Semesterarbeit

theoretical

Analysis and Machine Learning Modeling of Plastic Pyrolysis Data

Oskar Breitfeld

SA 1034

2026

Remove this and the next page in the print-out; This page is for the pdf-version only.
Diese Seite und die darauffolgende in der gedruckten Version entfernen; nur für die PDF-Version.

Remove this page in the print version.

Diese Seite in der Druckversion entfernen.

Lehrstuhl für Energiesysteme

Title of Semesterarbeit:	Analysis and Machine Learning Modeling of Plastic Pyrolysis Data
Author:	Oskar Breitfeld
Enrolment Number:	[redacted]
Supervisor:	Lukas Martetschläger
Assigned by:	Prof. Dr.-Ing. Hartmut Spliethoff Technical University of Munich Chair of Energy Systems Boltzmannstr. 15 85748 Garching
Assigned:	15.10.2025
Handed in:	15.04.2026

Statutory Declaration

I hereby declare that I have prepared the present work independently and without help of others. Thoughts and quotes that have been taken over directly or indirectly from other sources are marked as such. This work has not been submitted to any examining authority in the same or similar form and has not been published yet.

I hereby agree that the work can be made available by the Institute of Energy Systems to the public.

Use of AI Tools. The following AI tools were used in the preparation of this work:

- **GitHub Copilot**: support for code development and automated error identification in Python scripts.
- **Claude** (Anthropic): language proofreading, text editing, and LaTeX formatting assistance.
- **Google Gemini 2.5 pro API**: automated paper evaluation script.
- **Perplexity AI**: research help, including retrieval of waste statistics and literature comparisons.

All AI-generated outputs were critically reviewed, verified, and substantively adapted by the author. The intellectual content, scientific analysis, and conclusions of this work are the author's own.

_____, the _____ _____

Abstract

Accurate prediction of plastic pyrolysis product yields is essential for process design and optimization, yet existing machine learning studies in this domain suffer from unrecognized methodological shortcomings. This work presents a systematic investigation into machine learning-based yield prediction for non-catalytic plastic pyrolysis, built upon a unified database of 1,072 experimental data points compiled from 127 peer-reviewed publications. After systematic preprocessing - including mass balance validation, and feature harmonization - 795 data points remained for model training and evaluation.

A central contribution of this thesis is the identification and quantification of data leakage in pyrolysis machine learning. This work demonstrates that conventional random cross-validation inflates model accuracy by approximately 50 % to over 100 % in terms of mean absolute error (MAE) because data points from the same publication share latent correlations. To address this, Paper-Level GroupKFold cross-validation is introduced as the first application of grouped splitting by publication source in the pyrolysis ML literature - ensuring that no paper contributes data to both training and test folds simultaneously.

Under this rigorous evaluation protocol, the best-performing gradient-boosted models achieve MAE values of 16.41 wt% (oil), 11.97 wt% (gas), and 11.08 wt% (char) for XGBoost, which was selected as the primary model for its exact SHAP implementation and literature comparability. SHAP (SHapley Additive exPlanations) analysis confirms that the dominant feature - pyrolysis temperature - exhibits physically consistent directional effects across all three product fractions, validating the model's reliance on genuine thermochemical relationships rather than statistical artifacts. The residual prediction error is shown to approach the inter-laboratory noise floor inherent in heterogeneous literature data.

Keywords: Plastic Pyrolysis, Machine Learning, Database Integration, Yield Prediction

Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Meaning
ABS	Acrylonitrile Butadiene Styrene
ANN	Artificial Neural Network
BMBF	Bundesministerium für Bildung und Forschung
BTX	Benzene, Toluene, Xylene
CB	CatBoost
CHNS	Carbon, Hydrogen, Nitrogen, Sulfur
CSBR	Conical Spouted Bed Reactor
CV	Cross-Validation
DB	Database
DNN	Deep Neural Network
GBR	Gradient Boosting Regressor
GKF	GroupKFold
HDPE	High-Density Polyethylene
HIPS	High-Impact Polystyrene
KS	Kolmogorov–Smirnov
LDPE	Low-Density Polyethylene
LES	Lehrstuhl für Energiesysteme (Chair of Energy Systems)
LLDPE	Linear Low-Density Polyethylene
MAE	Mean Absolute Error
MB	Mass Balance
ML	Machine Learning
MNAR	Missing Not At Random
MSW	Municipal Solid Waste
PCA	Principal Component Analysis
PE	Polyethylene
PET	Polyethylene Terephthalate
PP	Polypropylene
PS	Polystyrene
PVC	Polyvinyl Chloride
RF	Random Forest
RMSE	Root Mean Square Error
SHAP	SHapley Additive exPlanations
SVR	Support Vector Regression
TGA	Thermogravimetric Analysis
TUM	Technical University of Munich
WoS	Web of Science
wt%	Weight Percent
XGB	Extreme Gradient Boosting (XGBoost)

Notation

Latin Symbols

Symbol	Unit	Explanation
k	—	Number of cross-validation folds
m	—	Number of permutation repeats
n	—	Number of data points / samples
p	—	Number of input features
r	—	Pearson correlation coefficient
R^2	—	Coefficient of determination
Y_{total}	wt %	Sum of oil, gas, and char yields
y_i	wt %	Measured product yield of sample i
$\hat{y}_i$	wt %	Predicted product yield of sample i
$\mathbf{X}$	—	Feature matrix

Greek Symbols

Symbol	Unit	Explanation
α	—	Regularization strength (Elastic Net)
$\boldsymbol{\beta}$	—	Regression coefficient vector
ε	wt %	Tolerance margin (SVR)
λ	—	L1/L2 mixing ratio (Elastic Net)
ϕ_j	wt %	SHAP value for feature j
ϕ_0	wt %	SHAP base value (expected prediction)

Indices

Symbol	Explanation
i	Sample index
j	Feature index
oil, gas, char	Target product fraction
train, val	Training and validation set

1 Introduction

1.1 Motivation

Global plastic production has grown at an unprecedented rate over the past seven decades, reaching approximately 400 million tonnes per year as of 2022 [40]. Since the onset of large-scale polymer manufacturing in the 1950s, cumulative production has exceeded 10 billion tonnes, of which an estimated 60% has accumulated in landfills or the natural environment [20]. Despite increasing public awareness and regulatory efforts, the global recycling rate for plastic waste remains below 10%, with the vast majority of post-consumer plastics either incinerated for energy recovery or disposed of in landfills [37]. This linear production-consumption-disposal pattern poses severe environmental challenges, including marine pollution, microplastic contamination, and the release of greenhouse gases from both incineration and degradation processes.

Germany is frequently cited as a leader in waste management. The official recycling rate for plastic packaging stands at 70.8%, as reported by the Zentrale Stelle Verpackungsregister [54]. However, this figure is measured on an *input* basis - meaning, it refers to the mass of collected material *sent to* recycling facilities, not to the mass of secondary raw material actually recovered. When measured at the output of recycling plants, the actual material recovery rate drops to approximately 33-38% [49, 13]. This discrepancy arises from substantial material losses during sorting and reprocessing: contaminated fractions, multi-layer packaging, and non-target polymers are rejected at various stages and diverted to thermal recovery.

The scale of Germany's plastic waste challenge is considerable. In 2023, total plastic waste generation amounted to 5.91 million tonnes [13]. Of this volume, only 38.4% was materially recycled (output-based), while 61.1% was directed to thermal recovery - that is, incineration with energy recovery [50, 13]. Several structural factors contribute to this gap between headline recycling rates and actual material recovery. First, Germany's dual waste collection system (*Gelber Sack* / Duales System) collects mixed post-consumer plastic packaging, but an estimated 30% of the collected material is misclassified or non-recyclable [49]. Second, deposit-return systems for PET beverage bottles (*Pfandflaschen*) achieve high collection rates and are included in the overall recycling statistics, thereby inflating the aggregate figure for mixed plastic waste streams where actual recycling rates are considerably lower [50]. Third, the distinction between "recycled" and "sent to recycling" in official statistics masks material losses occurring during sorting and reprocessing, which can amount to 50% or more of the input mass for mixed plastic waste [42, 13]. Taken together, the effective mechanical recycling rate for mixed, contaminated, or multi-layer plastic waste in Germany is substantially below the headline figures reported in official statistics.

