modified regular solution theory as a function of molar volume of component *i*, difference between solubility parameter of the mixture (average) and component *i*, and volume fraction of component *i*. Liquid solubility parameter is modeled as a function of melting enthalpy, temperature, and volume. Solid solubility parameter is calculated based on modified cohesive energy correlation, as a function of phase change enthalpies, temperature, and volume. Melting temperature, melting enthalpy, and molar volume are calculated using the following correlations and expressed as a function of MW:

$$T^f = 374.5 + 0.02617 * MW - \frac{20172}{MW}$$ Equation 2.1 (Won, 1986)

$$\Delta H^f = 0.1426 * MW * T^f$$ Equation 2.2 (Won, 1986)

$$v = \frac{MW}{0.8155 + 0.6272 * 10^{-4} * MW - \frac{13.06}{MW}}$$ Equation 2.3 (Won, 1986)

Vapor-Liquid equilibrium is estimated using Soave modified Redlich-Kwong EOS, where vapor-liquid equilibrium coefficient is expressed as ratio of volume fraction coefficients in liquid to vapor phases.

Hansen et al. (1988) have tested and reported WAP for 17 stabilized crude oil samples from the North Sea. They have compared the resultant WAP values with Won (1986) modeling results and found Won (1986) to overestimate WAP values. For 17 samples they reported average deviation of 17% or 53°K (max 26% or 74°K). This led them suggesting corrections to Won, 1986 model, by lowering activity coefficients, another alternative would be reducing melting enthalpies. Hansen et al. (1988) suggest that Won (1986) relate melting temperature of *n*-paraffins to carbon number. However, in crude oil *n*-paraffins content reduction is observed

with increase in carbon number. Hence, they have proposed the following revision to melting temperature; while, keeping melting enthalpy calculation same as proposed by Won (1986):

$$T^f = 402.4 - 0.01896 * MW - \frac{21709}{MW}$$ Equation 2.4 (Hansen et al., 1988)

Hansen et al. (1988) proposed a different method to model liquid phase by using polymer solution theory of Flory. They also suggested to treat solid phase as ideal mixture and activity coefficient for solid equal to 1. With above adjustments, the newly calculated WAP for 17 crude oil samples were reported within 2% (6.35°K) deviation if consider Gibbs excess energy from wax phase and 3% (11.35°K) if the Gibbs excess energy is neglected. Hansen et al. (1988) concluded that while they were able to get impressive results with WAP calculations they were not able to test wax precipitation curves they have developed because of limited test data.

Won (1989) has revisited his previous 1986 model by bringing back the heat capacity term that was ignored in the 1986 model, modeling solid phase as ideal mixture, and further evolved liquid phase activity coefficient calculation. Won has run detailed study of temperature on liquid activity coefficient behavior and revealed that behavior of both "a thermal solution" (activity coefficient is less than 1) and "regular solution" noted (activity coefficient value reduction with temperature rise). Hence, he suggested to model liquid activity coefficient as sum of both of above effects:

$$\ln \gamma = ln\gamma^A + ln\gamma^R$$ Equation 2.5 (Won, 1989)

The regular solution is modeled same as reported by Won (1986). A thermal solution is suggested to model as function of ratio of size parameters which can be related to ratio of molar volumes in power of N, where N is in range 0–1. Six different methods were compared: 1) ideal solubility, 2) liquid volume V^L, 3) $(V^L)^{2/3}$, 4) UNIFAC Van der Waals volume, $(V^*)^{2/3}$,

5) Hildebrand molar volume $VH^{0.7}$, and 6) $VH^{2/3}$. The sixth method was selected as the most accurate for solubility estimation.

UNIFAC: UNIQUAC Functional-group Activity Coefficients

UNIQUAC: UNIversal QUAsi-Chemical model

The research work reported by Hansen et al. (1988), was limited to WAP calculation and could not study wax precipitation as a function temperature in details because of limited laboratory data. With new laboratory data reported by Pedersen et al. (1991a) for North Sea crude oil, Pedersen et al. (1991a) were able to revisit their published work to update the thermodynamic models proposed by Won (1986) and Hansen et al. (1988) both for WAP and wax precipitation, focusing on the latter. Comparing the new laboratory data of wax precipitation with these two models (Won, 1986 and Hansen et al., 1988) they have demonstrated that the calculated results were way out of acceptable range of deviation, hence a need for revision of existing models.

As basis for their work Pedersen et al. (1991a) have selected Won (1986) model, which applies regular solution theory both to liquid and solid phase, whereas model proposed by Hansen et al. (1988), which is based on polymer solution theory is seen more complex and not adding much accuracy to the results compared to Won (1986). The modification proposed to Won (1986) model was to increase solubility parameters by following correlations for liquid-phase solubility, δ_i^L and solid-phase solubility, δ_i^S:

$$\delta_i^L = 7.41 + \alpha_1(lnC_N - ln7) \qquad \text{Equation 2.6 (Pedersen et al., 1991a)}$$

$$\delta_i^S = 8.50 + \alpha_2(lnC_N - ln7) \qquad \text{Equation 2.7 (Pedersen et al., 1991a)}$$

Moreover, solubilities for naphthene and aromatics are simply increased by 20% than those for paraffins. For melting enthalpy, Pedersen et al. (1991a) suggest increasing enthalpy

values of pure *n*-paraffins by coefficient $\alpha_3 = 0.5148$. Similar as above, the melting enthalpy for naphthene and aromatics are simply increased by 50% than those for paraffins. Additionally, Pedersen et al. (1991a) suggest heat capacities to be included in calculations and propose to estimate heat capacity difference as function of molecular weight and temperature, ΔC_{pi}:

$$\Delta C_{pi} = \alpha_4 MW_i + \alpha_5 MW_i T \qquad \text{Equation 2.8 (Pedersen et al., 1991a)}$$

Coefficients are calculated applying least-squares fit to laboratory data, as following: $\alpha_1 = 0.5914$ $(cal/cm^3)^{0.5}$, $\alpha_2 = 5.763$ $(cal/cm^3)^{0.5}$, $\alpha_3 = 0.5148$, $\alpha_4 = 0.3033$ cal/(g×K), $\alpha_5 = -4.635 \times 10^{-4}$ cal/(g×K^2).

This study demonstrates notable improvements in wax amount estimates. But our parameter of interest is WAP. Pedersen et al. (1991a) have compared laboratory WAP and calculated value from new model and reported Average Absolute Deviation (AAD) of 10.19°C (or 29%), while maximum deviation is 25°C (or 71%), which are very high.

Pedersen (1993) reported assessing the existing methods that model three phases vapor-liquid-solid equilibria to study wax precipitation. The author disagrees with some of previous research where vapor-liquid and liquid-solid equilibria were modeled with different methods. Vapor-liquid equilibria historically is modeled using cubic equation of state (EOS), either Soave-Redlick-Kwong (SRK) or Peng-Robinson (PR) EOS. To model liquid-solid equilibria Won (1986 and 1989) and Pedersen et al. (1991a) applied regular solution theory and Hansen et al. (1988) applied polymer solution theory. Pedersen (1993) suggested that using two different methods for modeling same liquid phase, conditional to vapor-liquid or liquid-solid equilibria is thermodynamically inconsistent and might lead to convergence issues. Hence, Pedersen (1993) proposed to use same method for both vapor-liquid and liquid-solid equilibria

modeling: cubic EOS. While modeling vapor-liquid equilibria using EOS is straight forward, for modeling liquid-solid equilibria, Pedersen (1993) proposes an algorithm to be used instead.

Liquid fugacity is expressed as a function of liquid phase mole fraction, fugacity coefficient, and system pressure. For solid phase, ideal solid phase mixture is assumed, and solid phase fugacity is defined as product of solid phase mole fraction and solid standard state fugacity. Solid phase standard state fugacity is related to liquid phase standard state fugacity by molar change in Gibbs free energy, which can be defined as difference between change in enthalpy and product of temperature and change in entropy. Liquid phase standard state fugacity is expressed as a function of liquid phase fugacity coefficient of pure component and system pressure. Fugacity coefficient of pure component is determined using cubic EOS.

Another point Pedersen (1993) disagrees on with previous studies is defining components that will precipitate in wax. Won (1986 and 1989), Pedersen et al. (1991a) and Hansen et al. (1988) works suggest all hydrocarbon (HC) components form wax. While Pedersen (1993) agrees that *n*-paraffins may fall in wax phase indifferent of their molecular weight, other components may not necessarily form wax phase, specifically aromatics will likely fall in asphaltenes, but not wax, despite very heavy MW. Presence of other heavy paraffin and cyclo-paraffin compound is unlikely because of structural challenge. Hence, the most occurred components to form wax would be *n*-paraffins, iso-paraffins, and naphthene. Pedersen (1993) suggested that assumption of all HC compounds forming wax in previous works of Won (1986 and 1989), Pedersen et al. (1991a) and Hansen et al. (1988) might have led to over-estimation of WAP and wax amount precipitation. Hansen et al. (1988) attempted to adjust the numbers to fit laboratory results by reducing the liquid phase activity coefficients, which the Pedersen (1993) sees as fundamentally not justified and Pedersen et al. (1991a) modeled ideal liquid phase and non-ideal solid phase. Pedersen (1993) suggests a fundamentally correct way

to proceed by modeling only those components that will form wax. However, this is not practical, as detailed Paraffins, Naphthenes, and Aromatics (PNA) analysis is not common. Hence, Pedersen (1993) suggested an empirical approach of defining components that will likely form wax. The selection of components is based on approach proposed by Pedersen et al. (1991a), which covers a range of carbon number fractions up to C_{80}. Further, Pedersen (1993) proposed use of expressions of solid phase mole fraction as a function of molecular weight and density at standard conditions of component and normal paraffin, related by coefficients, A, B and C. Density of normal paraffin at standard conditions is also expressed as function of molecular weight. Coefficients A, B, and C are found by tuning results to laboratory data reported by Ronningsen et al. (1991) and Pedersen et al. (1991a). Calculation results compared with laboratory data, 16 WAP measurements, reported by Pedersen et al. (1991b) demonstrate AAD of 2.63°C (or 8%); while the maximum deviation is 8°C (or 36%) which is much lower than reported by previous study, Pedersen et al. (1991a).

Previous work by Pedersen (1993) suggested that selected components only precipitate in solid wax phase, and this explained why all models prior Pedersen (1993) were overestimating the amount of wax precipitation. Analyzing laboratory data from Pedersen et al. (1991a) and more recent studies from Snyder et al. (1992, 1993 and 1994) that summarize the observations that solid wax phase mostly constitute of pure components, Lira-Galeana et al. (1996) have come to conclusion that only mutually immiscible pure, (pseudo) components precipitate in solid wax phase. Lira-Galeana et al. (1996) proposed to use stability analysis to define which components will precipitate in pure solid form, which is based on stability condition obtained from Michelsen (1982): if the fugacity of specific component is equal to or higher than solid phase fugacity of that component in pure form, then the component will precipitate. Lira-Galeana et al. (1996) have selected PR-EOS to calculate fugacity of

component in mixture. Solid phase fugacity of pure component is calculated by relating to liquid phase fugacity of pure component as a function of melting temperature, melting enthalpy, change in heat capacity, and the system temperature. Liquid phase fugacity of pure component is calculated as product of fugacity coefficient of pure component and system pressure.

Lira-Galeana et al. (1996) proposed a correlation based on laboratory measurements of C_6-C_{30} *n*-paraffins, naphthene/alkylcycloalkanes, and aromatic components/alkylbenzenes from API RP 44, 1964:

$$T_i^f = 333.46 - 419.01\exp(-0.008546MW_i)$$ Equation 2.9 (Lira-Galeana et al., 1996)

Lira-Galeana et al. (1996) also suggested to use enthalpy reduction similar to Pedersen et al. (1991a) as opposed to Won (1986) model:

$$\Delta H_i^f = 0.05276MW_iT_i^f$$ Equation 2.10 (Lira-Galeana et al., 1996)

Pedersen et al. (1991a) model was used to calculate the heat capacity of fusion:

$$\Delta Cp_i = 0.3033MW_i - 4.635x10^{-4}MW_iT_i^f$$ Equation 2.11 (Lira-Galeana et al., 1996)

Lira-Galeana et al. (1996) have compared their modeling results with laboratory data from Pedersen et al. (1991b) for 8 oil samples. Wax precipitation amount was in good match with laboratory data and calculated WAP too was found within average deviation of 2.14°K (or 0.7%); while the maximum deviation is 4.15°K (or 1.4%) which is significantly lower than previous reported models.

Coutinho et al. (1995) categorized all of the previous works reported on modeling wax appearance under following groups: ideal liquid phase (Won, 1985), regular solution theory adaptations (Won, 1986 and Pedersen et al., 1991a), Flory-Huggins model for activity

coefficients (Hansen et al., 1988). Coutinho et al. (1995) find most of these studies as essentially empirical correlations linking WAT to fluid properties and they argue that methods that lack fundamental basis and are developed by simply matching to laboratory data will be limited in predictive capacity. Coutinho et al. (1995) further are set to develop a fundamentally strong thermodynamic model for WAT estimation and the scope of subject work is limited to evaluation and comparison of existing activity coefficient-based models such as Flory-Huggins, modified UNIFAC and free volume models. Coutinho et al. (1995) suggest that all studies modeling solid-liquid equilibrium at low pressures originate from equating liquid and solid phase fugacity and can be expressed with generic form:

$$ln\frac{f^{ol}}{f^{os}} = \frac{\Delta h_m}{RT_m}\left(\frac{T_m}{T}\right) - 1 + \frac{\Delta h_t}{RT_t}\left(\frac{T_t}{T} - 1\right) - \frac{\Delta C_{pm}}{R}\left(\ln\frac{T}{T_m} + \frac{T_m}{T} - 1\right)$$ Equation 2.12 (Coutinho et al., 1995)

The differences between the abovementioned studies are because of the assumptions made to simplify the calculations, methods to estimate fusion, and transition enthalpies, temperatures, and heat capacity, and, most importantly, activity coefficients modeling approaches. Generalized expression for activity coefficient will be a product of combinatorial factor (to address variances in sizes and shapes of molecules), residual factor (to address energetic interactions) and free-volume factor (address major size differences between components in mixtures). In hydrocarbons, some of these factors can be ignored. Hence, the differences in modeling approaches: GCFLORY account for all 3 factors, Entropic free-volume and Flory free-volume ignores residual factor, Modified UNIFAC ignores free-volume, Flory-Huggins ignores both residual and free-volume factors. Coutinho et al. (1995) run different models and compare the results with laboratory data gathered from various sources, 60 binary compounds with more than 1,000 data points.

Interesting observations made from these comparisons. The models that exclude free-volume factor have accuracy that deteriorates with increasing carbon number, AAD is less than 0.7% for C_{20} and lower. The average AAD for Ideal solution: 0.628%, the Flory-Huggins: 1.048%, and for the Modified UNIFAC: 0.452%. Models that consider free-volume factor AAD, show less to no dependency on carbon number. These models have lower AAD than those that exclude free-volume factor. The average AAD for GCFLORY: 0.35%, Flory-FV: 0.33-0.695%, Entropic-FV: 0.338-0.683%. Flory-Huggins model showed the worst results, performance is lower than that of the ideal solution. Hence this model and studies that were based on this model are not recommended, such as Hansen et al. (1988). GCFLORY model despite strong fundamental basis, incorporating all three factors, yet does not display incremental value compared to Flory-FV and Entropic-FV models, which can be explained by limited impact of the residual factor.

Regular solution theory (RST) based models have not been included in detailed analysis by Coutinho et al. (1995), because of authors' disagreement with use of theory for wax modeling. RST relies on enthalpic changes and excludes effect of entropy changes. Coutinho et al. (1995) claim that for alkanes phase non-ideality result from entropic changes. Hence, they suggested that studies based on RST such as Won (1985, 1986), Pedersen et al. (1991a) will perform worse than the ideal solution model. Extension of the model to incorporate Flory-Huggins activity coefficient approach such as Won (1989) model is also criticized by Coutinho et al. (1995) that expected lower accuracy. Often models neglect heat capacity term, Coutinho et al. (1995) have analyzed the impact of excluding heat capacity term using 3 models and found difference in AAD within 0.024% (GCFLORY: 0.004%, Entropic-FV: 0.01%, Flory-FV: 0.024%). Another neglected term is solid phase transition, same 3 models were compared and have shown different results, GCFLORY: 0.027%, Flory-FV: 0.027%, Entropic-FV:

0.13%. Coutinho et al. (1995) reemphasize that these models should only be used at atmospheric-low pressures and in high pressure systems EOS should be used over the activity coefficient model.

2.3 WDT Empirical Methods

Different empirical studies have been conducted for WDT determination. Researchers have performed WDT measurements for various oil compositions and compared the results with the published thermodynamic models. Some researchers have developed their own thermodynamic models and compared the modeling results with laboratory data. Additionally, some of the researchers have extended their studies analyzing the effect of system pressure on WDT and compared thermodynamic modeling results with laboratory data. The most notable empirical studies performed on WDT determination are summarized in Table 2.3.

Table 2.3. Summary of empirical studies performed on WDT determination.

Empirical Model	Sample Composition	Pressure Range	Thermodynamic models compared	Results
Pauly et al. (1998)	3 synthetic mixures of C_{10} + heavy normal paraffins C_{18}-C_{30}	0.1 MPa	1) Ideal solution model, 2) Won (1986) model with regular solution, 3) Hansen et al. (1988) model based on Flory's generalized polymer solution for liquid phase, and solid phase is modeled ideal, 4) Pedersen et al. (1991a) based on Won (1986) model w/ heat capacity, 5) Coutinho et al. (1995, 1996) model: liquid phase non-ideality is	Deviation from WAT laboratory: 1) Ideal model: 4°K to 7°K 2) Won (1986): 4.1°K to 7.2°K 3) Hansen et al. (1988): 2°K to 5°K 4) Pedersen et al. (1991a): 2.2°K to 5.1°K 5) Coutinho et al. (1995, 1996): 0.6°K to 1°K 6) Ungerer et al. (1995): -0.1 to -3.1°K

			handled by free-volume model, solid state is defined by Wilson local composition model, 6) Ungerer et al. (1995), solid phase model for each component as pure, liquid phase based on PR-EOS with original mixing rules	Most accurate model -Coutinho et al. (1995, 1996), max dev 1°K
Metivaud et al. (1999)	ternary systems of n-alkanes: C_{14}+C_{15}+C_{16}, C_{16}+C_{17}+C_{18}, C_{18}+C_{19}+C_{20}, C_{19}+C_{20}+C_{21}	0.1 MPa	Binary systems: LIQFIT is used for 11 binary systems, Ternary systems: Txy-CALC is used for 11 ternary system compositions, Gibbs energy was handled in 2 ways: temperature independent and dependent.	The maximum average difference being 0.3°K
Dauphin et al. (1999)	5 systems of C_{10}+*n*-alkanes C_{18}-C_{36}	0.1 MPa	For solid phase, Coutinho's solid solution model w/ 2 options for modeling non-ideality in solid phase modeling: Wilson local composition model and UNIQUAC. For liquid phase non-ideality, Flory-free volume model is used.	Deviation from laboratory WAT: 1) Wilson equation: up to - 0.7°K 2) UNIQUAC: up to - 2.1°K For overall percentage of paraffins crystalized: graphical comparison shows UNIQUAC more accurate description of bimodal systems.
Vafaie-Sefti et al. (2000)	3 oil mixtures C_1-C_{20}, 3 oil mixtures C_{10}-C_{50} (w/ split for PNA)	Exp: 0.1 MPa Model: 0.1-100-	New model is developed - modified multi-solid phase model: Vapor and Liquid phase by PR-EOS, solid phase modeled either with average parameters method or PNA analysis.	Average absolute error from Laboratory WAT for oil samples 1-3, 1) Lira-Galeana et al. (1996): 0.09%,

		200-300 MPa	Existing models are compared: Oil samples1-3 with average parameters method compared with 1) Lira-Galeana et al. (1996), 2) Pedersen et al. (1991a), 3) Pedersen (1993), 4) Won (1986). Oil samples 4-7 with PNA analysis compared with Pan et al. (1997).	2) Pedersen et al. (1991a): 7.77%, 3) Pedersen (1993): 1.2%, 4) Won (1986): 13.93%, Max deviation of Vafaie-Sefti et al. (2000) model from Pan et al. (1997) model for oil samples 4-7: • Atmospheric: 0.42%, • 100 MPa: 0.5%, • 200 MPa: 0.55%, • 300 MPa: 0.41%
Pauly et al. (2000)	Light gas + Heavy HC 10 Binary systems, 5 Ternary systems, 4 mixtures of C_1+C_{10}+$C_{18\text{-}30}$	Up to 200 MPa	Method is developed based on LCVM model for describing fluid-fluid, fluid-solid equilibria at high pressures. Cubic EOS cannot be used as does not differentiate fugacitiies at equilibria when solid phase is introduced, excess Gibbs energy model is used instead coupled with EOS, various EOS were reviewed and LCVM is selected.	Absolute average deviation: Binary systems: C_1+$C_{16,20,22,24}$, AAD 1.13-2.02 C_2+$C_{16,28,20,22,23}$, AAD 1.04-3.07 C_3+C_{34}, AAD 2.07 Ternary systems: C_1+C_3+C_{24}, AAD 1.83 C_1+C_{10}+$C_{22,32}$, AAD: 0.34-2.09 C_2+C_{10}+C_{32}, AAD: 2.58 C_1+C_{22}+C_{24}, AAD: 2.14 4 mixtures of C_1+C_{10}+$C_{18\text{-}30}$, AAD: 0.14-1.05
Pauly et al. (2004)	5 synthetic mixtures of C_{10} + (C_{24}-C_{25}-C_{26})	0.1 MPa	Cautinho et al. (1996), based on Flory free-volume equation, is considered limited to low pressures, Pauly et al. (2000): Soave cubic EOS to extend to high pressures,	Deviation from Laboratory WAT for 5 samples: 1) Coutinho et al. (1996): within 0.5°K,

			with LCVM mixing rule with UNIFAC group contribution mode.	2) Pauly et al. (2000): within 0.8°K. Absolute average deviation % from laboratory solid deposit quantity for 5 mixtures: 1) Coutinho et al. (1996): within 9.1%, 2) Pauly et al. (2000): within 10%.
Milhet et al. (2005)	14 compositions of C_{14}+C_{16}, 10 compositions of C_{14}+C_{15}	up to 100 MPa	Liquid phase: PR-EOS, Solid phases: Gibbs energy models, for modeling activity coefficients of Rotator phase Chain Delta Lattice Parameter model and Wilson equation results were compared with laboratory data, both models gave similar accuracies by CDLP was selected due to predictability, for Triclinic phase, Margules model was used.	Average absolute deviations: Rotator phase: CDLP: C_{14}+C_{16} = 0.4°K, C_{14}+C_{15} = 0.3°K, Wilson: C_{14}+C_{16} = 0.2°K, C_{14}+C_{15} = 0.3°K, Triclinic phase - Morgules method: C_{14}+C_{16} = 0.2°K, C_{14}+C_{15} = 0.4°K

LCVM: Linear Combination of Vidal and Michelsen models

The empirical studies presented in Table 2.3 additionally provide the following insights:

Pauly et al. (1998) compared the most fundamental and commonly known thermodynamic models with laboratory data obtained from synthetic mixtures of decane and heavy hydrocarbon chain C_{18}-C_{30} at atmospheric pressure: The ideal solution model, Won model of regular solution (1986), Hansen et al model (1988), Pedersen et al. (1991a), Coutinho et al. (1995, 1996), and Ungerer et al. (1995) models. Of all these reviewed Coutinho et al.

(1995, 1996) model with Wilson local composition equation performs better than all listed with deviation from laboratory data within 1°K.

Metivaud et al. (1999) evaluated four ternary systems of consecutive $_n$-alkanes at atmospheric pressure. Their work distinguishes from others with in-depth study of crystal structures of $_n$-alkanes. They suggested while odd-numbered alkanes with orthorhombic crystal structure (Oi) at low temperatures take Rotator I (RI) form at higher temperatures, even-numbered alkanes are defined with Triclinic structure (Tp) with RI form characterized as metastable. Binary systems solid-liquid phase equilibria are modeled with LIQFIT program that evaluates excess Gibbs energy. Ternary systems are modeled with Txy-CALC program that is also based on excess Gibbs energy estimates.

Dauphin et al. (1999) investigated five systems of solvent (decane) and heavy hydrocarbon chain C_{18}-C_{36} at atmospheric pressure. Dauphin et al. (1999) support use of Cautinho et al. (1997) model with Wilson equation to characterize solid phase for narrow paraffin distribution window. However, for wider paraffin distribution they suggest use of UNIQUAC model instead of Wilson equation as proposed by Cautinho et al. (1998). For liquid phase Flory free volume model is used. The results showed that the Wilson model's accuracy for WAT estimation and the UNIQUAC's preference for solid deposit calculation.

Vafaie-Sefti et al. (2000) challenged use of multi-solid models as only liquid and solid phases are evaluated and introducing vapor phase. They suggested impact of pressure and composition on WAT is not taken into consideration and suggested use of models that consider all three phases. They used PR-EOS for vapor-liquid equilibria characterization. For liquid-solid equilibria they have evaluated two approaches: average properties method and PNA analysis. They have compared their work with the following thermodynamic models: Lira-Galeana et al. (1996), Pedersen et al. (1991a), Pedersen (1993), Won (1986) and laboratory

data at atmospheric pressure and found Lira-Galeana et al. (1996) demonstrating outstanding performance with AAD of 0.09%. There is no laboratory data available for higher system pressure, instead Vafaie-Sefti et al. (2000) model is compared with Pan et al. (1997), discrepancies falling within 0.55%. They concluded that the model proposed by Pan et al. (1997), validates Vafaie-Sefti et al. (2000) method.

Pauly et al. (2000) took into consideration the impact of presence of light gas. Binary and ternary systems and mixtures where solvent is present by methane, ethane, propane, and decane and heavy HC chain covers range of C_{16}-C_{34}. They noted that high system pressure at presence of light gas improves heavy HC solubility in liquid, that may reduce at some cases WAT by 15°K when pressure is increased from atmospheric to saturation pressure. Pauly et al. (2000) have developed their own model which is based on LCVM calculation to characterize liquid-solid equilibria at high pressures. Their work shows AAD within 3.07°K for binary systems, 2.58°K for ternary systems and 1.05°K for mixtures.

Pauly et al. (2004) work on 5 systems of decane solvent and heavy HC chain of C_{24}-C_{25}-C_{26} at atmospheric pressure evaluates Coutinho et al. (1996) s model versus Pauly et al (2004) model. Both models present agreeable results; while Coutinho et al. (1996) model accuracy prevails.

Milhet et al. (2005) have proposed a model that is based on PR-EOS for liquid phase characterization and the excess Gibbs energy models for solid phase description. They compared CDLP and Wilson calculation for activity coefficients modeling for Rotator phase and concluded that both showing similar performance, AAD: 0.2-0.4°K (Wilson marginally better, AAD: 0.2-0.3°K) but suggested use of CDLP for predictable results. Most models will not cover Triclinic phase and Milhet et al. (2005) suggested use of Morgules method (AAD within 0.4°K).

2.4 WDT Intelligent Modeling

Six available studies reported in the literature were reviewed on ML-based data modeling to determine WDT. Main highlights from these works are provided in Table 2.4, where these studies are summarized. In the consequent sections, a discussion of datasets, models, and the optimizing techniques used along with performance metrics are presented.

Table 2.4. Summary of intelligent ML-based data modeling studies on WDT prediction.

Reference	Data source	Dataset size	Model used	Accuracy
JPSE - July 2013 - Moradi et al - Prediction of wax disappearance temperature using artificial neural networks	Robbles et al. (1996) - 8 Metivaud et al. (1999) - 56 Daridon et al. (2002) - 54 Ji et al. (2004) - 58 Milhet et al. (2005) - 130 Vafaie-Sefti et al. (2007)	306	Two-layer network with various hidden neurons was examined and multiple optimizing algorithms were tested: LM, SCG, GDA, BR. Network w/ 16 hidden neurons and LM optimizer was selected.	MRE = 0.38%
Energy & Fuels - February 2019 - Bian et al - Prediction of Wax Disappearance Temperature by Intelligent Models	Milhet et al. (2005) - 144 Ji et al. (2004) - 74 Daridon et al, 2002 – 54	272	GWO-SVM, LS-SVM, GA-AN-FIS, PSO-AN-FIS. GWO-SVM is selected.	GVO-SVM: AARD = 0.7128%, SD = 0.0083, RMSE = 2.4208, R^2 = 0.9546, EP_{max} =

				1.7279%, EP_{min}= 0.0178%
Petroleum Science and Technology - February 2019 - Kamari et al - Evaluation of wax disappearance temperatures in hydrocarbon fluids using soft computing approaches	Robbles et al. (1996) Metivaud et al. (1999) Vafaie-Sefti et al. (2000) Daridon et al. (2002) Ji et al. (2004) Milhet et al. (2005)	254	ANN, LSSVM (least square support vector machine), DT (decision tree). DT is selected.	DT: AAPRE = 0.3%, APRE = -0.002%, SD = 0.003, RMSE = 1.5, R^2 = 0.97;
Energy & Fuels - October 2019 - Benamara et al - Modeling Wax Disappearance Temperature Using Advanced Intelligent Frameworks	Milhet et al. (2005) - 144 Ji et al. (2004) - 74 Daridon et al. (2002) - 54	272	RBFNN-GA, RBFNN-ABC, GMDH (RBFNN: radial basis function neural network, GA: genetic algorithm, ABC: artificial bee colony, GMDH: Group method of data handling)	RBFNN-ABC: AARD = 0.5402%, R^2 = 0.9706, RMSE = 1.9969, SD = 4.69×10^{-5}

	Metivaud et al. (1999) Vafaie-Sefti et al. (2000) Daridon et al. (2002) Ji et al. (2004) Milhet et al. (2005) Rizzo et al. (2007) Moradi et al. (2013) Behbahani et al. (2016) Mansourpoor et al. (2019a, 2019b) Kamari et al. (2019) Bian et al. (2019) Benamara et al. (2019)		RBF, MLP-LM/BR, ANFIS-CA/BBO/TLBO, DT, RF, ET, (RBF: radial basis function, MLP: multilayer perceptron, ANFIS: adaptive neuro-fuzzy interference system, CA: cultural algorithm, BBO: biogeography-based optimization, TLBO: teaching-learning-based-optimization, DT: decision tree, RF: random forest, ET: extra tree)	
JPSE - November 2021 - Amiri-Ramsheh et al - Modeling of wax disappearance temperature (WDT) using soft computing approaches: Tree-based models and hybrid models		346		RF has best results: APRE = 0; AAPRE = 0.246%; RMSE = 1.01; SD = 0.003; R^2 = 0.99; Accuracy: RF > DT > MLP-BR > RBF > ANFIS-BBO > MLP-LM > ANFIS-CA > ET > ANFIS-TLBO;
Fuel - August 2024 - Nait Amar et al - Modeling wax disappearance temperature using robust	Milhet et al. (2005) - 144 Ji et al. (2004) - 74	272	GEP (gene expression programming)	R^2 = 0.9647; RMSE = 2.1963; AARD = 0.5963%;

white-box machine learning	Daridon et al. (2002) - 54

Moradi et al. (2013) investigated two studies reported on thermodynamic modeling by Ghanaei et al. (2007) and by Ji et al. (2004). Ghanaei et al. (2007) have classified the existing models in 4 types (Model 1, 2, 3 without k_{ij} and Model 3 with k_{ij}) and have proposed their own model (New model). These models have been tested with laboratory data from literature reported by Milhet et al. (2005) data for C_{14}-C_{15} and C_{14}-C_{16} binary systems. Ji et al. (2004) investigated the ideal solution model, multi-pure-solid model, and Coutinho's UNIQUAC and have developed their own HWWAX model. They too have tested these models with published data reported by Robles et al. (1996) for binary systems C_{17}–C_{19}, Metivaud et al. (2002) for ternary systems C_{14}–C_{21} and against their own laboratory data, Ji et al. (2004) for various mixtures C_7–C_{36}.

Moradi et al. (2013) have run their newly developed ANN model separately for each dataset referenced above and compared their results against laboratory data and calculated values by Ghanaei et al. (2007) and Ji et al. (2004). This comparison is outlined in Table 2.3. Their ANN model has shown better results than Ji et al. (2004) referenced thermodynamic models and own developed model HWWAX and has shown acceptable results compared with Ghanaie et al. (2007) work. Moradi et al. (2013) have reviewed 18 studies reported from 1990 through 2011 and demonstrated how ANN application was adopted in different fields from pH response modeling to bio- and chemical reactors modeling and design. Based on success of these applications, Moradi et al. (2013) have aimed to apply ANN modeling for WDT prediction and have compared four different optimizing algorithms of Levenberg-Marquardt (LM), Scaled Conjugate Gradient (SCG), Gradient Descent with Adaptive learning rate (GDA), and Bayesian Regulation (BR). LM application has demonstrated superior results.

Bian et al. (2019) investigated application of various MS-based data modeling highlighting work by Moradi et al. (2013) as described above for application of ANN for modeling WDT. Kamari et al. (2013) have reported a research work on wax precipitation estimation using least-square support vector machine (LS-SVM) accompanied with coupled simulated annealing (CSA). Menad et al. (2018) reported a temperature-based oil-water relative permeability modeling using radial basis function neural network (RBFNN) with grey-wolf optimizer (GWO). Zhou et al. (2018) reported application of differential evolution-SVM to develop color differentiation program for printing and dyeing. Jafari et al. (2019) has reported use of adaptive network-based fuzzy interference system (ANFIS) enhanced with GWO to model landslide potential. Based on performance of these models in various applications, Bian et al. (2019) selected and compared 4 methods for WDT prediction: LSSVM, GA-ANFIS, PSO-ANFIS, and GWO-SVM. The GWO-SVM model performed better than the other three models and is noted to work well with small datasets.

Bian et al. (2019) analyzed Pearson correlation impact of carbon atom number, MW, and pressure on WDT prediction. While, MW and pressure are known to be main contributors, a good insight was a spread between different carbon atom numbers. Standard set of assessment criteria were applied to compare performance of the four methods: average absolute relative deviation (AARD), root-means-square error (RMSE), standard deviation (SD), determination coefficient (R^2), and error percentage (EP). Additionally, William's plot was generated to investigate stability and robustness of the methods that demonstrated GWO-SVM's superiority over other methods in both cases. In multi-component systems with 57 datapoints negligible doubtful data was captured and no doubtful data was captured in datasets with 272 datapoints.

Kamari et al. (2019) applied three different methods of multilayer perceptron (MLP) with LM optimizer, LS-SVM combined with CSA, and decision tree (DT) to estimate WDT

and have concluded that the DT model performed better than the other two methods. The assessment criteria were based on similar statistical parameters as above: average absolute percentage relative error (AAPRE), R^2, SD, and RMSE.

Benamara et al. (2019) discussed earlier published works on data-driven methods development highlighting Moradi et al. (2013) ANN model, Bian et al, 2019 GWO-SVM, and three methods from Kamari et al. (2019) MLP, LSSVM and DT, and proposed their own development of RBFNN with genetic algorithm (GA) and artificial bee colony (ABC) optimizers and development of Group Method of Data Handling (GMDH) derived correlation.

Few approaches are used to evaluate performance of newly developed algorithms. Standard set of statistical parameters: AARD, R^2, RMSE and SD is analyzed. RBFNN-ABC showing superior performance. Cross-plots are generated for unit-slope assessment, all methods show unit-slope behavior, while RBFNN-ABC demonstrating more uniform spread. Error distribution charts are plotted, RBFNN-ABC demonstrated near 0 pattern. Performance is compared with previous models. A stability analysis was done on RBFNN-ABC model, and only 1 data point was found out of range/an outlier, suggesting a data reliability of 99.63% and RBFNN-ABC's model stability and robustness.

Amiri-Ramsheh et al. (2021) paid due credits to previous works done by Moradi et al. (2013), Kamari et al. (2019), Bian et al. (2019) on building data-driven models to estimate WDT. Amiri-Ramsheh et al. (2021) have extended the work further by compiling wide range of models to investigate and compare their capacity for WDT estimation namely MLP with two different algorithms: LM and Bayesian Regularization (BR), RBF, ANFIS enforced with three different optimizers: cultural algorithm (CA), bio-geography based optimization (BBO) and teaching-learning based optimization (TLBO), as well as decision tree models: DT, random forest (RF) and extra tree (ET).

Models' performance assessment was performed similar to previous studies. Statistical analysis: AAPRE %, APRE %, RMSE, SD, and R^2 were calculated for all of the nine methods. RF has demonstrated best performance with AAPRE of 0.246%, RMSE of 1.01, and SD of 0.003. AAPRE was compared for dependency to pressure: RF model has shown lowest AAPRE with no change with pressure and to molar mass: RF has demonstrated similar AAPRE except for MW exceeding 250 g/mol, where AAPRE is highest for all nine models. Cross-plots: all nine models demonstrate unit slope trend, yet RF model plot shows outstanding fit to unit slope line, compared to other eight models, uniform, and well concentrated on unit slope line. Error distribution charts is plotted for RF and demonstrate near 0 concentration. Comparison with previous studies showed that the above-mentioned selected RF model was compared with previous study reported by Benamara et al. (2019) and has demonstrated superior performance. Although the numbers quoted by Amiri-Ramsheh et al. (2021) vary from reported by Benamara et al. (2019), the above statement still holds true where RF performance is reported higher than Benamara et al. (2019): AARD of 0.5402%, R^2 of 0.9706, RMSE of 1.9969, and SD of 4.69×10^{-5}. Cumulative frequency plot is generated for APRE% for four best performers: MLP-BR, ANFIS-BBO, DT, RF and Benamara et al. (2019). RF demonstrates best performance with 80% of output within APRE 0.4% and 90% within 0.6%. The plot suggests less than 10% of results within APRE of 1% for Benamara et al. (2019). Accuracy variation with temperature (laboratory WDT) were compared for four best performers: RF, DT, MLP-BR, and RBF within range of 270-315°K. While 56 data points from Metivaud et al. (1999) display accuracy within +/-0.3°K, the complete dataset accuracy variation range is extended to +/-2°K. Additionally, a detailed uncertainty analysis was performed for RF for the complete dataset.

Trend analysis was conducted to check the WDT's dependency on pressure and molar mass based both on laboratory data and RF modeling. WDT linearly increases with pressure

and non-linearly increases with molar mass. Sensitivity analysis was run on the RF model based on relevancy factor assessment that concluded higher contribution of pressure to the WDT value than of molar mass. Stability analysis was performed on the RF model based on Leverage approach with Williams plot that concluded no data out of leverage and only 6 data points found out of suspected limits out of 346 data points, which corresponds to 98.2% of laboratory data reliability and suggests RF model's robustness.

Nait-Amar et al. (2024) analyzed the literature with regards to ML-based methods application for WDT estimation such as Moradi et al. (2013), Kamari et al. (2019), Bian et al. (2019), Benamara et al. (2019) and Amiri-Ramsheh et al. (2021). They suggested that there is still room for further improvement on WDT estimation and proposed a new explicit correlation for WDT calculation based on gene expression programming (GEP).

When applying GEP and tuning the model, Nait-Amar et al. (2024) reported WDT as a function of molar mass and pressure should actually be represented as combination of two correlations that are applied conditionally depending on value of molar mass. This discontinuous relation of WDT to molar mass can be observed in WDT trend analysis where WDT is linearly related to pressure, and there is non-linear relation to molar mass and two slopes are observed both in current study and in research works by Benamara et al. (2021). While Nait-Amar et al. (2021) selected border point of molar mass as 210 g/mol based on GEP tuning, in works by Benamara et al. (2019) this break in linear trend is observed at 140 g/mol and in WDT trend plot by Nait-Amar et al. (2021) slope change is observed around same value of 120–140 g/mol. The GEP model was compared to GMDH explicit model by Benamara et al. (2021) and with thermodynamic modeling outputs by Benamara et al. (2021): The ideal solid model by Hansen et al. (1988), multi-pure-solid model by Lira-Galeana et al. (1996), and the UNIQUAC model by Coutinho et al. (1998). All these works use a similar the input data. The GEP model performed better than all four models (one white box GMDH and three thermodynamic).

The GEP model results were evaluated based on following aspects: Statistical analysis where AARD = 0.5963%, RMSE = 2.1963, and R^2 = 0.9647. Cross-plots showed that most of the data lie on the unit slop line. Error distribution chart showed the results are evenly distributed around zero-error line within 2% ARD. Cumulative frequency plot indicated 50% of output within 0.5% ARD, and 80% of data within 1% ARD. Shapley Additive Explanation graph (SHAP) was used to assess correlation with principal physics, applied to the GEP results. The diagram indicates WDT's direct dependency on pressure and molar mass. The SHAP values further analyzed against each input separately, pressure and molar mass indicating a linear relation to pressure and a non-linear relation to molar mass with slope discontinuity. The latter trend shows 3 slopes: 100–125, flat 125–200, and 200–275. Plots were built for trend analysis of WDT's dependency on molar mass at different pressures show linear relationship but with different slopes for each pressure. It should also be noted here that molar mass range at excess of 204 g/mol is analyzed that may not show the full picture.

3 METHODOLOGY

In this chapter, ML approaches in general and decision tree (DT) methods in specific are briefly discussed and historical development chronology of DT-based ML-algorithms is presented. Among DT methods the selected four models with boosting algorithms are reviewed: AdaBoost, GBM, XGBoost, and CatBoost, latter two being recent developments. Comparative analysis of these four methods is provided with field of application suggested for each. Tuning mechanism for these approaches with hyper-parameters is presented in detail. The proposed research workflow is clearly formulated and described with thorough input analysis and the ML methods fine-tuning settings to be used are highlighted.

3.1 Machine Learning Methods

Machine learning methods are mainly classified in four groups by the way they handle input the data and the algorithms used: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning (Acharya, 2021).

3.1.1 Supervised learning

The data with output values are fed in training process so that the machine can establish a relationship between inputs and outputs and apply this relationship to generate outputs for the input data fed in without output values. Data processing is either for data classification or regression. Schrider et al. (2018) share practical example of supervised learning for email classification to identify spam emails using the input data such as domain name, IP address, host name, and format. Chinnamgar (2019) suggests application of supervised learning methods in cases where possible output categories or continuous values are known but corresponding outputs for given inputs are missing.

3.1.2 Unsupervised learning

The data are fed in training process without output or target class. The machine establishes patterns in the input data by building relationship between features of the input data. Methods used with examples

by Geron (2019) and Zhang and Zhang (2003): clustering with example of customer segmentation, with K-Means, DBSCAN algorithms; Hierarchical Clustering, dimensionality reduction w/ example of visualization, feature selection, with PCA, t-SNE algorithms; anomaly detection w/ examples of fraud detection and cybersecurity, with Isolation Forest, One-Class SVM algorithms; and association rule learning with an example of market basket analysis, recommendation systems with example Apriori algorithm.

3.1.3 Semi-supervised learning

In this method, both labeled (supervised) and unlabeled (unsupervised) data are combined with the aim to improve the learning efficiency. It is used when there is abundance of unlabeled data, as labeling might be time-consuming or unaffordable. Van Engelen and Hoos (2020), share examples of computer-aided diagnosis, drug discovery or speech tagging process. Geron (2019) shares clear practical application of Google photo of family members, where names tagged in few pictures will allow machine to recognize them in the rest of the pictures.

The following methods are used in subject approach: predicting labels of unlabeled data and retraining, with examples of image classification, speech recognition and algorithm of self-training, graph based label spreading with examples of fraud detection, text classification and algorithm of Label spreading; improved label propagation with examples of fake news detection, bioinformatics and algorithm of Label propagation, generating labeled data with generative adversarial networks (GAN) with examples of medical imaging, natural language processing (NLP) and algorithm of Semi-supervised GANs, two classifiers reinforcing each other with examples of sentiment analysis, text mining and algorithm of Co-training.

3.1.4 Reinforcement learning

Machine interacts with environment and receives rewards or penalties when making decisions, the aim is to maximize the cumulative rewards with time as per Mohri et al. (2018). Geron (2019) shares example of Deepmind's Alphago program where the machine learnts to play Go

by trial and error and eventually has beaten human. Following methods could be listed for subject approach: table-based for small environments like grid-world, basic robotics with algorithm of Q-learning, value-based for video games, simple robotics with algorithm of DQN (deep q-learning), policy-based for continuous actions such as robotics, trading with algorithm Policy gradient (REINFORCE), or advanced AI agents such as Dota-2, Chess AI and example algorithm of Proximal Policy Optimization (PRO).

3.2 Supervised Learning ML Methods

ML methods for supervised learning approach can be further classified into the following groups: Support Vector Machines with example algorithms of LSSVM (Least Squares Support Vector Machine) and GWO-SVM (Grey Wolf Optimizer-based SVM). Decision Tree based methods with example algorithms of DT (decision tree), RF (random forest), and ET (extra tree). Artificial Neural Networks and its variations with example algorithms of ANN (Artificial Neural Network), MLP (Multi-Layer Perceptron), and RBFNN (Radial Basis Function Neural Network). Other supervised methods such as white box models GMDH (Group Method of Data Handling) and GEP (Gene Expression Programming), and ANFIS (Adaptive Neuro-Fuzzy Interference System).