Mechanical recycling, while effective for clean, single-polymer waste streams such as PET bottles or HDPE containers, faces fundamental limitations when applied to mixed or contaminated post-consumer plastics. Polymer blends, multi-layer packaging, and the presence of additives, pigments, or food residues lead to progressive quality degradation with each recycling cycle - a phenomenon commonly referred to as downcycling [42, 3]. Consequently, a significant

fraction of plastic waste that cannot be mechanically recycled at acceptable quality levels currently has no established recycling pathway and defaults to incineration or landfilling.

Chemical recycling through pyrolysis offers a promising complementary route for precisely these non-mechanically-recyclable fractions. Pyrolysis is a thermochemical conversion process that decomposes polymeric materials in an oxygen-free atmosphere at elevated temperatures, typically between 400 and 700 °C, yielding a product spectrum of pyrolysis oil (liquid), non-condensable gas, and solid residue (char) [47]. The liquid fraction can serve as feedstock for petrochemical processes, effectively closing the loop from waste plastic to virgin-grade monomers or fuels [30]. However, the product distribution is highly sensitive to a complex interplay of process parameters - including temperature, heating rate, residence time, reactor type, and feedstock composition - making process optimization and scale-up challenging [34].

Given the large number of interacting variables and the cost of systematic experimental campaigns, predictive models that can estimate product yields from process parameters and feedstock characteristics are of substantial practical value. Such models could inform reactor design, guide parameter optimization, and reduce the number of experiments required during commissioning and operation of pyrolysis plants. Machine learning approaches, which can capture nonlinear relationships without requiring explicit mechanistic formulations, are particularly well suited for this task.

1.2 Problem Statement

The scientific literature contains hundreds of experimental studies on plastic pyrolysis, collectively reporting yield data for a wide range of feedstocks, reactor configurations, and operating conditions. However, these data are scattered across publications with inconsistent reporting conventions: yield definitions vary (e.g., oil vs. liquid vs. condensable fraction), mass balances are not always closed, and process parameters are reported at different levels of detail [47]. Compiling a unified database from such heterogeneous sources requires careful harmonization and quality control, making database construction a research contribution in its own right.

Several recent studies have applied machine learning to compiled literature databases for plastic pyrolysis yield prediction, reporting promising results. Reported coefficients of determination (R^2) range from 0.85 to 0.99, with XGBoost consistently emerging as the best-performing model across studies [6, 55, 12, 27]. These results suggest that ML models can, in principle, learn the complex parameter-yield relationships governing pyrolysis processes from literature data alone.

However, a critical methodological concern arises from the cross-validation strategies employed in all published ML-pyrolysis studies identified in this work. Without exception, these studies use random train/test splits or standard k-fold cross-validation, which randomly distribute individual data points across training and test folds regardless of their paper of origin. This practice introduces what Kapoor and Narayanan [25] classify as type L3.2 data leakage: non-independent train/test splits arising from grouped data structures.

The leakage mechanism is straightforward. A single paper typically reports multiple experiments conducted on the same reactor, using the same feedstock batches, prepared by

the same experimentalists, and varying only one or two parameters (e.g., temperature). When data points from such a paper appear in both the training and test sets, the model can exploit shared latent information - reactor-specific biases, measurement idiosyncrasies, or feedstock batch effects - rather than learning genuinely generalizable parameter-yield relationships. The resulting performance metrics are systematically inflated and do not reflect the model's ability to predict yields for truly new, unseen experimental setups.

A systematic survey of the ML-pyrolysis literature from 2022 to 2025 (see Table 2.1) revealed that none of the identified studies employs group-based cross-validation (e.g., GroupKFold), leave-one-study-out validation, or any other strategy that prevents data points from the same paper from appearing simultaneously in training and test sets. This represents a significant methodological gap that calls into question the generalizability claims of existing models.