3.3 DT Algorithms

Algorithms for Decision Tree (DT) approaches have evolved with time as shown in Table 3.1 where some of these evolvements are presented in chronological order.

Table 3.1. Decision Tree algorithms' evolvement.

Year	Algorithm	Developer	Main Features	Improvements	Advantage	Application	Reference
1966	**CLS** (Concept Learning System)	Hunt et al.	First formalized decision tree method.	First attempt at decision tree learning.	Basis for future decision trees.	Early AI experiments.	Hunt et al. (1966)
1985	**ID3** (Iterative Dichotomiser 3)	Ross Quinlan	Uses **Information Gain** to split nodes.	More efficient than CLS.	Easy to interpret, handles categorical data.	Simple classification problems.	Quinlan (1986)
1984	**CART** (Classification and Regression Trees)	Breiman et al.	Uses **Gini Index**, supports **regression trees**, allows pruning.	Supports both classification & regression.	Handles missing values, binary splits improve performance.	Finance, risk assessment.	Breiman et al. (1984)
1993	**C4.5**	Ross Quinlan	Uses **Gain Ratio**, supports continuous data & pruning.	Handles missing values, better splitting criteria.	More stable than ID3, widely used.	Medical diagnosis, credit scoring.	Quinlan (1993)
1995	**C5.0**	Ross Quinlan	Improved C4.5 with **boosting** & faster execution.	More efficient, reduces overfitting.	Generates smaller trees, better accuracy.	Telecom, banking, fraud detection.	Quinlan (1998)
1995	**AdaBoost** (Adaptive Boosting)	Freund & Schapire	Combines weak learners using	First major **boosting algorithm**.	Reduces bias, improves accuracy.	Face detection,	Freund & Schapire (1997)

			adaptive weighting.			fraud detection.	
1999	**GBM** (Gradient Boosting)	Jerome Friedman	Uses **gradient descent** to minimize loss.	More flexible than AdaBoost.	Works with **various loss functions**.	Risk modeling, healthcare, finance.	Friedman (2001)
2001	**RF** (Random Forest)	Leo Breiman	Uses **bagging & multiple decision trees**.	Avoids overfitting of single trees.	Handles large datasets, more robust.	Fraud detection, customer segmentation.	Breiman (2001)
2006	**Extra Trees** (Extremely Randomized Trees)	Geurts et al.	Randomized feature splits, reduces variance.	More randomized than Random Forest.	Faster, less prone to overfitting.	High-dimensional datasets, big data.	Geurts et al. (2006)
2014	**XGBoost** (Extreme Gradient Boosting)	Tianqi Chen	Optimized GBM with **regularization & parallelization**.	Faster & more accurate than GBM.	Reduces overfitting, highly scalable.	Kaggle competitions, AI research.	Chen & Guestrin (2016)
2017	**LightGBM** (Light Gradient Boosting Machine)	Microsoft (Ke et al.)	**Leaf-wise splitting** instead of level-wise.	Faster than XGBoost for large datasets.	Handles **millions of samples** efficiently.	High-performance ML tasks.	Ke et al. (2017)
2018	**CatBoost** (Categorical Boosting)	Yandex (Prokhorenkova et al.)	Handles **categorical features natively**.	Avoids the need for one-hot encoding.	Best for categorical data, prevents overfitting.	E-commerce, finance, NLP.	Prokhorenkova et al. (2018)

Previous ML modeling works discussed in Chapter 2 have used SVM, ANN variations, GMDH, GEP, and ANFIS and Kamari et al. (2019) and Amiri-Ramsheh et al. (2021) have used DT methods with RF and ET algorithms. This work aims to investigate application of DT method with boosting mechanism for WDT modeling: AdaBoost, Gradient Boosting Machines, XGBoost, and CatBoost. DT methods with boosting mechanisms are compared with LR and KNN.

3.3.1 AdaBoost

Adaptive Boosting was introduced by Yoav Freund and Robert Schapire in 1995. It combines multiple weak learners to produce a strong output using adaptive weighting mechanism where misclassified learners receive higher weights in the next iteration round. Sequential training process introduced sequential ensemble method for later boosting algorithms. The method is less prone to overfitting and manages imbalanced datasets well. However, it is sensitive to outliers and is not efficient for complex problems.

3.3.2 GBM

GBM algorithm was developed in 1999 by Stanford University professor Jerome H. Friedman and is designed to generate competitive, highly robust, and interpretable models for both regression and classification tasks (Friedman, 2001). Its mechanism is based on ensemble approach that combines multiple base learners to improve the predictive accuracy. The GBM was derived from AdaBoost (Brownlee, 2016) and laid foundation for XGBoost and CatBoost. Unlike AdaBoost instead of assigning weights, the GBM minimizes loss function using gradient descent, i.e. the output of each base learner is boosted by factor of a fixed learning rate and then fed into the prediction made by the previous base learner. The output from each learner is summed up to calculate model's predicted value. Base learners serve as the fundamental building blocks of the final model and are often referred to as weak learners since their individual predictions are only slightly better than random guessing. The GBM models can be constructed using various types of base learners: Linear, Smooth, and DT models. DT

based learning model will be used for GBM base learners in this research work. As mentioned above, the GBM may address both classification and regression tasks and it uses different algorithms for each of the tasks. The GBM further differentiates into 3 groups based on the the input data volume: Batch GBM, Mini-batch GBM, and Stochastic GBM.

3.3.3 XGBoost

XGBoost is extreme gradient boosting algorithm to perform classification, regression, and ranking tasks that was introduced in 2016 by Chen, T. and Guestrin, C. It is based on gradient boosting mechanism and inherits its mechanism of populating base learners building them into assembled model. The XGBoost follows sequential ensemble method where additive classified is produced. Regularized objective function of the XGBoost can be expressed as following:

$$L(\varphi) = \sum_i l(\tilde{y}_i, y_i) + \sum_k \Omega(f_k) \qquad \text{Equation 3.1 (Acharya, 2021)}$$

where: $l(\tilde{y}_i, y_i)$ is the loss function. Various loss function options are available such as hinge loss, logistic loss, cross entropy loss, and exponential loss. However, the commonly used are mean squared error and logistic loss. $\Omega(f_k)$ is regularization component to avoid data over-fitting in training process.

The XGBoost improvements over the GBM is high processing speed (Chen and He (2014) report 10 times higher speed for the XGBoost than for the GBM) and forecast accuracy. These two improvements are achieved with following features of the XGBoost: Shrinkage sizes newly introduced weights by factor, η – learning rate. Column sub-sampling allows random features subsets selection to avoid hurdle with the algorithm having to consider all features in the dataset when training. This feature significantly reduces training time. The best split candidate is located applying either exact greedy algorithm (when complete dataset is fit in one location) or approximate algorithm (when dataset is spread across different locations). Sparse data management allows large complex datasets processing with missing attributes or multiple

zero values. Out-of-core computations allows the XGBoost to process data from secondary memory locations if dataset does not fit in primary location and is segmented in different memory locations. This directly affects and improves the processing speed.

3.3.4 CatBoost

CatBoost is Categorical Boosting algorithm introduced by Prokhorenkova et al. (2017). The algorithm is based on gradient boosting mechanism and works well both with categorical and numerical datasets. Below features shape up the CatBoost as one of the best performers in ML-based models. The CatBoost allows categorical attributes to be converted into numerical data in the data pre-processing using encoding technique, where instances of categorical examples are counted and used as numerical value. This significantly improves algorithm performance. To address overfitting issue, the CatBoost applies special approach to process categorical values where attributes are permuted or shuffled, and an average value is calculated and used in each run during training. Feature combination is main differentiator of the CatBoost over other algorithms. It allows the CatBoost to produce powerful attribute by combining the existing data features. For this CatBoost uses greedy approach, where it misses out first round (first split in the tree) but compiles all combinations and categorical attributes in current tree with categorical attributes of the whole dataset. Another differentiator of the CatBoost over other algorithms is "Fighting Gradient Bias", where the approach for building tree structure and setting leaf values differ. It uses modified gradient-based DT in previous process and traditional gradient boost DT in latter.

3.3.4.1 Boosting Algorithms Comparison

A comparison between four boosting algorithms used in this research work for WDT modeling are presented in Table 3.2.

Table 3.2. Boosting algorithms comparison.

Algorithm	Speed	Missing Data Management	Large Datasets	Categorical Data Management	Tuning Complexity
AdaBoost	Slow	No	No	No	Low
GBM	Moderate	No	No	No	High
XGBoost	Fast	Yes	Yes	No	High
CatBoost	Fast	Yes	Yes	Yes	Medium

3.4 Hyper-parameters

Hyperparameters are setting specifications for boosting algorithm that control its performance, speed, and accuracy. As an example, if tree grows into width or depth, it may face overfitting issues during the training process. Hence, the tree growth parameters should be controlled pruning method and parameters used would be tree size and depth. These parameters are tuned to yield high performance of the algorithm. The hyper-parameters used in boosting techniques in this research work are described in the following sections.

3.4.1 Learning Rate

Gradient boosting algorithms apply sequential ensemble modeling, where the results of gradient/tree/base learner is multiplied by learning rate to define the step size for the next gradient/tree/base learner definition. The range is defined between 0 to 1. The default values for the XGBoost is 0.1 and for the CatBoost is 0.03. If the values are set too low training time is increased. The time reduction is achieved by increased learning rate and reduced number of trees. Parameter referral: *learning_rate, shrinkage_rate*

3.4.2 Number of Trees

The parameter stands for number of boosting iterations, number of trees, or number of base learners generated. The range for the XGBoost is 100 to 5000 and the default is 100. The range for the CatBoost is not limited and the default is 1,000. Parameter referral: *n_estimators, num_iterations, iterations*

3.4.3 Maximum depth

This parameter refers to the distance from root node to the end leaves or sum of the splitting nodes from root node, which is measured as number of levels from where decision branches separate. The maximum feasible could be the number of attributes minus one. However, increased depth complicates the process and runs into risk of over-fitting. Hence, an optimal number is found by tuning. Default value for XGBoost and CatBoost is 6. Parameter referral: *max_depth, depth*

3.4.4 Minimum child weight

A threshold value that ensures the leaf is formed only if minimum number of samples is achieved. This helps to prevent overfitting. The default value for the XGBoost is 1, if dataset is large value is set between 5 to 10. The parameter can be increased to reduce the training time.

3.4.5 Gamma

This parameter is the XGBoost's pruning criterion. It is lower than the gain calculated previously then pruning takes place; otherwise, no further pruning required. This parameter controls overfitting.

3.4.6 Subsample or Colsample

Processing complete dataset increases the training time and reduces the efficiency. A solution is proposed where part of data is only used in each iteration round at a time. Subsample is splitting dataset into sub-sets and Colsample is selection of specific sub-set for training iteration round. It ranges from 0.1 to 1, complete dataset representing 1.

3.4.7 Regularization parameters (alpha and lambda)

These parameters address the overfitting issue. If the dataset has multiple attributes and they all are used during the training process, then this might lead to overfitting. To prevent this,

reduced number of attributes are used in training and this regularization helps to define those impactful features. Regularization reduces evenly the impact weight of those less impactful features. Alpha and lambda are referred to as L1 and L2 regularization parameters, consecutively.

3.4.8 Random strength

This parameter controls the CatBoost overfitting. On each iteration round various splitting scenarios are developed and scored based on the loss function reduction. The split with lowest loss function is selected. Parameter referral: *random_strength*

3.4.9 Bagging temperature

This parameter controls settings of the Bayesian bootstrap in the CatBoost which assigns random weights to objects. It is applied by default in both classification and regression modes, helping to improve model stability and performance by introducing randomness in weight assignment. Parameter referral: *bagging_temperature*

3.4.10 Border count

This parameter in the CatBoost defines the number of splits for numerical features, playing a key role in controlling model complexity. By limiting the number of splits, it helps to prevent overfitting and improves the model's generalization to new data. Parameter referral: *border_count*

3.4.11 Tree growing policy

Different tree growing strategies exist. However, the default for CatBoost is Symmetric Tree, which is reported to have 10 times higher processing speed than non-symmetric trees. On each iteration leaves are split with the same condition that supports tree symmetry.

Table 3.3. Hyper-parameters used in selected boosting algorithms with respective value ranges.

Hyper-parameter	AdaBoost range	GBM range	XGBoost range	CatBoost range
n-estimators, iterations	50 – 500	100 – 500	100 – 1,000	500 – 5,000
learning_rate	0.01 – 1.0	0.01 – 0.2	0.01 – 0.2	0.01 – 0.2
base_estimator	Decision Tree Classifier			
algorithm	SAMME.R			
random_state	Any integer			
max_depth, depth		3 – 10	3 – 10	4 – 12
min_samples_split		2 – 10		
min_samples_leaf		1 – 10		
subsample		0.5 – 1	0.5 – 1	
loss (deviance, exponential)		deviance		
min_child_weight			1 – 10	
gamma			0 – 5	
colsample_bytree			0.5 – 1	
alpha			0 – 10	
lambda, l2_leaf_reg			0 – 10	3 – 10
random_strength				1 – 10
bagging_temperature				0 – 1
border_count				32 – 255

3.4.12 Hyper-parameters search methods

Hyper-parameters optimization is process of locating the best set of hyper-parameters to yield the best results. Locating the best set induces multiple iterations of trial and error with manual involvement. For the sake of efficiency that process is automated using various techniques such as Grid Search, Random Search, and Bayesian Optimization.

3.4.12.1 Grid Search

Every possible scenario is checked using grid search. This selects the best but is time consuming.

3.4.12.2 Random Search

This function reduces the number of scenarios by randomly selecting sets. Bergstra and Bengio (2012) suggest similar to Grid Search results are obtained but in less time. However, the method might miss out the best solution.

3.4.12.3 Bayesian Optimization

This method can be used to build probabilistic models of error function. The used combination of hyper-parameters is saved and take part in the probabilistic approach reducing the processing time.

3.5 Data Modeling Approach

3.5.1 Input data sources

Publications with WDT measurements were reviewed and the data reported was analyzed and the results are compiled in Table 3.4. Robles et al. (1995) listed phase transition temperatures for solid-solid and solid-liquid systems in Tableau 1 of original publication. Following Ji et al. (2019) review, solid-liquid transition temperatures are used, Tsol for 0% and 100% C_{19} molar composition and Tliq for in-between values. Pauly et al. (2001) reported data for PC% of paraffins crystalized and the WDT was not directly measured. Hence, the dataset is not used. Daridon et al. (2002) reported data for solvent mono-component *n*-tridecane which is not used, only data reported for 9 mixtures is used. The data published by Ji et al. (2003) in Figures 9 and Tables 1 and 2 of original publication contain the data reported by Robles et al. (1995) and Metivaud et al. (1999) and were not used to avoid duplication. The data presented in Figures 11, 12, 13 and 14 of original publication were not used as the reported parameter is wax precipitation amount, mass %, not WDT. The data reported by Milhet et al. (2005) presented

in Table 4 of original publication is not used as reported for metastable rotator phase. The data reported by Rizzo et al. (2007) in Figure 7 of original publication have mass composition that totals to 90% only. The mass composition was increased in proportion. The data reported by Monsourpoor et al. (2019) as presented in Tables 3, 4, 5 and 6 of original publication was not used as data was previously reported by Metivaud et al. (1999). The data presented in Table 1 and Table 9 of original publication was also not used as WDT is not directly measured but rather estimated based on measured WAT data. Table 9 of original publication contains measured WDT both with viscometer and DSC. The DSC data is used based on Monsourpoor et al. (2019) observation, page 10: "DSC has higher accuracy in comparison with viscometry".

Table 3.4. Review of the input data sources.

Reference	Samples	Data points total 380	Composition	Pressure	Outcome
Robles et al. (1995)	Tableau 1	8	Binary system C_{17}+C_{19}	Atmospheric	WDT
Metivaud et al. (1999)	Table 4 (11): C_{14}-C_{15}-C_{16} Table 5 (11): C_{16}-C_{17}-C_{18} Table 6 (18): C_{18}-C_{19}-C_{20} Table 7 (16): C_{19}-C_{20}-C_{21}	56	Four ternary systems C_{14}-C_{21}: C_{14}+C_{15}+C_{16}, 11 data points C_{16}+C_{17}+C_{18}, 11 data points C_{18}+C_{19}+C_{20}, 18 data points C_{19}+C_{20}+C_{21}, 16 data points.	Atmospheric	WDT
Dauphin et al. (1999)	Table 1 and Table 2	5	Table 1 - 5 mixtures, mass % nC_{10}, nC_{18}-nC_{36}, MW g/mol Table 2 - WDT for 5 mixtures	Atmospheric	WDT
Pauly et al. (2001)	Table 3	Not used, 92, PC% of paraffins crystallized	nC_{13}-nC_{36}	0.1, 10, 30, 50 MPa	PC%
Daridon et al. (2002)	Table 2 (Fig.3)	Not used, 6, not wax	*n*-tridecane (not wax)	6 data pairs, 0.1 MPa - 98.3 MPa	WDT
	Table 3, 4	54	9 mixtures, nC_{13}-nC_{24}	6 data pairs, 0.1 MPa - 98.3 MPa	WDT
Ji et al. (2003)	Fig.5	10	$C_{6+}C_{16}$ and C_6+C_{17} binaries, xi for C_{16} and C_{17}, 5x2 = 10 points	Atmospheric	WDT
	Fig.6	12 – 1 =11	C_{16}+C_{18} binary, xi for C_{18}, 12 points	Atmospheric	WDT
	Fig.7	10 – 1 = 9	C_{16}+C_{20} binary, xi for C_{20}, 10 points	Atmospheric	WDT
	Fig.8	7 – 2 =5	C_{15}+C_{19} binary, xi for C_{19}, 7 points	Atmospheric	WDT
	Fig.9 (Robles)	Not used, 8 (Robles)	C_{17}+C_{19} binary, xi for C_{19}, 8 points	Atmospheric	WDT
	Table 1	Not used, 11 (Metivaud)	xi for C_{14}+C_{15}+C_{16} ternary, 11 points	0,1MPa	WDT
	Table 2	Not used, 18 (Metivaud)	xi for C_{18}+C_{19}+C_{20} ternary, 18 points	0,1MPa	WDT
	Table 3	3	xi for C_6+C_{16}+C_{17} ternary, 3 points	0,1MPa	WDT
	Table 4 and Fig.10	19	Composition for 3 mixtures: A, B, C	A-7 values, B-7 values, C-5 values	WDT
	Fig.11	Not used		0.1MPa and 50MPa	Wax, mass %
	Fig.12	Not used	Carbon number, 31 points (20 zero values)	0.1MPa	Wax, mass %
	Fig.13	Not used	Carbon number	0.1MPa	Wax, mass %
	Fig.14	Not used	Carbon number	0.1MPA - 11 non-zero values, 50MPa - 14 non-zero values	Wax, mass %
Milhet et al. (2005)	Table 1	84	nC_{14}+nC_{16}, 14 compositions	6 data columns, 0.1-20-40-60-80-100 MPa	WDT
	Table 2	60	nC_{14}+nC_{15}, 10 compositions	6 data columns,	WDT

				0.1-20-40-60-80-100 MPa	
	Table 4	Not used, 33, rotator	Metastable rotator phase: nC_{14}+nC_{16}, 7 compositions	6 data rows, 0.1-20-40-60-80-100 MPa	WDT
Rizzo et al. (2007)	Fig.5	17	Octadecane from Wurflinger-5, Nelson-3, Domanska-4, Rizzo-5	0.1-150 MPa	WDT
	Fig.7	9	Pauly (5), Rizzo (4) synthetic tetradecane + wax, nC_{20}-nC_{42}, mass composition issue 83.66% + 6.34% = 90%	0.1-100 MPa	WDT
Mansourpoor et al. (2019)	Tables 3, 4, 5, 6 (Metivaud et al)	Not used, Repeats Metivaud's data	Sys#1: 11 mixtures, $C_{14+}C_{15}$+C_{16}, mol%, Sys#2: 11 mixtures, C_{16}+$C_{17+}C_{18}$, mol%, Sys#3: 18 mixtures, $C_{18+}C_{19}$+C_{20}, mol%, Sys#4: 16 mixtures, C_{19}+$C_{20+}C_{21}$, mol%,	Atmospheric	WDT
	Table 1, Table 9	9	12 Iranian oils: Table 1 - composition, Table 9 – WDT	Atmospheric	WDT (WAT)
Shariatrad et al. (2022)	Table 1 and Fig 1	30	30 mixtures of quarternary system, weight fraction, $nC_{11}H_{24+}nC_{14}H_{30}$ +$nC_{16}H_{34+}nC_{18}H_{38}$	0.9 bar	WDT

3.5.2 Input parameters

As discussed earlier in theoretical review section, Wax Disappearance Temperature is affected by fluid composition, system pressure, and intermolecular interactions because of their influence on wax solubility and phase behavior. While fluid composition is critical component incorporating it as input parameter is not straight forward. Hence, molar weight is proposed as simplified representation. Wax precipitation is primarily influenced by heavier hydrocarbons and molar weight sufficiently captures its effect. System pressure effect on WDT is well explained in theoretical section. Lighter hydrocarbons remain in solution at higher pressure that delays wax precipitation resulting at lower WDT. As pressure reduces, the lighter hydrocarbons escape the solution to gas phase and wax precipitation is expedited leading to higher WDT. While, advanced thermodynamic models incorporate molecular interactions, the basic models assume this effect is indirectly managed by molar weight. Thus, all the WDT intelligent data modeling works reviewed earlier use molar weight and system pressure as the input parameters for WDT prediction as these are believed to consider primary thermodynamic influences on

wax solubility and precipitation resulting in computationally efficient and highly accurate approach for predicting WDT. In scope of this research work, it is proposed to base the WDT modeling onto molar weight, MW in g/mol and system pressure, P in MPa.

3.5.3 Input data analysis

3.5.3.1 *Dataset statistical description*

Statistical description of the database consisting of 380 datapoints assembled for the purpose of this research work is summarized in Table 3.5.

Table 3.4. Input data statistical description.

Parameter	Min	Avg	Max
Input, P (MPa)	0.09	30.37	144.4
Input, MW (g/mol)	88	214.99	294.46
Output, WDT (K)	247.65	294.05	333.16

3.5.3.2 *Histograms*

Distribution of the data is further presented using histograms for each input parameter. MW: The distribution is left-skewed suggesting a bi-modal distribution which might fit well with categorical separation. Feature transformation such as log transformation can be considered for data normalization. Pressure: The distribution is highly right-skewed, which is expected with most of published data reported at atmospheric pressure. Categorical classification could be considered. Log transformation might be required if regression model to be used. WDT: The data is normally distributed. No transformation is required, and a linear model can be used.

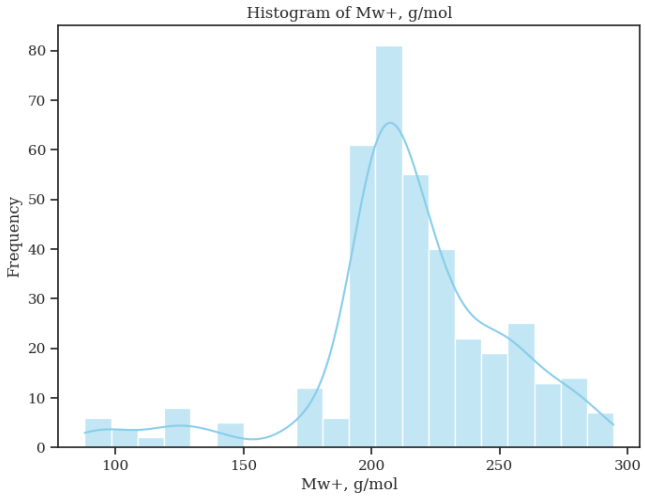

Figure 3.1. Histogram for molar weight data, g/mol.

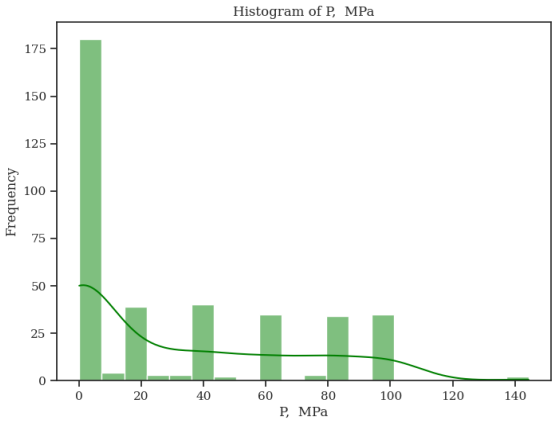

Figure 3.2. Histogram for pressure data, MPa.

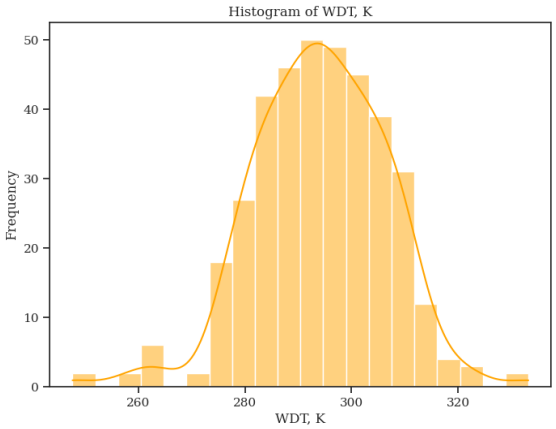

Figure 3.3. Histogram for WDT data, °K.

3.5.3.3 *Box plots*

Similar to histograms, box plots can provide further understanding of data distribution and characteristics such as central tendency (median), spread, outliers, and skewness that may require some data pre-processing procedures such as outlier management, scaling (normalization/standardization), and log transformation.

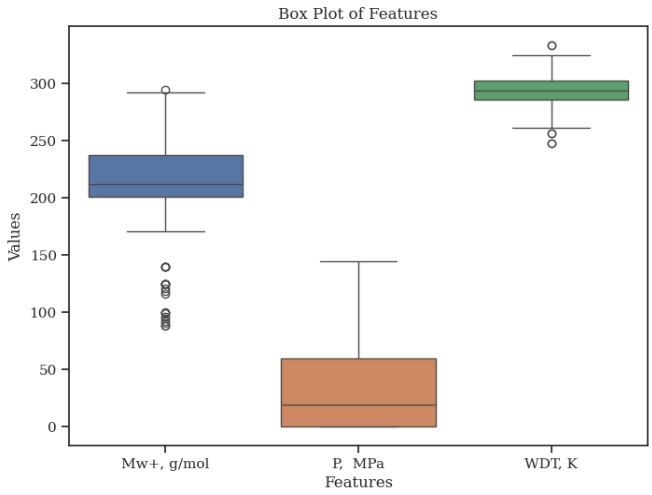

Figure 3.4. Box plots of MW, P, and WDT.

MW: The data is centered around 200-250 g/mol. Some outliers are spotted below 150 g/mol. These outliers can be managed either by clipping or transformation. The distribution is right-skewed. P: The distribution is highly right-skewed with most values at atmospheric pressure. Log transformation might be required. WDT: There is a normal distribution. Few outliers are spotted but no further pre-processing is required.

3.5.3.4 *Pair plots and Pair plots with KDE*

For further detailed visualization, the data is also plotted in form of scatter plots with histograms (pair plots) and pair plots with Kernel Density Estimation (pair plots with KDE) as shown in Figures 3.5 and 3.6. These plots help to detect if there are any linear or non-linear relationships between the the input parameters or to identify if any clusters or patterns present within dataset, suggest if any outliers, as well if some the input parameters are strongly correlated with one of these parameters that can be

removed to avoid redundancy and to improve the modeling efficiency. As one can see from Figure 3.5, the pair plot suggests a strong positive correlation between MW and WDT, categorical spread between MW and P, and non-linear relationship between P and WDT. In addition, the pair plot with KDE suggests a high skewness for P (mostly atmospheric data) and a normal distribution for WDT (Figure 3.6).

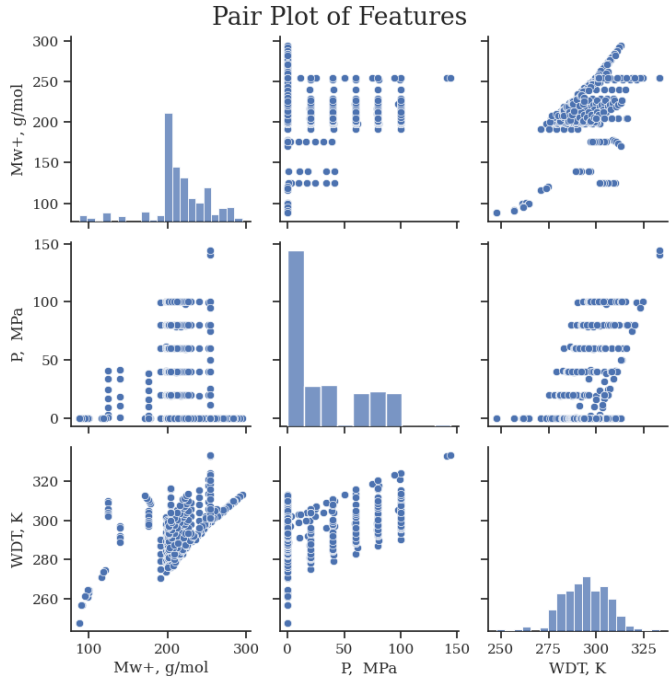

Figure 3.5. Pair plot of the the input parameters.

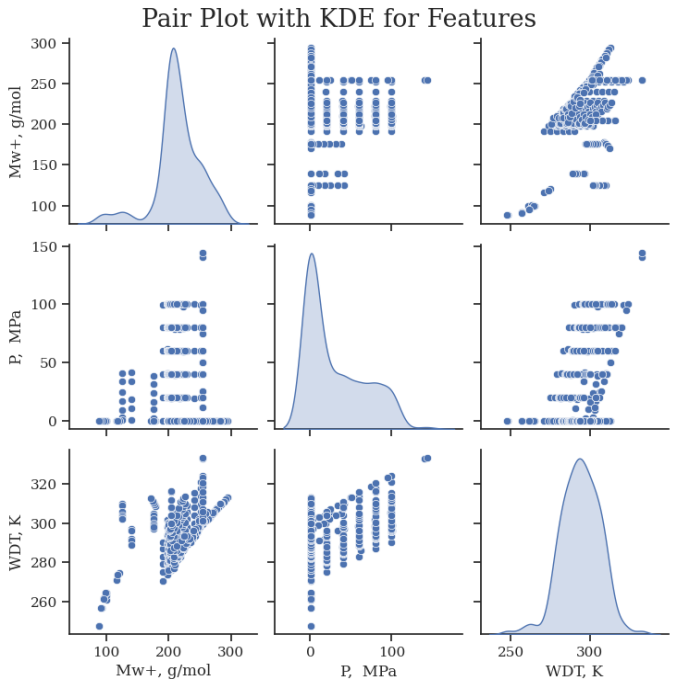

Figure 3.6. Pair plot with KDE of the the input parameters.

3.5.3.5 *Correlation heatmap*

This chart is generated between three variables: MW, P, and WDT as presented in Figure 3.7. The following conclusions can be drawn from the heat map: A positive correlation is present between MW and P with WDT. MW and WDT display the strongest relation and P has a weak relation to both parameters.

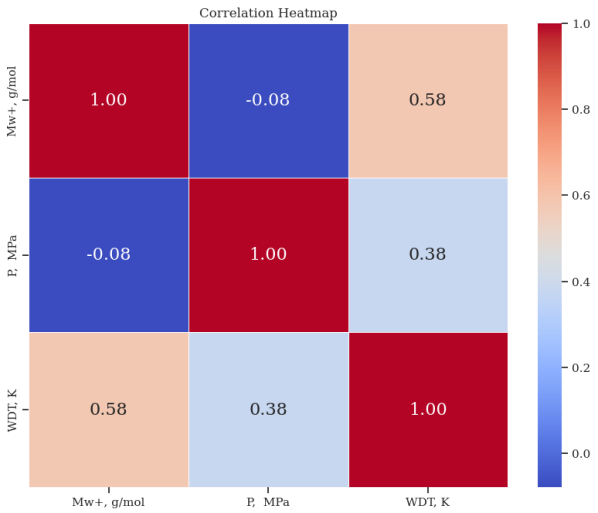

Figure 3.7. Correlation heatmap between the the input parameters.

3.5.3.6 *Violin plot*

Violin plot is another way to present the data combining both data statistics as in box plot and data distribution as in density plot. Similar to previous forms of data presentation, violin plots help to assess skewness and detect if any multi-modal distribution is present, and to check for outliers (wider sections for highly dense data and narrow sections for sparse data). MW: Violin plot presents bi-modal distribution suggesting two categories in dataset in terms of MW (Figure 3.8). P: Violin plot presents right-skewed distribution with long tail due to mostly atmospheric measurements reported (Figure 3.9). Data transformation could be reviewed for data processing, extreme values are to be assessed prior linear regression application. WDT: Violin plot presents normal distribution fit for use in linear models (Figure 3.10).

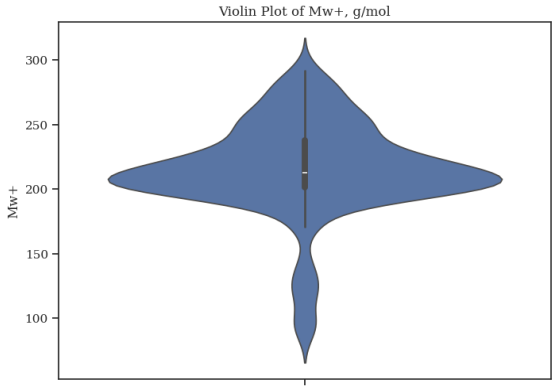

Figure 3.5. Violin plot for MW.

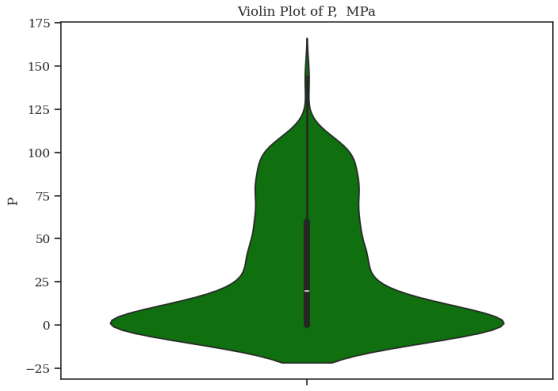

Figure 3.6. Violin plot for P.

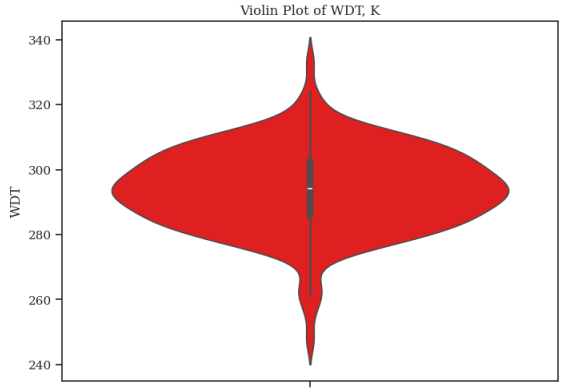

Figure 3.7. Violin plot for WDT.

3.5.3.7 *Feature importance chart*

This chart (Figure 3.11) helps to assess contribution of each of the parameter for output determination. The CatBoost uses two methods to calculate feature importance: a) analyzes a change in model prediction performance if either of the input variables are removed and b) analyzes change in loss function if either of the input variables are removed. This chart helps to exclude irrelevant or redundant parameters and reduces overfitting. Less important data can be missed out to gain on processing speed. MW shows the highest importance of ~70% and P shows lower importance of ~30% (Figure 3.11).

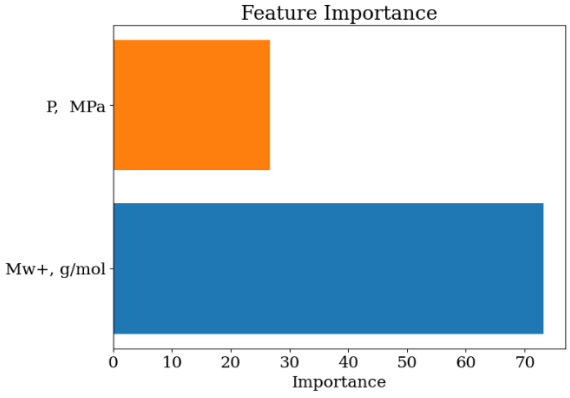

Figure 3.8. Feature importance chart computed by the CatBoost algorithm.

3.5.4 ML methods deployment

The DT methods used in this research work benefit from boosting mechanisms for WDT modeling: AdaBoost, Gradient Boosting Machines, XGBoost, and CatBoost. These methods are compared with LR and KNN for reference.

3.5.5 Training and Testing

The datasets for training and testing of the models are split using the 80-20 strategy where 80% of the data was used for training the model and 20% for testing. Applying the same setting to all four DT models assures fair comparison between model performance.

3.5.6 Hyper-parameters

The hyper-parameters used to tune the four developed DT models are presented in Table 3.6. Here, parameters are set same between models to ensure consistency and fair comparison.

Table 3.5. Hyper-parameters used for tuning the models with set values.

Hyper-parameter	AdaBoost range	GBM range	XGBoost range	CatBoost range
n-estimators, iterations	500	3,000	3,000	3,000
learning_rate	0.1	0.1	0.1	0.1
base_estimator	Decision Tree Regressor ()			
algorithm	SAMME.R			
random_state	42	42	42	42
max_depth, depth	6	6	6	6
min_samples_split		•		
min_samples_leaf		•		
subsample	0.8	0.8	0.8	0.8
loss (deviance, exponential)		Deviance		
min_child_weight			1	
gamma			0	
colsample_bytree			0.8	
alpha			0.1	

lambda, l2_leaf_reg			1	•
random_strength				•
bagging_temperature				•
border_count				•

3.5.7 Methodology workflow

Below flow chart outlines methodology workflow for WDT modeling.

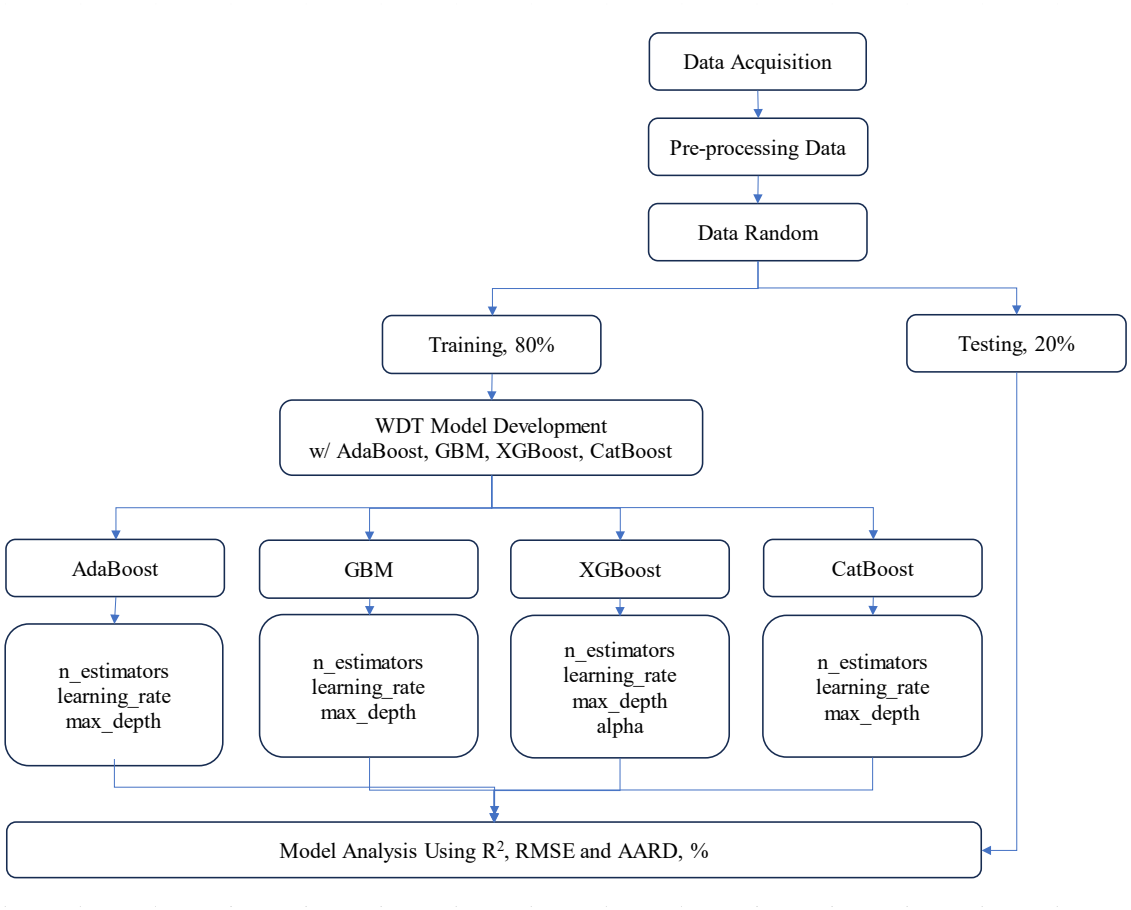

Figure 3.9. Methodology workflow WDT model development.

3.5.8 Evaluation metrics

The following metrics will be used to compare the modeling results for the developed models: DT methods such as AdaBoost, Gradient Boosting, XGBoost, and CatBoost along with KNN and LR approaches.

3.5.8.1 *Parity plot*

Parity plot is a cross plot that can be used to compare the predicted WDT values with actual measured WDT values. The plot can provide information on model accuracy and suggests systematic bias is present if data is evenly shifted away from diagonal line, assesses model's performance across the range for any non-consistent behavior, and helps to detect if any outliers are present in the dataset. The data plotted along the 45 degrees line suggests model's high performance. Scattered and deviated data suggests errors, specifically the scattered data with no pattern suggests predictions are random and are not reliable. Non-linear pattern may require feature revisit or non-linear modeling. If the parity plot shows model underperformance, then the following should be investigated: Revisiting the input parameters, possibly some variables were missed out. Revisiting model settings and tuning the hyper-parameters. If a non-linear trend is observed, then DT methods such as XGBoost and CatBoost or neural networks might offer a better fit. Outliers could be addressed with either removal from the dataset or log transformations.

3.5.8.2 *Residual plot*

Residual plot is an error distribution graph that can be used to visualize how errors are distributed across the value range. The plot helps to assess model performance and select appropriate solution for improvement. Errors scattered away from 0 line suggest presence of a systematic bias. If errors or residuals are distributed evenly, then the data is good for linear regression. Uneven behavior might require data transformation. If a pattern is observed in error distribution, then non-linear method should be selected. Large errors suggest presence of outliers. An acceptable residual plot with well-behaved errors will have errors spread on zero line without any pattern or curve and variance across predictions will remain constant. A non-acceptable residual plot where model issues are encountered might display the following patterns: U-shaped curve suggests a non-linear relationship, a cone-shaped error distribution suggests heteroscedasticity that requires log transformation, clusters or trends in error

distribution suggest biased model and may require input parameter revision or different method selection, and outliers if present should be addressed.

3.5.8.3 *Standardized Residuals vs. Leverage plot*

In Standardized Residuals versus Leverage plot also known as Williams' plot, standardized residuals are errors scaled by standard deviation on y-axis and leverage on x-axis indicates contribution of data on the model. Chart helps to detect outliers (residuals high on y-axis) and identify high contributors (high leverage on x-axis). Residual limits are set +/-3. Threshold limit is set at twice input parameter type divided by number of data points.

3.5.8.4 *SHAP graph*

SHAP graphs are popular with black box methods, such as XGBoost, CatBoost, Deep learning. SHAP provides best understanding of the model processing, why would this model make this specific prediction by quantifying the contribution of each feature on end result. This approach also helps to detect bias in the model. There are different types of SHAP graphs, such as, Bar, Summary, Waterfall, Force and Dependence plots. Bar Plot ranks feature importance helping to screen reducing those low contributing features, Summary Plot provides overview of each feature importance thus identifying key contributors, Waterfall Plot provides detailed breakdown of a single prediction instance SHAP value, Force Plot visualizes individual predictions, thus explaining a single prediction instance, Dependence Plot helps to understand interactions between features. In scope of subject work it is proposed to limit evaluation metrics to Bar and Summary Plots for XGBoost and CatBoost.

4 RESULTS AND DISCUSSION

To recap from previous sections, two the input parameters were selected for WDT modeling: molar weight and pressure, as these include effects of fluid composition, system pressure, and intermolecular forces on WDT. The literature has been reviewed and a dataset with total of 380 datapoints was built from published studies, where each data source was thoroughly assessed as detailed in Chapter 3. Of this dataset 80% of data has been used for training and 20% for testing processes, a very common and effective data splitting strategy. Six different MD algorithms were applied for WDT modeling: KNN, LR, and four DT methods with Boosting mechanisms such as AdaBoost, Gradient Boosting Machines, XGBoost, and CatBoost. Hyperparameter tuning was applied on the latter four models and optimal values were selected as presented in Table 3.6.