1.3 Research Questions

This thesis addresses the following research questions:

1. **Generalization:** Can a machine learning model trained on heterogeneous literature data reliably predict plastic pyrolysis yields for new, previously unseen experimental setups?
2. **Data leakage quantification:** To what extent does paper-level data leakage inflate reported performance metrics in the existing ML-pyrolysis literature, and how large is the discrepancy between random-split and group-based cross-validation?
3. **Feature importance:** What are the dominant process parameters and feedstock characteristics controlling pyrolysis yield distributions, and are the learned feature importances physically consistent with established thermochemical knowledge?

1.4 Contributions

This work makes the following contributions to the field of machine-learning-based pyrolysis modeling:

1. **Largest compiled non-catalytic plastic pyrolysis database.** A database of 795 experimental data points extracted from 127 peer-reviewed publications is constructed, focusing exclusively on non-catalytic thermal pyrolysis. This database is, to the best of the author's knowledge, the largest of its kind in the published literature.
2. **First application of paper-level GroupKFold cross-validation.** For the first time in the pyrolysis ML literature, a group-based cross-validation strategy is employed that strictly prevents data points from the same source paper from appearing in both training and test folds. This yields performance estimates that reflect true generalization to new experimental setups.
3. **Quantification of the data leakage effect.** By comparing random-split and GroupKFold cross-validation on the same dataset and model, the magnitude of performance inflation due to paper-level data leakage is quantified. Results indicate that conventional random splits underestimate the mean absolute error by approximately 50 % to over

100 % depending on the target variable, providing a concrete measure of how severely published metrics overstate model generalizability (Section 5.4).

4. **SHAP-based feature importance with physical validation.** SHAP (SHapley Additive exPlanations) analysis is applied to identify the dominant features driving model predictions, and the learned feature-yield relationships are validated against established thermochemical principles to ensure physical plausibility.
5. **Systematic data quality assessment.** A paper-level error analysis methodology is developed that identifies outlier papers with disproportionately high prediction errors, enabling targeted quality assessment and providing a framework for iterative database refinement.

1.5 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2 provides the theoretical background and reviews the state of the art. It introduces the fundamental principles of plastic pyrolysis, defines the relevant product fractions, and surveys the key process parameters and reactor concepts. The chapter concludes with a review of existing machine learning approaches for pyrolysis yield prediction, highlighting the methodological gaps that motivate this work.

Chapter 3 describes the systematic literature search, screening, and data extraction workflow used to construct the pyrolysis database. It details the Web of Science query design, the multi-stage screening process including AI-assisted abstract filtering, and the data extraction protocol. The integration of data from predecessor databases and the schema for provenance tracking are also presented.

Chapter 4 details the machine learning methodology, including the data preprocessing pipeline (filtering, imputation, feature engineering), the cross-validation strategies evaluated - with emphasis on the paper-level GroupKFold approach central to this work - the model families considered (Elastic Net, Random Forest, XGBoost, CatBoost), the hyperparameter tuning procedure, and the evaluation metrics. The SHAP-based interpretability framework and the physical validation methodology are also introduced.

Chapter 5 presents the results, including the overall model performance comparison, the cross-validation strategy comparison quantifying the data leakage effect, feature importance rankings, and the paper-level error analysis.

Chapter 6 discusses the implications of the results in the context of the existing literature. It examines the data leakage problem as a field-wide issue through the lens of the Kapoor-Narayanan framework, discusses limitations of the present approach, and derives practical implications for the Burghausen H_2-Reallabor project.

Chapter 7 summarizes the key findings and provides an outlook on future work, including potential extensions to catalytic pyrolysis, transfer learning approaches, and deployment considerations.

1.6 Context: H_2-Reallabor Burghausen

This thesis is conducted within the framework of the H_2-Reallabor Burghausen project at the Chair of Energy Systems (Lehrstuhl für Energiesysteme), Technical University of Munich [26]. The H_2-Reallabor Burghausen is a large-scale demonstration project – funded by the German Federal Ministry of Education and Research (BMBF) with approximately 39 million euros – investigating hydrogen-based energy systems and associated value chains in the Burghausen-Chemiedreieck industrial region in southeastern Bavaria [9, 8].