In this section the results of data modeling work are presented. First, statistical, and graphical methods for model performance comparison is presented. Then, outlier detection is demonstrated using Williams' plot and bias and each input feature contribution is assessed with SHAP graphs. Additionally, model consistency and reliability are assessed by performing trend analysis, where sensitivity to molar weight at different pressures is analyzed. As well, a GEP model was developed similar to the GEP model presented by Nait-Amar et al. (2024). Lastly, the select models were compared with ideal solid and thermodynamic models such as multi-pure solid and Coutinho's UNIQUAC models and previous studies with ML models by Bian et al. (2019) – GWO-SVM, Benamara et al. (2019) – RBFNN-ABC, Amiri-Ramsheh et al. (2021) – RF, and Nait-Amar et al. (2024) – GEP.

4.1 Statistical performance analysis

The data modeling results were evaluated using the following statistical metrics: coefficient of determination, R^2, root-mean-square error, RMSE, and average absolute relative deviation, AARD. Metric calculation formulas are provided in the text below:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(WDT_i^{exp} - WDT_i^{calc})^2}{\sum_{i=1}^{N}(WDT_i^{calc} - \overline{WDT})^2} \quad \text{Equation 4.1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(WDT_i^{exp} - WDT_i^{calc})^2} \quad \text{Equation 4.2}$$

$$AARD, \% = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{WDT_i^{exp} - WDT_i^{calc}}{WDT_i^{exp}}\right| x\ 100 \quad \text{Equation 4.3}$$

where, N is the total number of measurements, WDT^{exp} is the measured value, WDT^{calc} is the calculated value, and $\overline{WDT}$ is the average WDT value. The results are tabulated in Table 4.1 for training part, testing part, and complete dataset.

Table 4.1. Statistical Performance Analysis.

Metrics		KNN	LR	AdaBoost	GBM	XGBoost	CatBoost
R^2	Train R^2	0.8792	0.498	0.9996	0.9996	0.9988	0.9996
	Test R^2	0.8758	0.5959	0.9667	0.9657	0.9806	0.9765
	Overall R^2	**0.8791**	**0.5153**	**0.9945**	**0.9943**	**0.996**	**0.996**
RMSE	Train RMSE	4.4334	9.0374	0.2465	0.2446	0.4381	0.2603
	Test RMSE	3.8911	7.018	2.0139	2.0436	1.5365	1.6917
	Overall RMSE	**4.3303**	**8.6712**	**0.9272**	**0.9397**	**0.791**	**0.7916**
AARD, %	Train AARD	0.9628%	2.3070%	0.0309%	0.0288%	0.0930%	0.0485%
	Test AARD	0.7928%	1.8157%	0.4154%	0.4175%	0.3373%	0.3560%
	Overall AARD	**0.9288%**	**2.2087%**	**0.1078%**	**0.1066%**	**0.1419%**	**0.1100%**

When evaluating R^2, both CatBoost and XGBoost models are most accurate with almost perfect fit, and AdaBoost with GBM demonstrate excellent performance, with moderate performance for KNN and very poor performance on LR. For RMSE, DT models perform superior to other techniques, CatBoost and XGBoost models demonstrating lowest error, AdaBoost and GBM acceptable low error values, while the KNN model shows high error and LR shows worst performance. This suggests both KNN and LR models are not fit for this regression task. For AARD, all DT models show low relative accuracy; while, the GBM model demonstrates best performance, the CatBoost model still offers best overall balance across all metrics.

Summarizing statistical evaluation of model performances: The CatBoost model is ranked top performer offering combination of accuracy and stability. The XGBoost model is offered second best performance with R^2 and RMSE equivalent to CatBoost model and slightly lower performance on AARD. The GBM model offers best AARD; while, demonstrating high performance on R^2 and RMSE. The AdaBoost offer high performance and can be a model of choice for simplicity without loss in performance. The KNN model has tolerable performance; while struggles with accuracy compared to top performers. The LR model should be avoided at all costs as could not address complex relationships of the model.

4.2 Parity plot

The parity plots (cross-plots) of estimated WDT versus measured WDT values for the 6 developed intelligent models: LR, KNN, AdaBoost, GBM, XGBoost, and CatBoost are presented in Figures 4.1 to 4.6. Parity or cross-plots are useful for graphical evaluation of model performance. Predicted values are plotted against actual WDT values and the distribution is compared to unit-slope ideal line. The closer the plotted data points are located to ideal unit-slope line the higher is uniformity of the distribution and reliability of the model. These plots

can also be analyzed for data spread and clustering and give insight into consistency comparing training and test data.

The LR model's distribution has a wide scattering, especially on the training data. And analyzing the distribution range, higher values are systematically underestimated. The model fails to handle complex data and is very inaccurate. The KNN model demonstrates slightly better performance than the LR model with better alignment across unit-slope line, yet still high spread, and similar to the LR model shows tendency of underfitting at higher WDT values. The AdaBoost model output sit on ideal line both for train and test data, with very minimal scatter suggesting strong model with minimal errors. The GBM model's performance is similar to the AdaBoost, very tight clustering data lies on unit-slope line, and little to no bias/overfitting. An acceptable, robust, and consistent model. The XGBoost model exhibits the best performance with best alignment across unit-slope line. Very little, an almost negligible deviation. Ideally fits across whole range. The CatBoost model shows identical performance to the XGBoost, with data sit well on ideal line. Overall, the DT-based models demonstrate superior performance in handling non-linear patterns in WDT modeling.

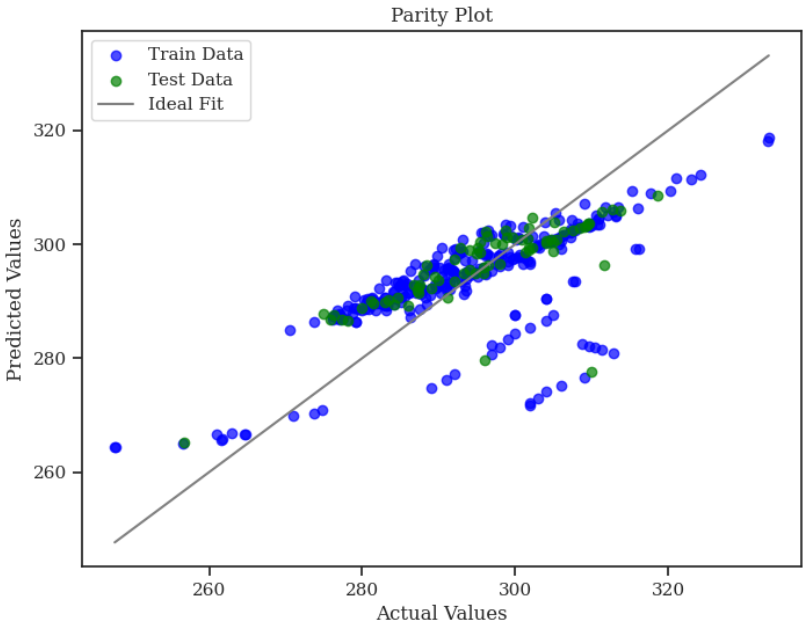

Figure 4.1. Parity plot for the developed LR model.

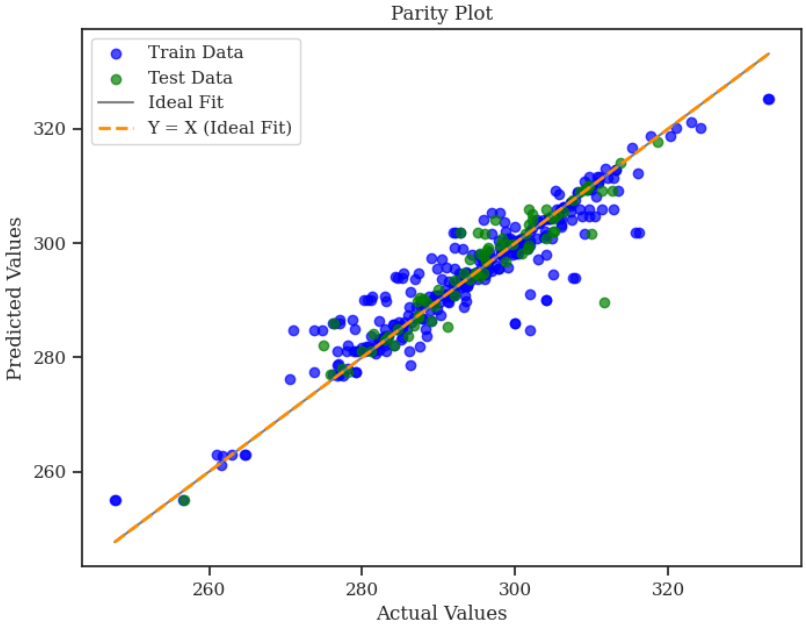

Figure 4.2. Parity plot for the developed KNN model.

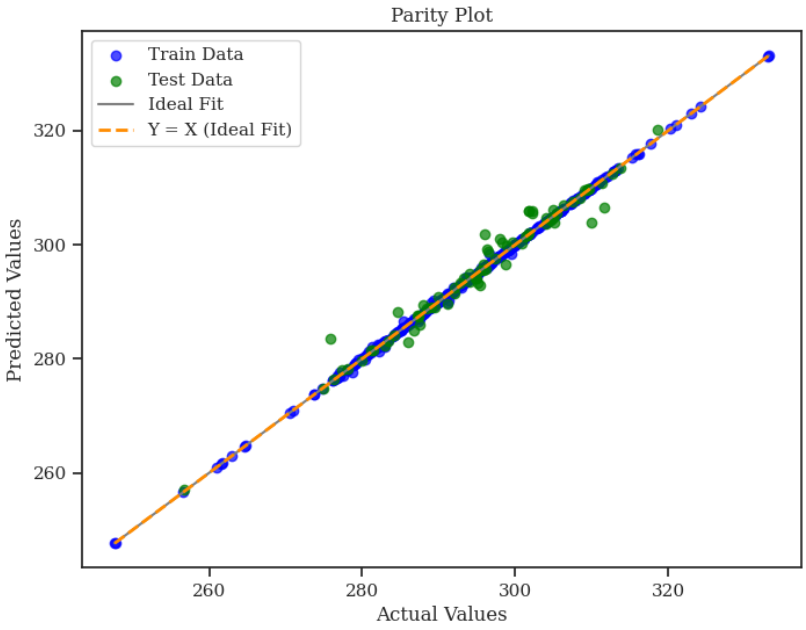

Figure 4.3. Parity plot for AdaBoost.

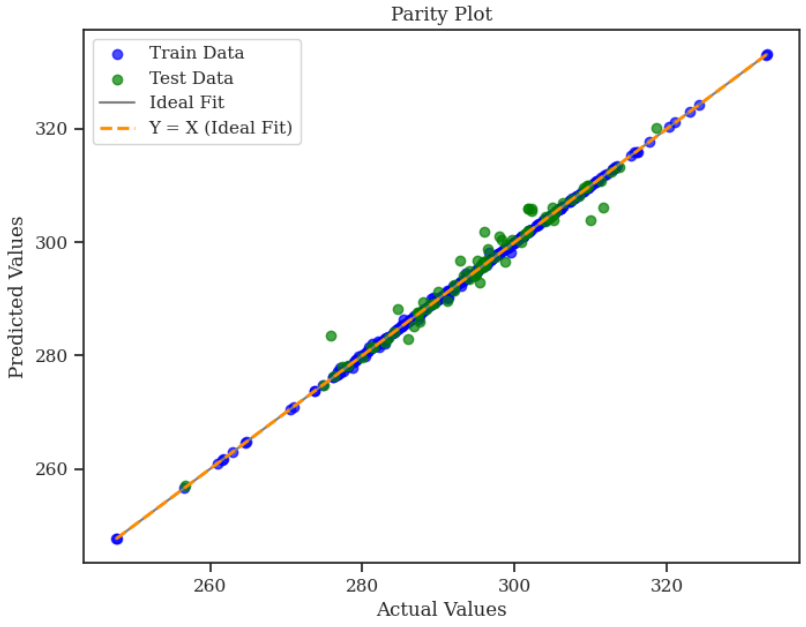

Figure 4.4. Parity plot the developed GBR model.

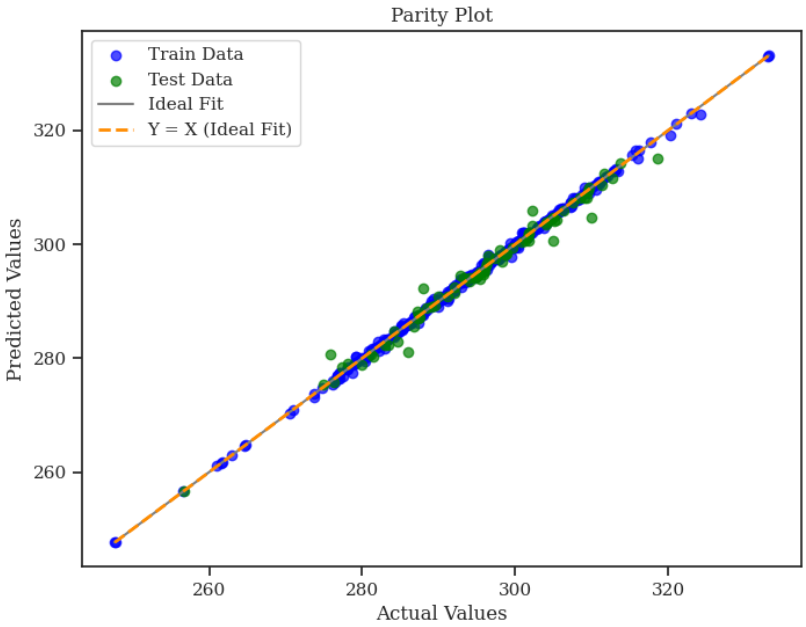

Figure 4.5. Parity plot for the developed XGBoost model.

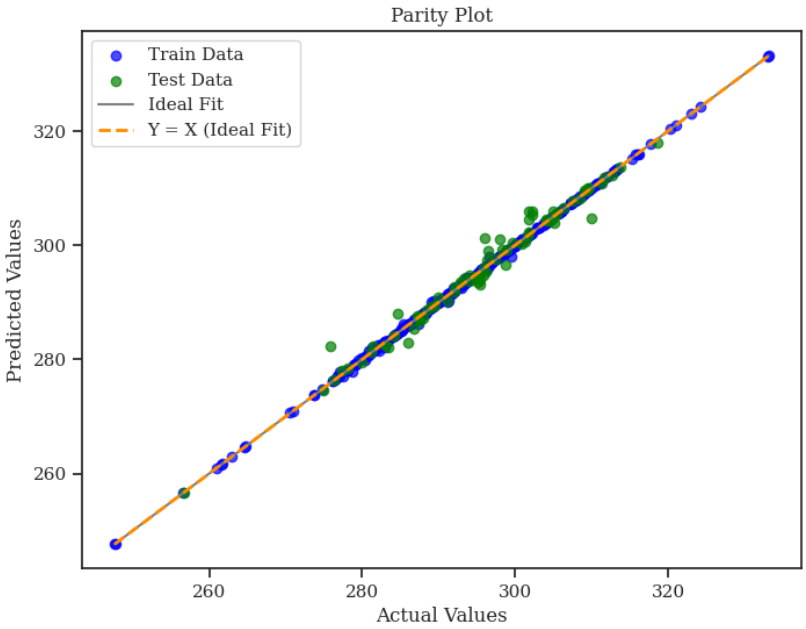

Figure 4.6. Parity plot for the developed CatBoost model.

4.3 Residual plot

Residual plots also known as error distribution graphs generated for the 6 developed ML models in this research work: LR, KNN, AdaBoost, GBM, XGBoost, and CatBoost are presented in Figures 4.7 to 4.12. These plots of relative deviation between actual and estimated WDT versus measured WDT values illustrate how errors are distributed across the whole range of WDT values. Uniform distribution across zero-error line would prove model accuracy and reliability. Even distribution across the range would suggest data fit for linear regression. Trend or pattern observed in the distribution would require non-linear methods. Outstanding deviation would suggest presence of outliers in dataset.

Analyzing the error distribution graphs suggests that the developed LR model error distribution indicates clear pattern presence and error increasing with increasing WDT values. The graph suggests that the model is biased and not fit for this non-linear task. The KNN plot shows more scattered error distribution, but this is random distribution, and no trend or pattern is observed, still some clustering can be observed, and much wider variance compared to the LR model. Deviations are higher in lower WDT values. The KNN model offers better performance than the LR model, but still lacks consistency and performance is unreliable. The AdaBoost model demonstrates error distribution centered around zero line, with very low deviation, without any clear trend or pattern suggesting good data generalization, minor outliers are present but overall good fit of data. The GBM model's errors are evenly plotted around zero line, with excellent consistency both for train and test data. The GBM model displays similar performance as the AdaBoost but more uniform distribution suggesting a strong, stable model with minimal bias. The XGBoost model demonstrates best performance with flat and tight data distribution across zero line, with no systematic error. Train and test data almost in line with each other. Reliable and well-calibrated model. The CatBoost model displays a well-

centered distribution similar to the XGBoost model but little higher deviations, with few outliers, though not systematics, suggesting second best performer. Clearly, the DT-based models demonstrate a far better error distribution, the KNN model shows inconsistent and less reliable distribution, while the LR model exhibit worst performance suggesting method not fit.

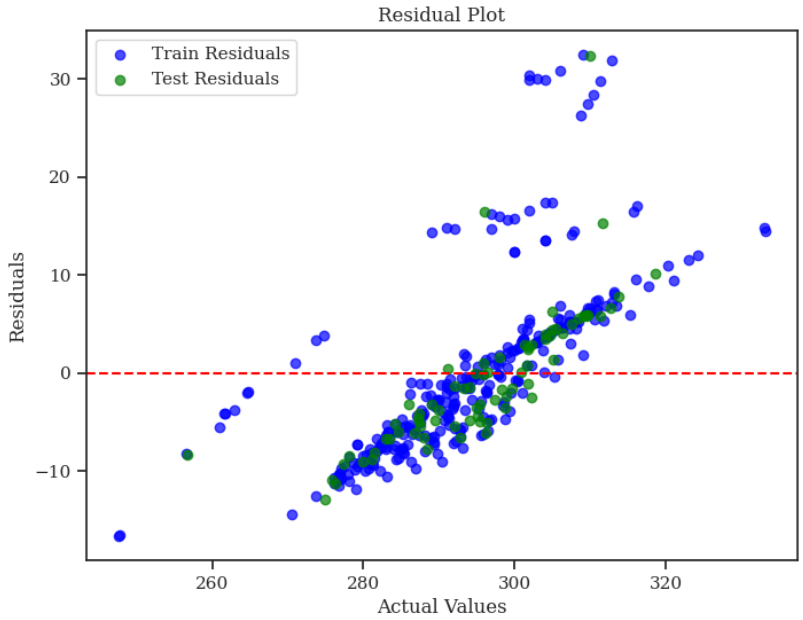

Figure 4.7. Residual plot for the developed LR model.

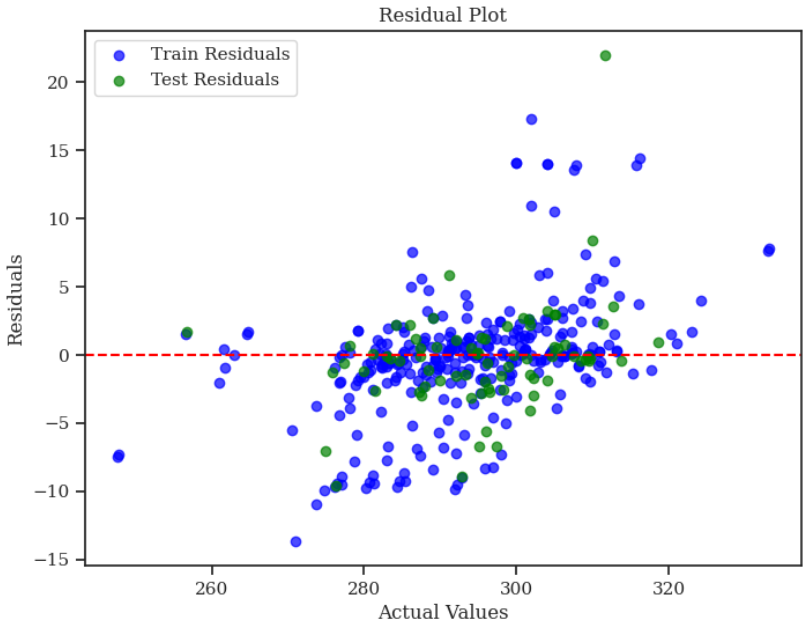

Figure 4.8. Residual plot for the developed KNN model.

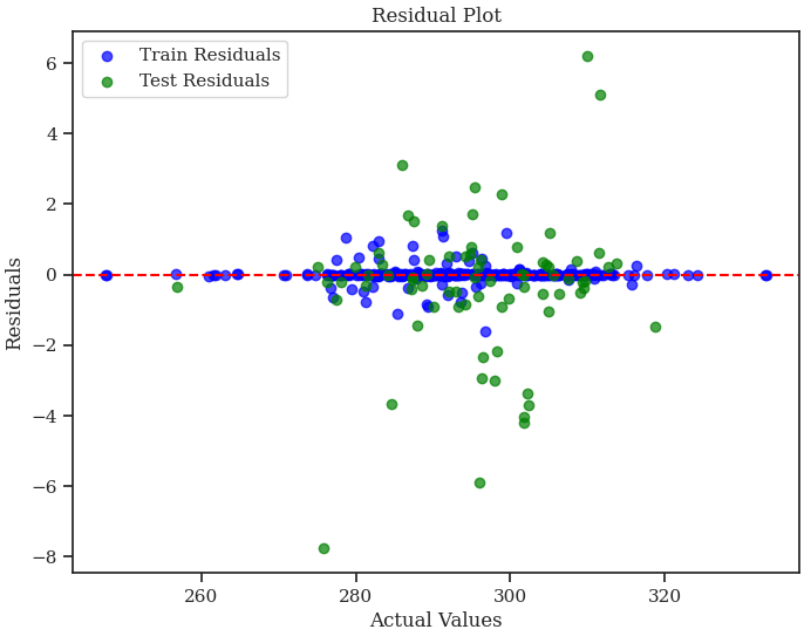

Figure 4.9. Residual plot for the developed AdaBoost model.

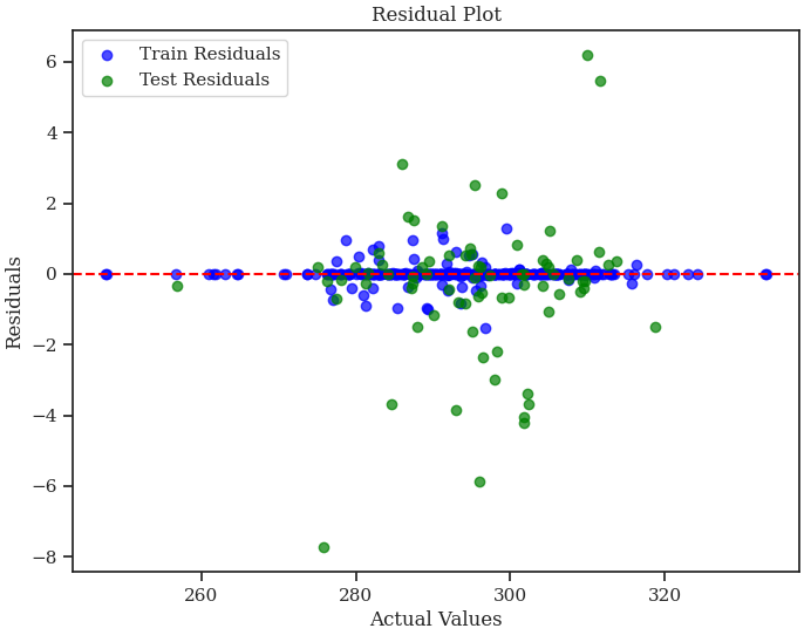

Figure 4.10. Residual plot for the developed GBM model.

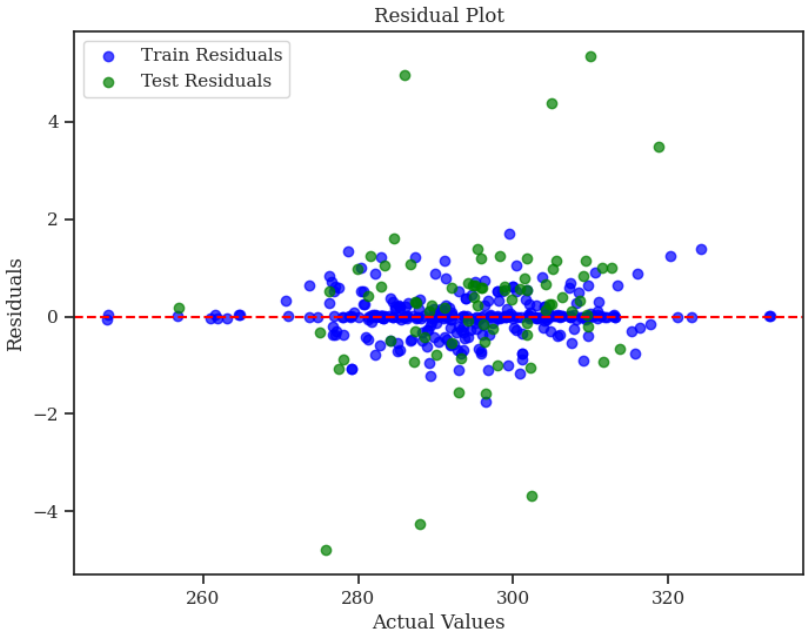

Figure 4.11. Residual plot for the developed XGBoost model.

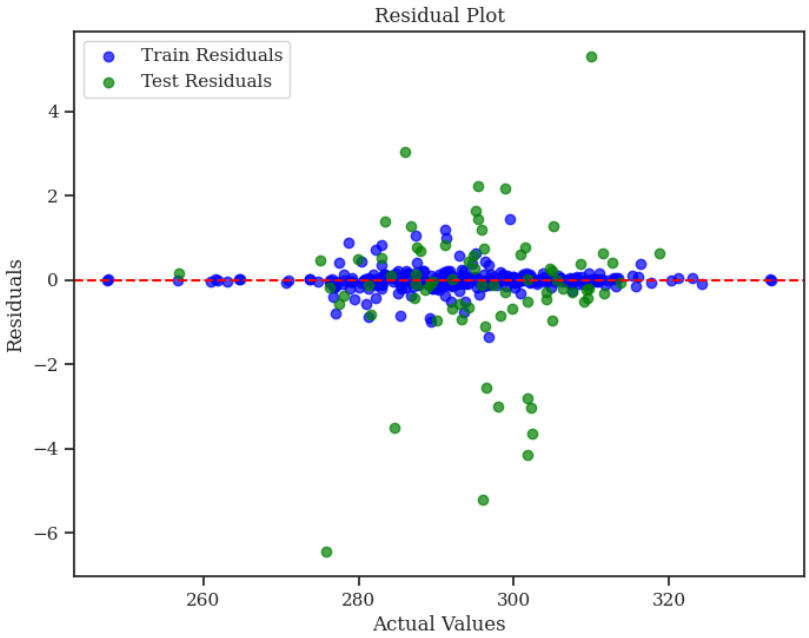

Figure 4.12. Residual plot for the developed CatBoost model.

4.4 Williams' plot

Both XGBoost and CatBoost were concluded as top performers from statistical analysis and parity and error distribution plots. Standardized residual versus Leverage plot (Williams' plot) was used to compare the XGBoost and CatBoost models as presented in Figures 4.13 and 4.14.

These plots demonstrate error scaled by standard deviation on y-axis vs their contribution to the model on x-axis with set threshold ranges +/-3. The plots allow to identify outliers and data with high impact in regression models and are used for model validation and diagnostics.

The XGBoost model shows the most errors within +/-3 threshold limits suggesting stable prediction errors, 3 data exceed the limit (+/-4) which is negligibly low and 11 high leverage data that are of high impact but are at zero line, i.e. no error, suggesting good control at high leverages. This suggest that the XGBoost model as a robust and reliable model. The CatBoost model shows slightly more residual outliers exceeding threshold limit, reaching +6. High leverage points are same as in the XGBoost model and are too at zero line which suggests good control too. Similar to the XGBoost model's performance but with more outliers.

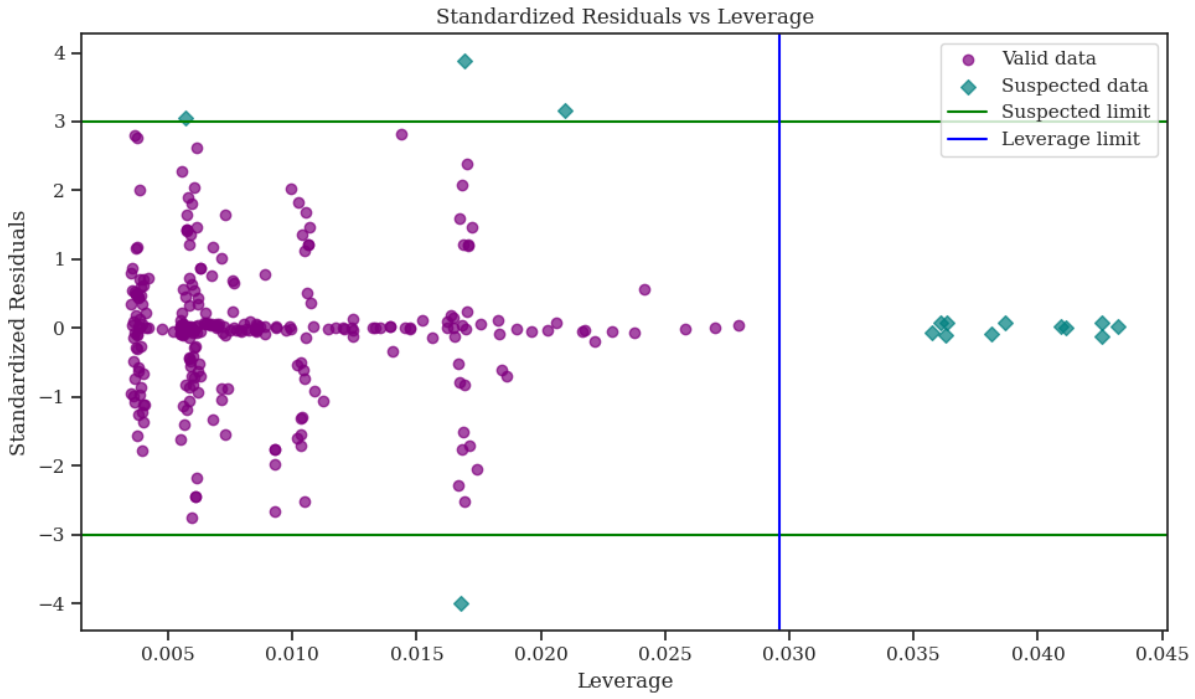

Figure 4.13. Standardized Residuals vs. Leverage plot for the developed XGBoost model.

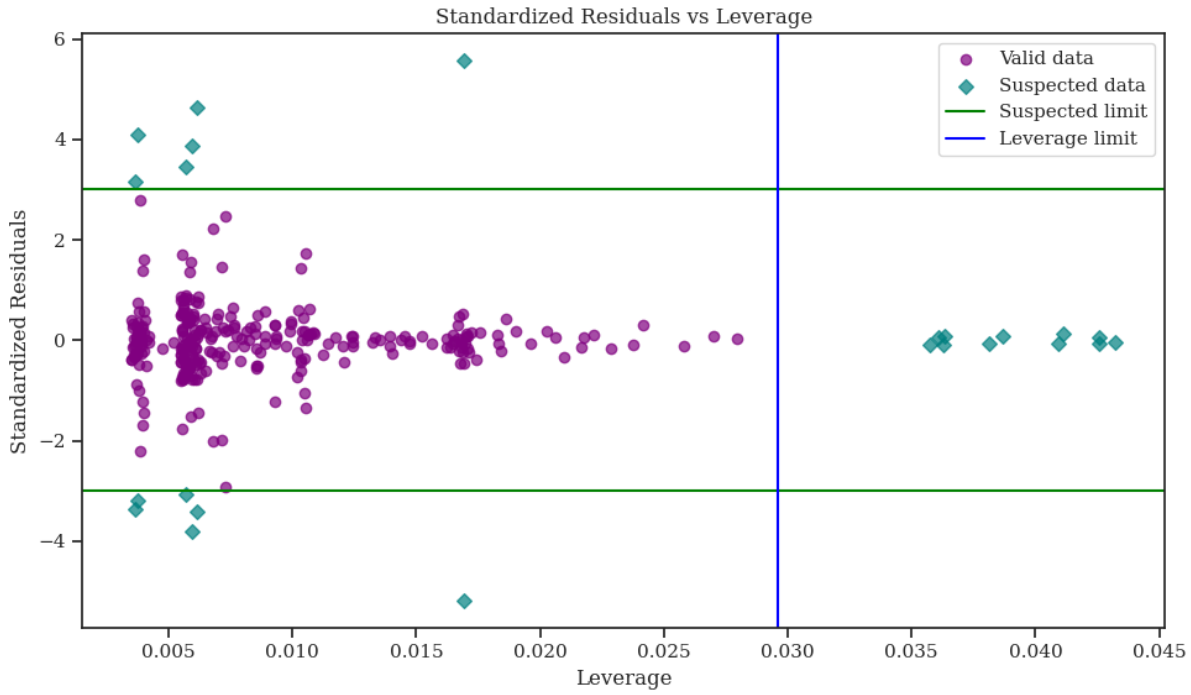

Figure 4.14. Standardized Residuals vs. Leverage plot for the developed CatBoost model.

4.5 SHAP plots

Two types of SHAP plots were used to assess the impact of each of the two features on the XGBoost and CatBoost on output modeling: Bar plots and Summary plots. The SHAP bar plots to analyze MW and P impact on WDT for the developed XGBoost and CatBoost models are presented in Figure 4.15 and Figure 4.16.

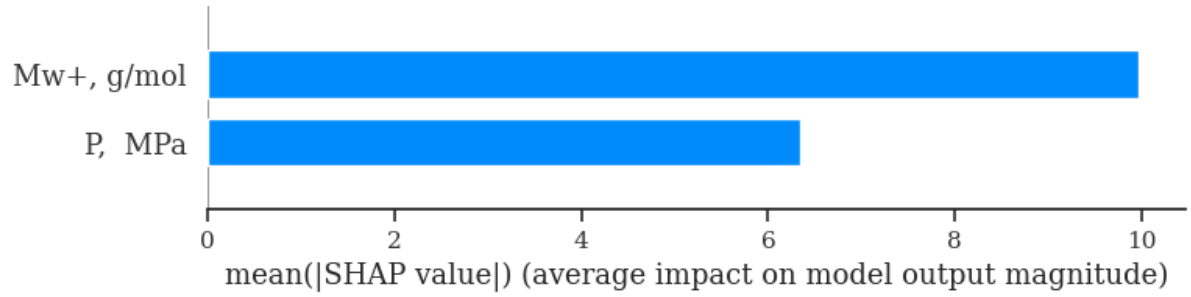

Figure 4.15. SHAP feature importance bar plot for the developed XGBoost model.

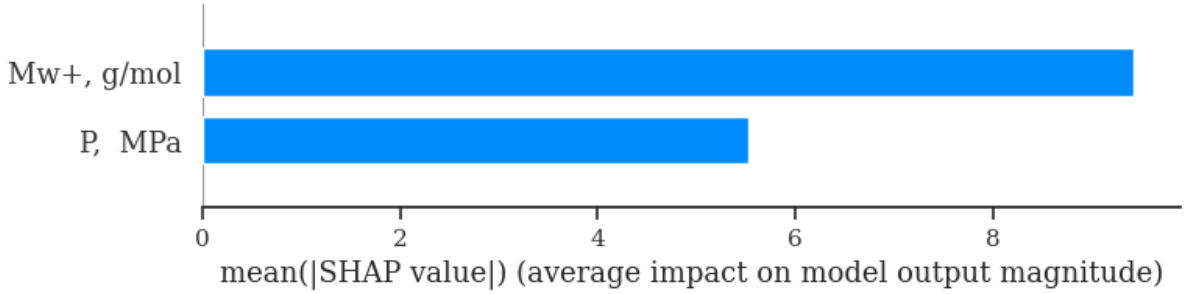

Figure 4.16. SHAP feature importance bar plot for the developed CatBoost model.

Both XGBoost and CatBoost confirm MW as most critical input feature (XGBoost ~ 10, CatBoost ~ 9.2) and see P as important feature but not as influential as MW. Here, the XGBoost model allocates more importance to P (~ 6.2) compared to the CatBoost (~ 5.6), that suggests that the XGBoost may model more nuanced interaction with P than the CatBoost.

SHAP Summary plot to illustrate distribution and direction of impact for the developed XGBoost and CatBoost models are presented in Figure 4.17 and Figure 4.18. Both XGBoost and CatBoost models show MW domination, high MW showing strong positive impact on WDT and low MW demonstrating strong negative impact with high influence of range +30/-40 SHAP. Both models show P as important feature, high P increasing WDT and low P slightly reducing WDT. Element of non-linearity is present both in MW (high value impact range of -15 to 30, SHAP is likely taking median as borderline for low/high feature value, this would explain non-linearity) and P (again P distribution suggest data clustering around atmospheric pressure that too would explain non-linearity, uneven spread -10 to 30).

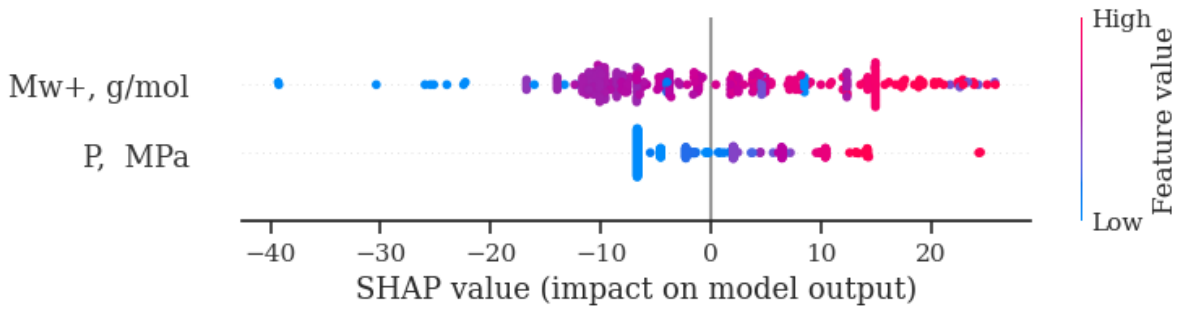

Figure 4.17. SHAP summary plot for the developed XGBoost model.

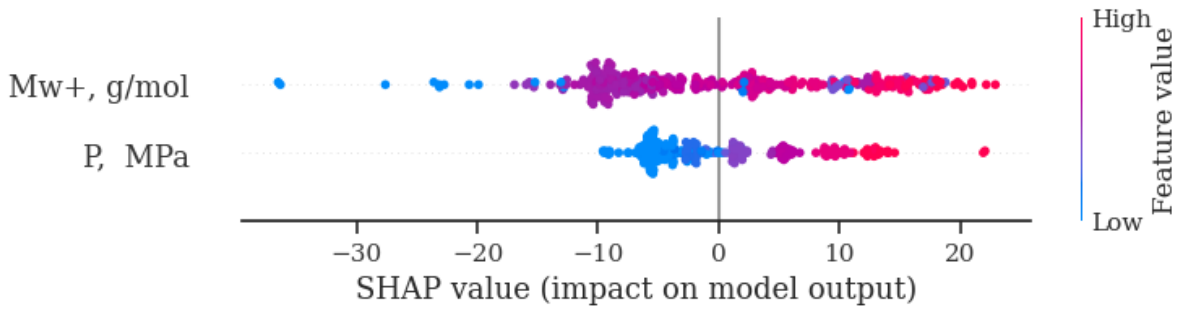

Figure 4.18. SHAP summary plot for the developed CatBoost model.

4.6 Trend Analysis

Trend analysis was performed to assess physical validation of the developed models by plotting WDT (laboratory data versus the XGBoost and CatBoost modeling outputs) vs. MW at different pressures ranging of 0.1 MPa, 20 MPa, 40 MPa, 60 MPa, 80 MPa, 100 MPa as presented in Figures 4.19 to 4.24. the trends suggest a strong dependance of WDT on MW; where, the WDT increases with higher MW. Physical justification of this observation is that higher MW results in higher viscosity that leads to higher WDT. The weaker dependance of WDT onto P is observed, where WDT seems to increase with higher P, which contradicts with earlier statements from different studies such as Bellarby (2009), Juyal et al. (2011), Pauly et al. (2000) that suggest higher pressure leads to lighter components dissolving that would suspend wax particles in solution reducing wax precipitation that's lowering WDT. This observation is supported by conclusion drawn from SHAP summary plot analysis where higher P was concluded to increase WDT.

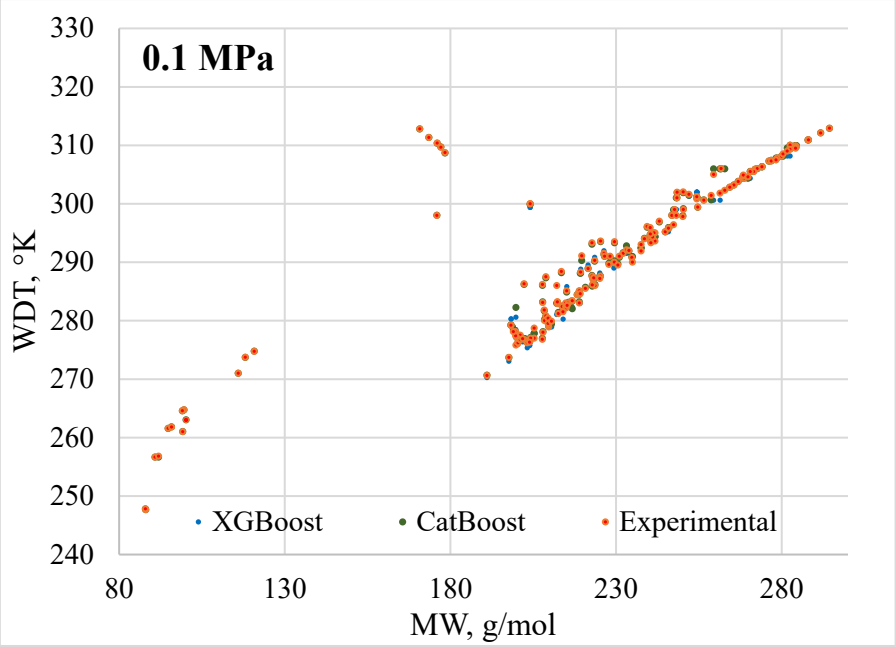

Figure 4.19. Trend Analysis: WDT vs. MW at 0.1 MPa.

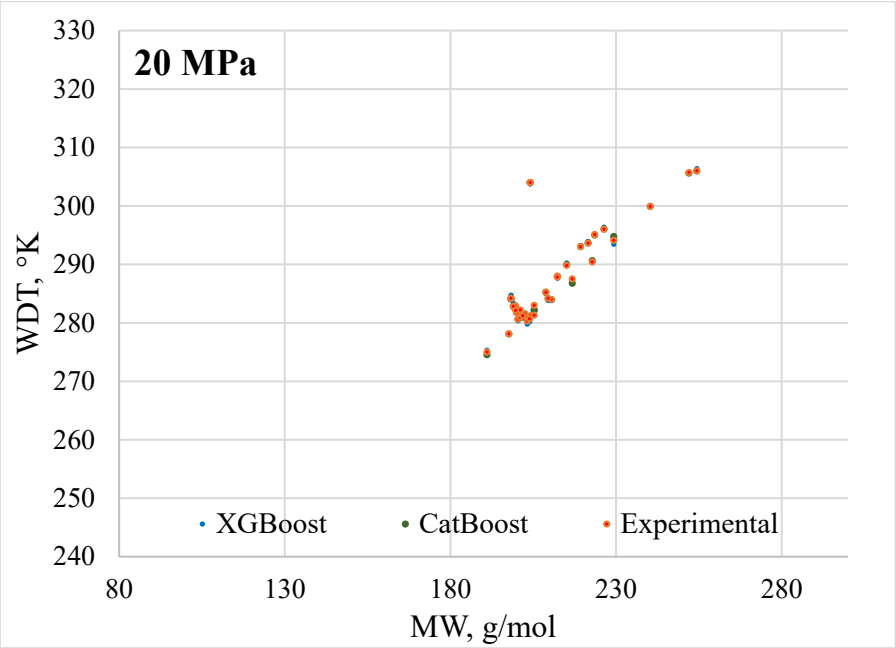

Figure 4.20. Trend Analysis: WDT vs. MW at 20 MPa.