One component of this project is the development of a mobile pyrolysis container unit equipped with a rotary kiln reactor for decentralized plastic waste valorization [26]. The concept envisions deployable units that can process locally collected, non-recyclable plastic waste into pyrolysis oil, which can subsequently be fed into existing petrochemical infrastructure or upgraded to hydrogen via reforming.

The machine learning model developed in this thesis serves as a decision-support tool within this project context. By predicting expected product yields as a function of feedstock composition and operating parameters, the model can inform the selection of process conditions for the rotary kiln reactor during commissioning and operation. Furthermore, the SHAP-based feature importance analysis identifies the most influential process parameters, thereby guiding the design of targeted experimental campaigns for the mobile unit.

It should be noted that the present database is compiled from literature data spanning a wide range of reactor types and scales, and the direct transferability of model predictions to the specific rotary kiln configuration in Burghausen requires careful validation. This limitation and potential mitigation strategies are discussed in Chapter 6.

2 Background and State of the Art

This chapter introduces the scientific and methodological foundations upon which the present work rests. Section 2.1 reviews the thermochemical principles of plastic pyrolysis, the product spectrum, the key process parameters, and the reactor technologies relevant to laboratory-scale experiments. Section 2.2 surveys the application of machine learning to pyrolysis yield prediction, comparing the approaches, datasets, and reported performance of recent studies. Section 2.3 introduces the concept of data leakage in machine-learning-based science, focusing on the taxonomy of Kapoor and Narayanan [25] and its direct applicability to the pyrolysis modeling domain.

2.1 Plastic Pyrolysis Fundamentals

2.1.1 Definition and Product Spectrum

Pyrolysis is the thermochemical decomposition of organic materials in the absence of oxygen, typically conducted at temperatures between 300 °C and 900 °C [47]. When applied to plastic waste, pyrolysis breaks down long-chain polymer molecules into shorter hydrocarbon fragments through a combination of random scission, end-chain scission, and depolymerization mechanisms. The relative contribution of each mechanism depends on the polymer type, the temperature, and the residence time of both vapor and solid phases [45, 1].

The products of plastic pyrolysis are conventionally classified into three fractions reported on a mass basis (wt %):

- **Liquid oil (or wax):** Condensable hydrocarbons collected by cooling the pyrolysis vapors. The liquid fraction comprises a complex mixture of aliphatic and aromatic compounds spanning a wide carbon-number range. At moderate temperatures (400 °C-500 °C), the condensable product from polyolefins is often a heavy wax (predominantly C_{20}-C_{50}), whereas at higher temperatures (500 °C-700 °C), secondary cracking shifts the distribution toward lighter oils (C_5-C_{20}) containing paraffins, olefins, naphthenes, and aromatics [33, 16].
- **Non-condensable gas:** Light hydrocarbons (C_1-C_4) and small amounts of hydrogen and carbon oxides that pass through the condensation system. The gas fraction increases with temperature as thermal cracking intensifies [52].
- **Solid residue (char):** Carbonaceous residue remaining in the reactor after pyrolysis. For pure thermoplastics, the char yield is typically low (1 %-5 %) at temperatures above 500 °C, though it can be significantly higher for polymers containing aromatic backbones (e.g., PET) or heteroatoms (e.g., PVC) [38, 47].

A critical challenge in compiling pyrolysis data from different laboratories is the inconsistency in how these fractions are defined and measured. The boundary between “oil” and “wax” varies with condenser temperature and collection protocol; some authors report a single “liquid” fraction while others distinguish between oil and wax by solvent extraction or filtration.

Similarly, the gas yield may be measured directly (e.g., by gas chromatography of the exit stream) or calculated by difference from the mass balance (gas = 100 − liquid − char). These reporting differences introduce systematic noise that any cross-study modeling effort must account for [6, 12].

2.1.2 Key Process Parameters

The yield distribution among liquid, gas, and solid products is governed by several interdependent process parameters. The four parameters most commonly reported in the literature and most relevant to the feature set used in this work are briefly reviewed below.