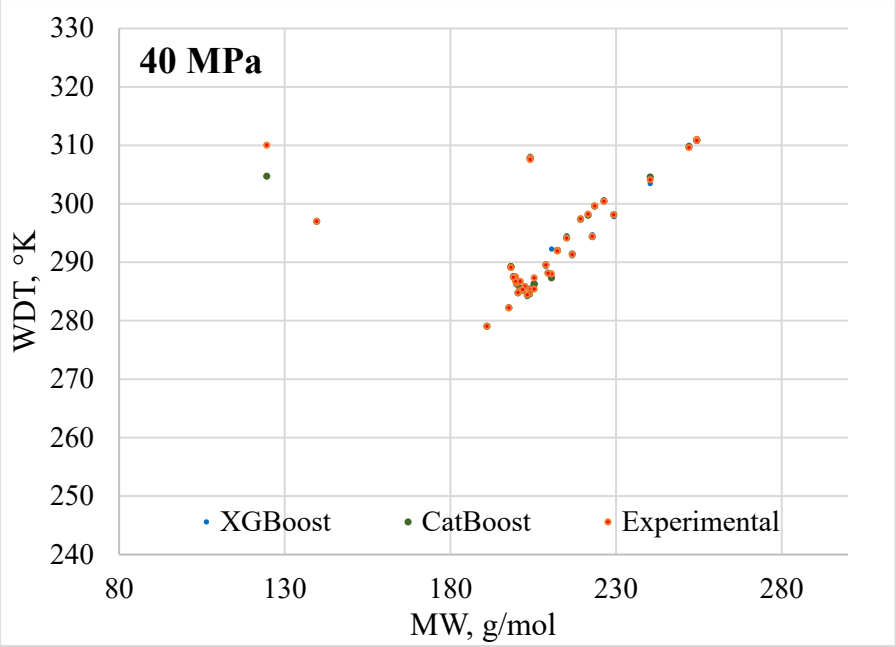

Figure 4.21. Trend Analysis: WDT vs. MW at 40 MPa.

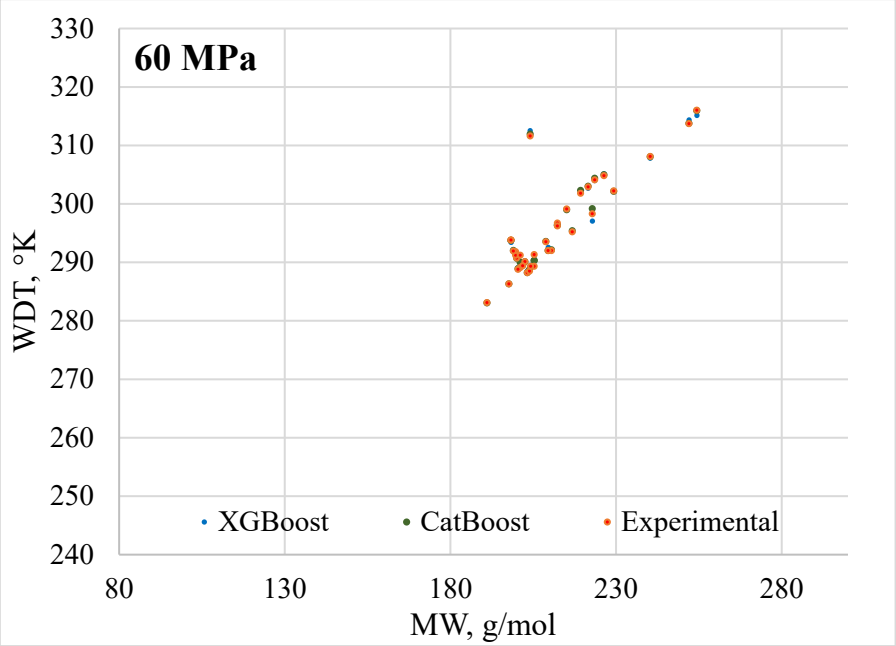

Figure 4.22. Trend Analysis: WDT vs. MW at 60 MPa.

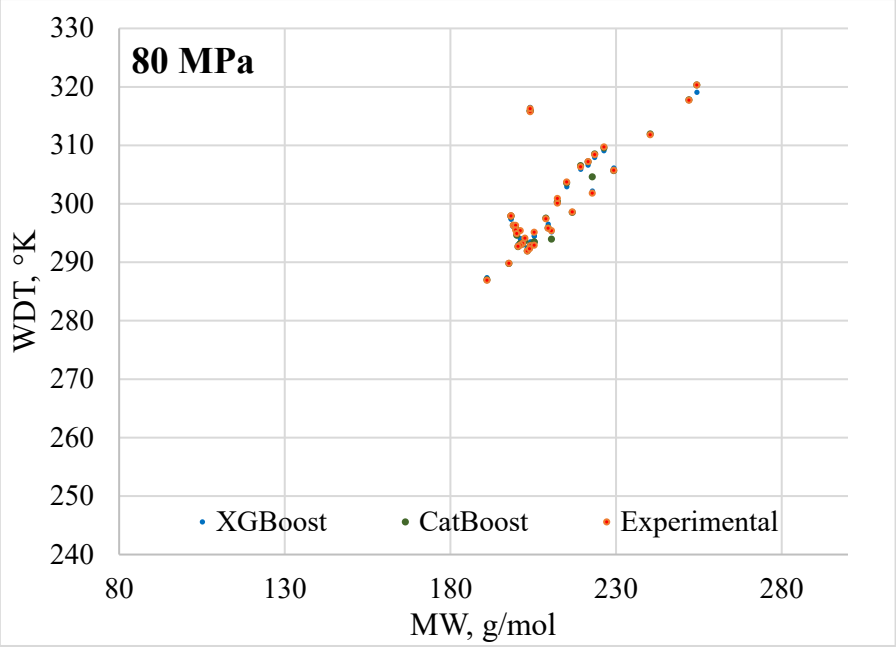

Figure 4.23. Trend Analysis: WDT vs. MW at 80 MPa.

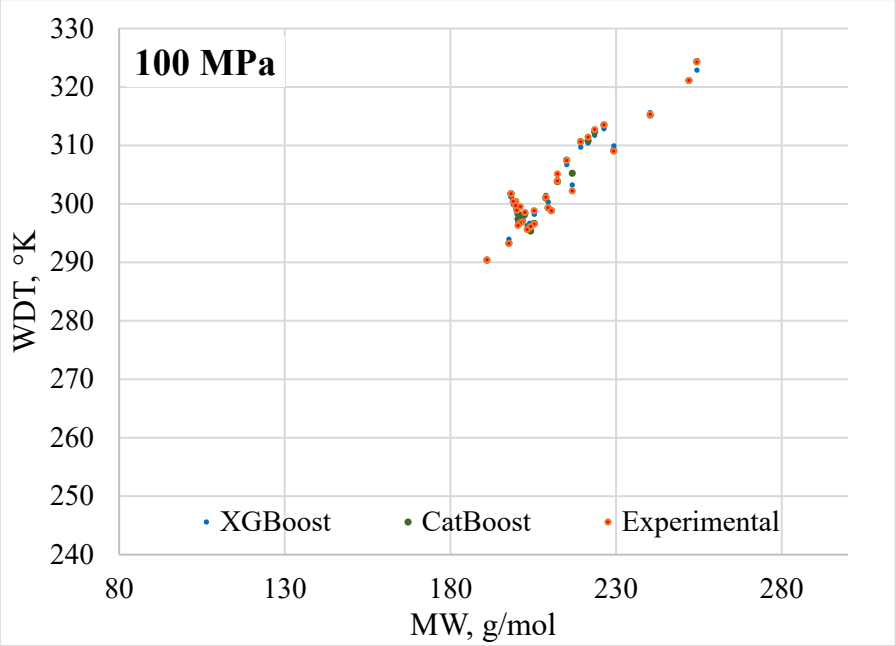

Figure 4.24. Trend Analysis: WDT vs. MW at 100 MPa.

4.7 GEP model development

Nait Amar et al. (2024) have developed white box model to estimate WDT with clear and explicit formulation as function of two parameters: MW and P. Nait-Amar et al. (2024) suggested that the tuning process concluded a need to split database into two groups according

to MW values, MW ≤ 210 g/mol and MW > 210 g/mol. A database similar to previous chapter with 272 data points was compiled and used with data reported by Milhet et al. (2004) – 144 data points, Daridon et al. (2002) – 54 data points and Ji et al. (2004) – 74 data points.

In this research work, GEP was used to develop a correlation. Two scenarios were modeled where the data was managed with one dataset and data was managed in two sub-sets of a dataset, similar to Nait Amar et al. (2024). The latter is supported by the input data analysis where Violin plot for MW data suggested bi-modal distribution hinting to 2 categories in dataset in terms of MW. The median of MW distribution was used as borderline 209.686 g/mol (~210 g/mol). The database used for the model was same as in previous chapter of 273 data points versus Nait Amar work with 272 data points with 80/20 split for training and testing. For database with 1 dataset, 218 data points were used for training and 55 data points for testing, for database with 2 sub-sets: first half of data, H1 (less than 210 g/mol) covers 137 data points of which 110 were used for training and 27 were used for testing purposes, second half of the data, H2 (more than 210 g/mol) covers 136 data points of which 109 were used for training and 27 for testing purposes. The background settings both reported by Nait Amar et al. (2024) and ones used for new GEP model development are presented in Table 4.2 and Figure 4.25, respectively.

Comparing two scenarios, one with dataset without splitting shows better and more consistent results than one with dataset split into two sub-sets. The newly developed GEP correlation is compared with the GEP correlation proposed by Nait-Amar et al. (2024) and the GMDH model proposed by Benamara et al. (2019) as shown in Table 4.3. As one can see from Table 4.3, performances of both scenarios of the newly develop GEP correlation unfortunately lag slightly behind both the GEP correlation proposed by Nait Amar et al. (2024) and the GMDH proposed by Benamara et al. (2019). Further tuning of the model is required.

Table 4.2. Control parameters for the developed GEP correlation.

Models	GEP by Nait Amar et al. (2024)	Current GEP 273 w one dataset	Current GEP 273 w two sub-sets – H1	Current GEP 273 w two sub-sets – H2
Individuals	500	500	150	300
Head size		7	7	7
Number of genes	10	5	3	3
Mutation rate	0.5	See Figure 4.25 below		
Cross over rate	0.95			
Inversion rate	0.06			
Operators	+, -, x, /	+, -, x, /, exp, log, $\sqrt[3]{}$		

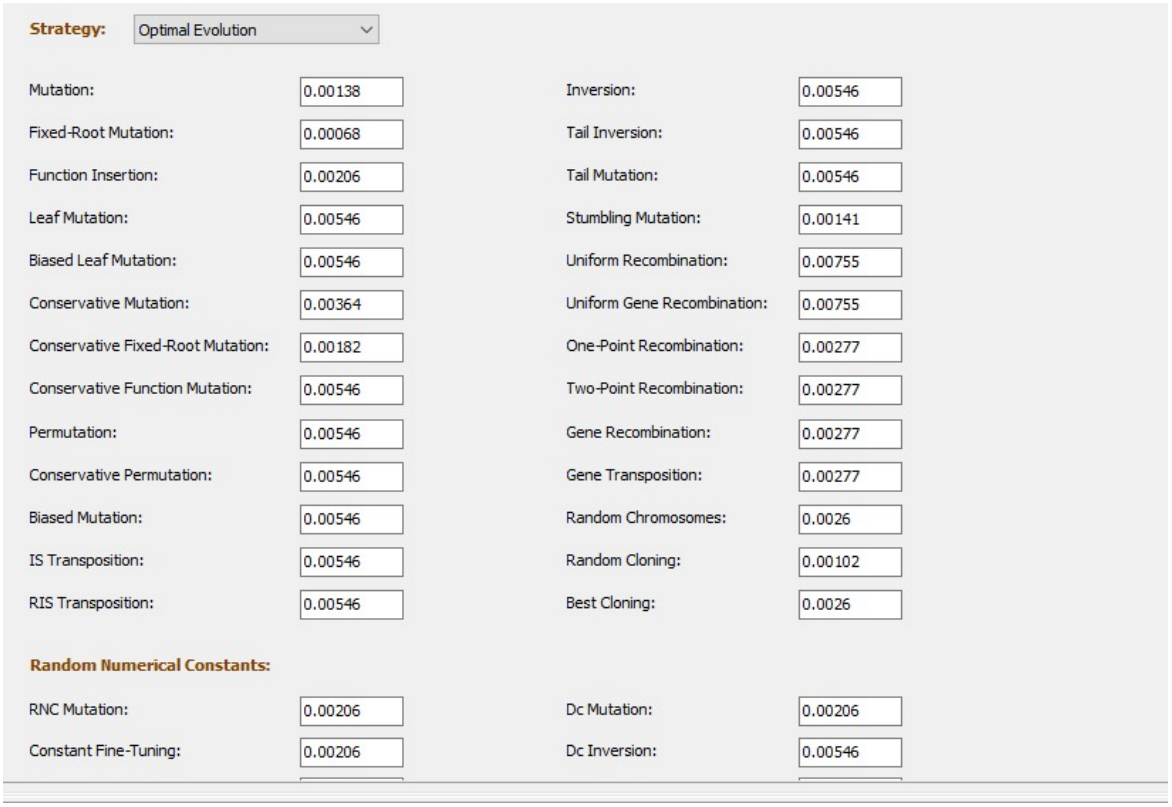

Figure 4.25. Background settings for the developed GEP correlation.

The resultant correlations are presented as below:

- for the database with 1 dataset:

$$WDT\ (MW,P) = [exp(exp(exp(\frac{1}{log(MW)+MW})))]^2 + [tan(C_1) + tan(\frac{C_1\sqrt[3]{MW}}{C_2})] +$$

$$tan(tan(C_3)) + [\frac{1}{exp(log(\frac{C_4+C_5}{P*C_6}))}] + [\sqrt[3]{MW} + \sqrt[3]{C_8} * MW + C_7] \quad \text{Equation 4.4}$$

C_1 = - 0.424889875745345

C_2 = -0.511447789595117

C_3 = 8.43338821009264

C_4 = 0.244103307114127

C_5 = 6.34723842089192

C_6 = 1.30549056118889

C_7 = 6.31333725846668

C_8 = 2.8082674121035 x 10^{-2}

- for the database with 2 sub-sets: first half of data, H1 (less than 210 g/mol) covers 137 data points of which 110 were used for training and 27 were used for testing purposes,

  for MW ≤ 210 g/mol:

$$WDT\ (MW,P) = ((\frac{C_1}{P} + P) + (\frac{C_2}{MW}) - C_3)^2 + ((C_4 - C_5) - (\frac{1}{(MW-C_6)})^2) + (\frac{{C_7}^2 * C_8}{C_8+MW})$$

Equation 4.5

C_1 = −1.81006409136559

C_2 = −2.07270242373579

C_3 = 8.74193263706319

C_4 = 4.50036089194547

C_5 = −9.98950428090209

C_6 = 7.3034528713825

C_7 = 13.0531483346575

C_8 = −8.70887930966409

- for MW > 210 g/mol:

$$WDT\ (MW,P)\ =\ C_1\ -\ (C_1 -\ log(\frac{exp(MW)*P}{log(10)})) + (C_2 - {C_4}^2) * C_3 + \frac{log(P)}{(\frac{\sqrt[3]{P}}{P} log(MW))}$$

Equation 4.6

C_1 = −8.28541116442893

C_2 = 5.29943081720178

C_3 = −9.44834942636326

C_4 = 5.28253934593358

Table 4.3. The developed GEP correlation's comparison with previous GMDH and GEP correlations.

Models	R^2	RMSE	AARD, %
GMDH, Benamara et al. (2019)	0.9582	2.3805	0.6888
GEP, Nait Amar et al. (2024)	0.9647	2.1963	0.5963
New GEP (Scenario 1 w/ one dataset)	0.9434	2.8367	0.8130
New GEP (Scenario 2 w/ two sub-sets)	0.9490	2.6679	0.7097

4.8 Comparison with previous models

The comparison is based on the dataset used by Bian et al, 2019 that included data reported in atmospheric pressure by Milhet et al. (2004) – two binary systems of C_{14}+C_{15} and C_{15}+C_{16} with 144 data points, Daridon et al. (2002) – range of mixture from C_{13} through C_{24} with 54 data

points, and Ji et al. (2004) – three ternary systems of C_{14}+C_{15}+C_{16}, C_{18}+C_{19}+C_{20}, C_6+C_{16}+C_{17} and six binary systems of C_{15}+C_{19}, C_{17}+C_{19}, C_6+C_{16}, C_6+C_{17}, C_{16}+C_{18}, C_{16}+C_{20} with 74 data points. The dataset was not explicitly reported in any of these works, and attempts was made to replicate the exact same dataset for sake of comparison with results of current works. While the data reported in works of Milhet et al. (2004) and Daridon et al. (2002) were replicated exact same, it was a challenge to reinstate data reported by Ji et al. (2004), and 75 data points were used instead of 74. The total size of the dataset modeled is 273 data points. The statistical performance metrics of the intelligent models developed in this research: the developed XGBoost model and the CatBoost model versus previous models published in literature using almost the same dataset with 273 reliable data points versus 272 data points from literature are presented in Table 4.4.

Table 4.4. Comparison with the previous models.

Models	**R^2**	**RMSE**	**AARD, %**	**References**
Ideal solid	0.9525	2.9208	0.8649	Benamara et al. (2019)
Multi-pure solid	0.9514	5.0012	1.1077	Benamara et al. (2019)
Coutinho's UNIQUAC	0.9572	2.3421	0.6841	Benamara et al. (2019)
Bian et al's GWO-SVM	0.9546	2.4208	0.7128	Bian et al. (2019)
Benamara et al.'s RBFNN-ABC	0.9706	1.9969	0.5402	Benamara et al. (2019)
Benamara et al.'s GMDH	0.9582	2.3805	0.6888	Benamara et al. (2019)
Nait Amar et al.'s GEP	0.9647	2.1963	0.5963	Nait Amar et al. (2024)
XGBoost (this study)	0.9960	0.7910	0.1419	This study
CatBoost (this study)	0.9960	0.7916	0.1100	This study

The following comparative analysis can be concluded based on above results. R2 values suggest that both of the developed CatBoost and XGBoost models dominate the ranking in goodness of fit. Benamara's RBFNN-ABC is the best among earlier works. Modern ML models show better curve-

fitting and accuracy. RMSE values obtained for the modern ML models of XGBoost and CatBoost are 2.5 times less than next best performer, Benamara's RBFNN-ABC. All other previous models perform at similar range of 2-3, and multi-pure solid model shows serious lag of 5. This shows that ensemble-based models are more precise and consistent. The calculated AARD values suggest that the developed XGBoost and CatBoost perform superior to the rest with AARD minimum 4 times less than second next performer of Benamara et al, RBFNN-ABC. All of the previous and traditional models perform at range of 0.54% to 0.865% with multi-pure solid showing clear lag with AARD 1.1%. The boosting models have superior relative accuracy followed by hybrid neural-model. Overall, the developed CatBoost and XGBoost intelligent models perform better than all earlier models, both traditional and data-driven. This includes a strong performance both for prediction and generalization. Benamara's RBFNN-ABC is the strongest non-boosting approach. Traditional thermodynamic models are consistently less accurate.

4.9 Does MW sufficiently characterize fluid composition for WDT estimation?

WDT is affected by fluid composition: components, binary/ternary systems, and carbon numbers. Both the earlier modeling works, and the present study use the variables use for prediction of WDT to MW and P. While, limiting parameters eases the process it might overly simplify the models. Pedersen (1993) state that not all HC components form wax, assuming all form wax might overestimate WDT and proposed to limit the wax forming components to n-paraffins, iso-paraffins, and naphthene. However, this is not practical as detailed PNA is not available, and it is proposed to use empirical methods as proposed by Pedersen et al. (1991a). These relationships relate mole fractions to molecular weight and density with coefficients that are obtained from empirical methods. The relationship between number of carbon atoms and MW is shown Figure 4.26. The linear relationship between carbon number and MW suggests that MW can indeed represent number of carbon atoms that in turn represents fluid composition. A more detailed sensitivity analysis of data-driven methods to various fluid properties could be run to confirm if MW sufficiently characterizes the fluid composition or additional properties needs to be considered.

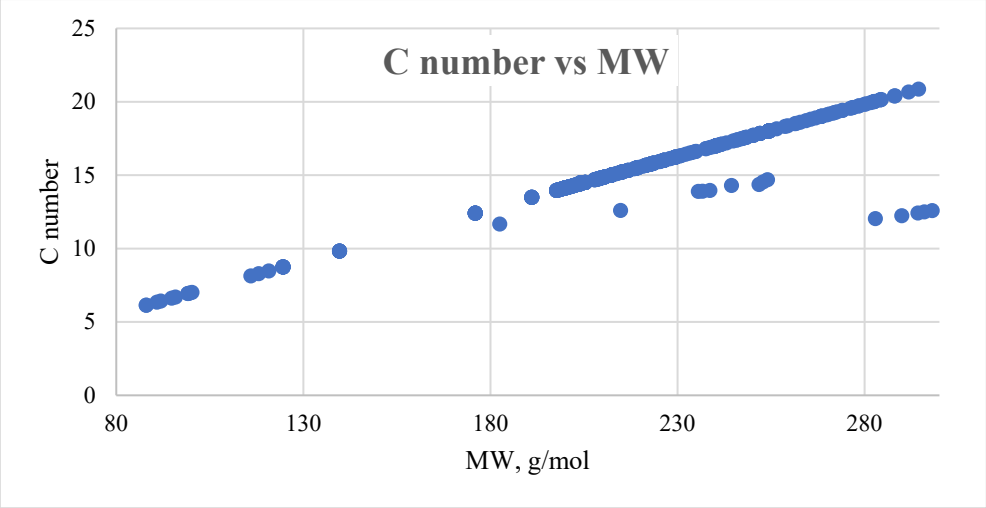

Figure 4.26. Carbon number vs. MW.

4.10 Concern of generalization of empirical and data-driven models.

According to Coutinho et al. (1995), most of thermodynamic studies are empirical correlations relating WDT to fluid properties, and they do not have the fundamental basis as they only match to laboratory data and are limited in predicting capacity. Same concern of limited generalization can be expressed for data-driven methods. When evaluating empirical methods, where researchers have analyzed various approaches and compared with obtained laboratory data, it should be noted that the method proposed by Coutinho et al. (1995), performed superior to all the empirical, even more recent, and complex correlations. These results support fundamental basis models such as ones proposed by Coutinho et al. (1995, 1996, 1998) as models of choice. However, in reality these require a lot of the input parameters and are complex which makes their application in practice very difficult. This withholds model's choice. If fundamental models such as ones proposed by Coutinho el at. (1995, 1996, 1998) could be adapted and merged with ML techniques, more strong data-driven models can be developed that both delivers high accuracy and yet is not limited in generalization.

4.11 Field data use for data-driven models

Current studies are limited to laboratory data, where laboratory data is limited, and experiment environment is well controlled with simplified fluid characterization and limited external

parameters such as limited pressure data. Field measurements would bring abundance of data and real-life observations unlike laboratory studies where conditions are controlled and may not represent the reservoir conditions.

Industry service companies have built devices that are based on Electron Paramagnetic Resonance (ERP) that quantify asphaltene deposition in real time (Abdallah et al., 2018 and Lovell et al., 2020). A ML learning platform was also developed that obtains asphaltene characteristics in 3-phase flow in real time (Lovell et al., 2021). The same companies are now working on expanding coverage of device to include wax deposition. Considering real-time crude out footprint obtained from flowing wells abundance of data could be made available and data-driven models incorporating these data could deliver practical and robust models, expanding application scope and addressing the generalization issue.

5 CONCLUSIONS AND RECOMMENDATIONS

In this chapter, the major conclusions derived from this research work are presented and further recommendations are made for future investigation in scope of WDT determination/prediction studies.

5.1 Conclusions

The objective of this research work was to implement recent developments in data-driven methods for WDT prediction aiming at brining added values to the existing literature in terms of predictive capacity, both accuracy and generalization capability with larger dataset than ones used in previous studies. Performance of four models developed based on DT ensemble methods: AdaBoost, GBM, XGBoost, and CatBoost were compared with the developed K-

nearest neighbor and linear regression models. The following six conclusions can be drawn from this research work:

1. Previous work suggested use of two the input parameters: molar weight and pressure, as contributing features to predict WDT as these include effects of fluid composition, system pressure, and intermolecular forces. The present study also confirmed the importance of these two variables.
2. Laboratory WDT data published in literature was analyzed and a reliable dataset of validated 380 data points from various sources was put together for model training and testing. Publications with WDT reporting only were reviewed, WAT reporting was excluded. Two out of ten publications on WDT were excluded because of indirect measurements (WAT and % of wax precipitation correlated to WDT). The previous modeling works were using often cited 3 studies with 272 datapoints, largest database used by Amiri-Ramsheh et al. (2009) with 346 data points. Increased dataset size extends model applicability and generalizability.
3. Input data analysis detected no major outliers, which is expected as literature published data was used, which is well screened. Few outliers were spotted for WDT. However, these were not excluded to avoid overfitting and no further data pre-processing was performed.
4. Performance of the developed intelligent models using DT ensemble methods with boosting algorithms: AdaBoost, GBM, XGBoost and CatBoost were compared with the developed KNN and LR models:
    a. From statistical metrics, both the developed CatBoost and XGBoost models demonstrates top performance in terms of R^2 and RMSE and there is slight lower performance in terms of AARD, % for the XGBoost (0.1419%) compared to the CatBoost (0.11%). The developed AdaBoost and GBM intelligent modles display

strong performance too, almost negligibly lagging behind top performers. The developed KNN model shows acceptable results in terms of R^2 (0.88) but not acceptable in terms of RMSE and AARD. The developed LR model delivers the lowest performance with R^2 at 0.5153 and high RMSE 8.67, suggesting non-linear relationship for WDT.

b. Graphical error assessment shows that the DT-based intelligent models with boosting optimizers display strong performance with alignment with unit-slope ideal line with minimal scattering on Parity or Cross-plots. The residual plots suggested that the error distribution aligned with zero line for all four ensemble models; while, the developed KNN model showed lack of trend, i.e. random distribution but no consistency is observed and the developed LR model showed clear trend that suggests systematic underfitting.

5. Top two selected methods: XGBoost and CatBoost were further analyzed:
   a. Williams' plot was used to analyze performance of the developed XGBoost and CatBoost models and suggested that the developed XGBoost model has a much narrow residual thresholds (+/-3) except for 3 outliers. Overall, suggesting the model's robustness; while, the developed CatBoost model showed slightly more outliers.
   b. SHAP analysis was performed on basis of SHAP bars and SHAP summary chart with outcome that MW is the main contributor to WDT prediction, and the developed XGBoost model has slightly higher importance allocated to P than the developed CatBoost model. An observation was made that both higher MW and higher P have a positive impact on WDT; while lower MW and lower P reduces the predicted WDT values. Non-linearity was observed in relation between MW-WDT and P-WDT.

c. Trend analysis was performed that suggested all models validate the laboratory data where higher the MW then higher the WDT. This supports the physical basis for the models and addresses the generalization issue. Dependance of WDT on P is of lower scale and not exactly consistent. However, it can be observed that higher P also led to WDT increase, which, is supported by SHAP summary analysis. However, this also contradicts some of the statements in literature. This should be further investigated.

d. Comparative analysis of the developed XGBoost and CatBoost models' performance with previous models in literature was performed. The dataset used for previous ML studies, namely, Benamara et al. (2019), Amiri-Ramsheh et al. (2021), and Nait Amar et al. (2024) with 272 data points was replicated to benefit from data comparison done in previous studies. From statistical metrics, the boosting supported DT models were found superior in terms of both accuracy and robustness when compared to the previous data-driven models and traditional thermodynamic models reported in the literature. The developed XGBoost and CatBoost models within DT-based ML methods have performed superior to all of other models from all three metrics of R^2, RMSE and AARD, %.

6. The developed GEP model building was attempted in two options: scenario 1 with complete dataset use without splitting and scenario 2 with 2 sub-sets split at MW median value of 210 g/mol. The results of both scenarios demonstrate good performance with R^2 higher than 0.9, but still lower than the GEP model published by Nait Amar et al. (2024) – $R^2 = 0.9647$ and the GMDH model published by Benamara et al. (2019) – $R^2 = 0.9582$. Further work with fine tuning of parameters is required.

Overall, the objective of this research work was met. Intelligent DT-based boosting algorithms proven to perform superior to all of the previously proposed data-driven models in literature studies for

prediction of WDT with two most recent developed models of XGBoost and CatBoost validated as top performers.

5.2 Recommendations

Within the scope of this research work, room for further improvement was observed and following recommendations are proposed for further enhancement of the current practice of WDT modeling efforts:

1. Input parameters could be revisited for more accurate characterization of fluid characteristics and composition and possible inclusion of other fluid properties beside MW and P. Clearly, this requires availability of additional reliable experimental and field data.
2. Correlation of WDT with P should be further investigated to fundamentally justify the trend observed in SHAP summary plot and Trend analysis that consider various thermodynamic studies along with data-driven methods including the present study.
3. Possibility to merge fundamental thermodynamic models such as Cautinho et al. (1995, 1996, 1998) with ML methods could be investigated to virtually develop a model that is not limited to training datasets,

4. Another measure to address generalization issue would be to expand the data for ML models trainings (currently limited to laboratory data, published in literature) by promoting ML direct measurements in field real-time that would present abundance of more practical data that already considers real-world field conditions (limited in laboratory environment).

REFERENCES

BHecker Jr, H. L. (2000). Asphaltene: To Treat or Not. SPE-59703-MS. *SPE Permian Basin Oil and Gas Recovery Conference.* Midland, TX, USA. 21 March 2000.

Bellarby, J. (2009). *Well Completion Design.* Elsevier.

Singh, P., Walker, J., Lee, H. S., et al. (2006). An Application of Vacuum Insulation Tubing (VIT) for Wax Control in an Arctic Environment. OTC 18316. *2006 Offshore Technology Conference.* Houston, TX, USA. 1-4 May 2006.

Biao, W. & Lijian, D. (1995). Paraffin Characteristics of Waxy Crude Oils in China and the Methods of Paraffin Removal and Inhibition. SPE 29954. *International Meeting on Petroleum Engineering.* Beijing, PR China. 14-17 November 1995.

Ji, H-Y., Tohidi, B., Danesh, A., & Todd, A. C. (2004). Wax phase equilibria: developing a thermodynamic model using a systematic approach. *Fluid Phase Equilibria*, *216.* 201–217.

Bian, X.-Q., Huang, Y.-W., Wang, Y., Liu, Y.-B., & Kasturiarachchi, D. T., K. (2019). Prediction of Wax Disappearance Temperature by Intelligent Models. *Energy & Fuels*, *33*. 2934-2949.

Boranbayeva, L., Boiko, G., Sharifullin, A., Lubchenko, N., Sarmurzina, R., Kozhamzharova, A. & Mombekov, S. (2024). Analysis of the Processes of Paraffin Deposition of Oil from the Kumkol Group of Fields in Kazakhstan. *Processes, 12, 1052*.

Kozhabekov, S., Zhubanov, A. & Toktarbay, Zh. (2019). Study the rheological properties of waxy oil with modified pour point depressants for the South Turgai oil field in Kazakhstan. *Oil and Gas Science and Technology – Rev. IFP Energies Nouvelles, 74.*

Monger-Mcclure, T.G., Tackett, J.E., & Merrill, L.S. (1999). Comparisons of Cloud Point Measurement and Paraffin Prediction Methods. *SPE production & facilities, 14*(1), 4-16.

Zhao, Y., Paso, K., Norrman, J., Ali, H., Sorland, G., & Sjoblom, J. (2015). Utilization of DSC, NIR, and NMR for wax appearance temperature and chemical additive performance characterization. *J Therm Anal Calorim*, *120,* 1427–1433.

Juyal, P., Cao, T., Yen, A., & Venkatesan, R. (2011). Study of live oil wax precipitation with high-pressure micro-differential scanning calorimetry. *Energy & Fuels, 25*(2), 568-572.

Alcazar-Vara, L.A., & Buenrostro-Gonzalez, E. (2013). Liquid-Solid Phase Equilibria of Paraffinic Systems by DSC Measurements. *In:* A. A. Elkordy, ed. *Applications of Calorimetry in a Wide Context - Differential Scanning Calorimetry, Isothermal Titration Calorimetry and Microcalorimetry.* IntechOpen, 253-276.

Chen, H, Yang, S., Nie, X., Zhang, H., Huang W., Wang, Z., & Hu, W. (2014). Ultrasonic Detection and Analysis of Wax Appearance Temperature of Kingfisher Live Oil. *Energy Fuels, 28,* 2422−2428.

Elsharkawy, A. M., & Al-Sahhaf, T. A. (2000). Wax deposition from Middle East crudes. *Fuel, 79(9),* 1047–1055.

Sarica, C., Zhang, J., & Volk, M. (2008). Experimental investigation of wax deposition and removal in flow loops. *Journal of Energy Resources Technology, 130(4),* 043102.

Dantas Neto, A. A., Gomes, E. A. S., Barros Neto, E. L., Dantas, T. N. C. & Moura, C. P. A. M. (2009). Determination of wax appearance temperature (WAT) in paraffin/solvent systems by photoelectric signal and viscosimetry. *Brazilian Journal of Petroleum and Gas, 3(4)*, 149-157.

Mehrotra, A. K. & Bhat, N. V. (2007). Modeling the Effect of Shear Stress on Deposition from "Waxy" Mixtures under Laminar Flow with Heat Transfer. *Energy & Fuels, 21,* 1277-1286.

Taraneh, J. B., Rahmatollah, G., Hassan, A., & Alireza, D. (2008). Effect of wax inhibitors on pour point and rheological properties of Iranian waxy crude oil. *Fuel Processing Technoloy, 89(10),* 973-977.

Chen, C., Zhang, J., Xie, Y., Huang, Q., Ding, Y., Zhuang, Y., Xu, M., Han, S., Li, Z., & Li, H. (2021). An investigation to the mechanism of the electrorheological behaviors of waxy oils. *Chemical Engineering Science, 239*.

Wang, Y., Liu, X., Huang, Z., Wang, Z., & Liu, Y. (2022). Characterization of Wax Precipitation and Deposition Behavior of Condensate Oil in Wellbore: A Comprehensive Review of Modeling, Experiment, and Molecular Dynamics Simulation. *Energies, 15(11)*.

Alnaimat, F., Ziauddin, M., & Mathew, B. (2020). Wax deposition in crude oil transport lines and wax estimation methods. *Intelligent System and Computing.*

Bhat, N.V., & Mehrotra, A.K. (2004). Measurement and prediction of the phase behavior of wax−solvent mixtures: significance of the wax disappearance temperature. *Industrial & Engineering Chemistry Research, 43*(13), 3451-3461.

American Society for the International Association for Testing and Materials (2021), ASTM D4419 – 90: Standard Test Method for Measurement of Transition Temperatures of Petroleum Waxes by Differential Scanning Calorimetry (DSC). ASTM Standards Online. Available at: https://www.astm.org/d4419-90r21.html (Accessed: 30.01.2025)

American Society for the International Association for Testing and Materials (2021), ASTM D5771 – 21: Standard Test Method for Cloud Point of Petroleum Products and Liquid

Fuels (Optical Detection Stepped Cooling Method). ASTM Standards Online. Available at: https://www.astm.org/d5771-21.html (Accessed: 30.01.2025)

American Society for the International Association for Testing and Materials (2021), ASTM D5773 – 21: Standard Test Method for Cloud Point of Petroleum Products and Liquid Fuels (Constant Cooling Rate Method). ASTM Standards Online. Available at: https://www.astm.org/d5773-21.html (Accessed: 30.01.2025)

American Society for the International Association for Testing and Materials (2021), ASTM D8420 – 21: Standard Test Method for Wax Appearance Temperature and Wax Disappearance Temperature of Petroleum Products and Liquid Fuels. ASTM Standards Online. Available at: https://www.astm.org/d8420-21.html (Accessed: 30.01.2025)

American Society for the International Association for Testing and Materials (2022), ASTM D97 – 17b: Standard Test Method for Pour Point of Petroleum Products. ASTM Standards Online. Available at: https://www.astm.org/d0097-17br22.html (Accessed: 30.01.2025)

Robles, L., Espeau, P., Mondieig, D., Haget, Y., & Oonk, H.A.J. (1996). Polymorphism and molecular alloys in the binary system $C_{17}H_{36}$-$C_{19}H_{40}$. *Thermochimica Acta, 274.* 61-72.

Metivaud, V., Rajabalee, F., Oonk, H. A. J., Mondieig, D., & Haget, Y. (1999). Complete determination of the solid (RI)–liquid equilibria of four consecutive n-alkane ternary systems in the range $C_{14}H_{30}$–$C_{21}H_{44}$ using only binary data. *Canadian Journal of Chemistry, 77*. 332-339.

Dauphin, C., Daridon, J.L., Coutinho, J., Baylere, P., & Potin-Gautier, M. (1999). Wax content measurements in partially frozen paraffinic systems. *Fluid Phase Equilibria*, *161.* 135–151.

Pauly, J., Daridon, J-L., & Coutinho, J.A.P. (2001). Measurement and prediction of temperature and pressure effect on wax content in a partially frozen paraffinic system. *Fluid Phase Equilibria, 187–188*. 71–82.

Daridon, J.L., Pauly, J., & Milhet, M. (2002). High pressure solid–liquid phase equilibria in synthetic waxes. *Phys. Chem. Chem. Phys., 4.* 4458–4461.

Milhet, M., Pauly, J., Coutinho, J.A.P., Dirand, M., & Daridon, J.L. (2005). Liquid–solid equilibria under high pressure of tetradecane+pentadecane and tetradecane+hexadecane binary systems. *Fluid Phase Equilibria*, *235.* 173–181.

Rizzo, A., Carrier, H., Castillo, J., Acevedo, S., & Pauly, J. (2007). A new experimental setup for the liquid–solid phase transition determination in crude oils under high pressure conditions. *Fuel*, *86.* 1758–1764.

Mansourpoor, M., Azin, R., Osfouri, S., & Izadpanah, A. A. (2019). Experimental measurement and modeling study for estimation of wax disappearance temperature. *Journal of Dispersion Science and Technology, 40:*2. 161-170.

Shariatrad, F., Javanmardi, J., Rasoolzadeh, A., & Mohammadi, A.H. (2007). Experimental Measurement and Thermodynamic Modeling of the Wax Disappearance Temperature (WDT) for a Quaternary System of Normal Paraffins. *ACS Omega 2022, 7.* 16928−16938.

Won, K. W. (1986). Thermodynamics for solid solution-liquid-vapor equilibria: wax phase formation from heavy hydrocarbon mixture. *Fluid Phase Equilibria, 30.* 265–279.

Hansen, A. B., Pedersen, K. S., & Rønningsen, H. P. (1988). A Thermodynamic Model for Predicting Wax Formation in Crude Oils. *AIChE Journal, 34/*12. 1937–1942.

Won, K. W. (1989). Thermodynamic calculation of cloud point temperatures and wax phase compositions of refined hydrocarbon mixtures. *Fluid Phase Equilibria, 53.* 377–396.

Pedersen, K. S., Skovborg, P., & Rønningsen, H. P. (1991a). Wax Precipitation from North Sea CrudeOils. 4. Thermodynamic Modeling. *Energy & Fuels, 5.* 924–932.

Pedersen, W. B., Hansen, A. B., Larsen, E., Nielsen, A. B. (1991b). Wax Precipitation from North Sea Crude Oils. 2. Solid-Phase Content as Function of Temperature Determined by Pulsed NMR. *Energy & Fuels*, *5*. 908-13.

Pedersen, K. S. (1993). Prediction of Cloud Point Temperatures and Amount of Wax Precipitation. *SPE Production & Facilities.* 46-49.

Lira-Galeana, C., Firoozabadi, A., & Prausnitz, J. (1996). Thermodynamics of Wax Precipitation in Petroleum Mixtures. *AIChE Journal, 42/*1. 239–248.

Coutinho, J. A. P., Andersen, S. I., & Stenby, E. H. (1995). Evaluation of activity coefficient models in prediction of alkane solid-liquid equilibria. *Fluid Phase Equilibria*, *103*. 29–39.

Coutinho, J. A. P., Knudsen, K., Andersen, S. I., & Stenby, E. H. (1996). A Local Composition Model for Paraffinic Solid Solutions. *Chemical Engineering Science*, *51/*12. 3273–3282.

Pauly, J., Dauphin, C., & Daridon, J-L. (1998). Liquid–solid equilibria in a decane + multi-paraffins system. *Fluid Phase Equilibria, 149.* 191–207.

Vafaie-Sefti, M., Mousavi-Dehghani, S.A., & Bahar, M. M-Z. (2000). Modification of multisolid phase model for prediction of wax precipitation: a new and effective solution method. *Fluid Phase Equilibria, 173.* 65-80.

Pauly, J., Daridon, J-L., & Coutinho, J.A.P. (2004). Solid deposition as a function of temperature in the nC10 + (nC24–nC25–nC26) system. *Fluid Phase Equilibria, 224.* 237–244.

Rønningsen, H. P., Bjørndal, B., Hansen, A. B., & Batsberg Pedersen, W. (1991). Wax precipitation from North Sea crude oils. 1. Crystallization and dissolution temperatures, and Newtonian and non-Newtonian flow properties. *Energy and Fuels, 5*(6). 895-908.

Moradi, G., Mohadesi, M, & Moradi, M. R. (2013). Prediction of wax disappearance temperature using artificial neural networks. *Journal of Petroleum Science and Engineering, 108.* 74–81.

Kamari, A., Rahimzadeh, A., Mohammadi A. H., & Ramjugernath, D. (2019). Evaluation of wax disappearance temperatures in hydrocarbon fluids using soft computing approaches. *Petroleum Science and Technology*, *37/*7. 829–836.

Benamara, C., Amar, M. N., Gharbi, K., & Hamada, B. (2019). Modeling Wax Disappearance Temperature Using Advanced Intelligent Frameworks. *Energy & Fuels*, *33.* 10959 – 10968.

Amiri-Ramsheh, B., Safaei-Farouji, M., Larestani, A., Zabihi, R., & Hemmati-Sarapardeh, A. (2021). Modeling of wax disappearance temperature (WDT) using soft computing approaches: Tree-based models and hybrid models. *Journal of Petroleum Science and Engineering,* 208.

Amar, M. N., Zeraibi, N., Benamara C., Djema, H., Saifi, R., & Gareche, M. (2024). Modeling wax disappearance temperature using robust white-box machine learning. *Fuel*, 376.

Acharya, S (2021). Comparative Analysis of classification accuracy for XGBoost, LightGBM, CatBoost, H2O, and Classifium. Thesis (Master). Østfold University College.

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, *34/*4, 301-312.

Chinnamgari, S. K. (2019). *R Machine Learning Projects: Implement supervised, unsupervised, and reinforcement learning techniques using R 3.5.* Packt Publishing Ltd.

Geron, A. (2019). *Hands-on machine learningwith Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media.

Zhang, C., & Zhang, S. (2003). Association rule mining: models and algorithms. *Springer*, *vol. 2307*.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109/*2, pp. 373-440.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning.* MIT press.

"Xgboost: A scalable tree boosting system" in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785-794.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist*. *29* (5) 1189 - 1232, October 2001.

Fürnkranz, J. (2011). *Decision Tree. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning.* Springer, Boston, MA.

Quinlan, J.R. (1986). Induction of decision trees. *Mach Learn 1*, 81–106.

Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). *Classification and Regression Trees (1st ed.).* Chapman and Hall/CRC.

Quinlan, J. R. (2014). *C4.5: programs for machine learning.* Elsevier.

Quinlan, J. R. (2004). *Data mining tools see5 and c5. O,* http://www.rulequest.com/see5-info.html

Freund, Y., & Schapire, R.E. (1997). A Decision -Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of computer and system sciences, 55*, 119-139.

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5–32.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Mach Learn 63*, 3–42.

Chen, T., & Guestrin, C. (2016) "Xgboost: A scalable tree boosting system," CoRR, vol. abs/1603.02754, 2016. arXiv: 1603.02754. [Online]. Available: http:/ /arxiv.org/abs/1603.02754

Meng, G.Q., Finley, T., Wang, T., Chen, W., Ma, W., Q. Ye, Q., & Liu, T.Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems,* 3146-3154.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., & Gulin, A. (2017). Catboost: Unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516, 2017.

Brownlee, J. (2016). A gentle introduction to the gradient boosting algorithm for machine learning. Available: https://machinelearningmastery.com

Chen, T., & He, T. (2015). Higgs boson discovery with boosted trees. *NIPS 2014 workshop on high-energy physics and machine learning*, pp. 69-80.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*/2.

Abdallah, D., Punnapala, S., Kulbrandstad, O., Godoy, M., Madem, S., Babakhani, A. & Lovell, J. (2018). Asphaltene Studies in On-Shore Abu Dhabi Fields, Part IV: Development of a Surface Sensor. SPE-191676-MS. *2018 SPE Annual Technical Conference and Exhibition.* Dallas, TX, USA. 24-26 September 2018.

Lovell, J., Abdallah, D., Punnapala, S., Al Daghar, K., Kulbrandstad, O., Madem, S. & Meza, D. (2020). A Chemical IoT System for Flow Assurance - From Single-Well Applications to Field Implementation. SPE-203286-MS. *Abu Dhabi International Petroleum Exhibition & Conference.* Abu Dhabi, UAE. 9 November 2020.

Lovell, J., Abdallah, D., Fonseca, R. M., Grutters, M., Punnapala, S., Kulbrandstad, O., Meza, D. & Baez, J. (2021). Interpretation Challenges and Solutions for Real-Time Asphaltene Paramagnetic Sensing at the Wellhead. SPE-207553-MS. *Abu Dhabi International Petroleum Exhibition & Conference.* Abu Dhabi, UAE. 15-18 November 2020.