Temperature. Pyrolysis temperature is the single most influential process variable. At temperatures below approximately 400 °C, polymer cracking is incomplete and the solid residue fraction is large. As temperature increases through the 450 °C-550 °C range, liquid yields reach a maximum for most polyolefins, reflecting the balance between primary cracking (which generates condensable fragments) and secondary cracking (which converts these fragments into lighter, non-condensable gases). Above 600 °C, the gas fraction dominates as secondary cracking reactions become increasingly severe [33, 52, 16]. For polystyrene, the temperature dependence is qualitatively different: styrene monomer recovery is already high at 400 °C-500 °C, and liquid yields above 80 % are commonly reported [35, 38].

Heating rate. The rate at which the feedstock reaches the target temperature affects the competition between primary and secondary reactions. Slow pyrolysis (heating rates of 1 °C/min-20 °C/min) allows extended contact between primary cracking products and the hot solid matrix, favoring char formation and secondary gas production. Fast pyrolysis (heating rates exceeding 100 °C/min, up to flash conditions of 1000 °C/s) rapidly ejects primary volatiles from the reaction zone, maximizing liquid yields [39, 47]. In practice, the effective heating rate depends on the reactor type, the particle size, and the heat transfer medium.

Residence time. Two distinct residence times are relevant: the *solid residence time* (the duration for which the feedstock remains at the reaction temperature in the reactor) and the *vapor residence time* (the time that primary cracking products spend in the hot zone before being quenched or exiting the reactor). Vapor residence time is particularly important because it controls the extent of secondary cracking: longer vapor residence times shift the product distribution from liquids toward gases [33]. In fluidized bed and spouted bed reactors, vapor residence time can be controlled independently of solid residence time through the carrier gas flow rate, providing an additional degree of freedom for product optimization [16, 23].

Particle size. The size of the feedstock particles affects heat transfer to the polymer melt. Larger particles experience internal temperature gradients, leading to slower and potentially incomplete decomposition. At laboratory scale, particle sizes typically range from 1 mm to 10 mm; at pilot scale, shredded plastic flakes of 5 mm-30 mm are common. Although particle size is recognized as a relevant variable, it is one of the most inconsistently reported parameters in the literature, with many studies omitting it entirely or reporting it only qualitatively [6].

2.1.3 Reactor Types

Laboratory-scale pyrolysis experiments employ a variety of reactor configurations that differ in heat transfer mechanism, contact pattern, and operating mode (batch vs. continuous). The reactor type influences the achievable heating rates, the residence time distributions, and the scale of the experiment, all of which affect product yields and thus represent important contextual variables for any cross-study database.

Batch reactors (autoclaves, fixed-bed reactors, and stirred reactors) are the most common configuration at laboratory scale. The feedstock is loaded into a closed vessel, heated to the target temperature under an inert atmosphere, and the volatile products are continuously swept out by a carrier gas stream and collected in a condensation train. Batch reactors are simple to operate and provide precise mass balances (the entire feedstock charge and all products can be weighed), but they impose long solid residence times and do not control vapor residence time independently. The majority of published pyrolysis datasets originate from batch experiments [38, 47].

Semi-batch reactors represent an intermediate category in which fresh feedstock is added continuously or periodically while the reactor temperature is maintained. This configuration is less common in the academic literature but appears in some temperature-programmed experiments.

Continuous reactors operate with steady-state feedstock feeding and product removal. Several designs are used at laboratory and pilot scale:

- *Fluidized bed reactors* use a bed of inert particles (typically sand) fluidized by the carrier gas to provide rapid heat transfer to the polymer feed. They achieve high heating rates and short, controllable vapor residence times, making them well-suited for maximizing liquid yields [52, 33, 23].
- *Conical spouted bed reactors* (CSBRs) employ a conical geometry that creates vigorous solid circulation, enabling processing of sticky polymer melts without agglomeration. The CSBR has been extensively studied for polyolefin pyrolysis by the research group of Olazar and Bilbao [16].
- *Rotary kilns and auger (screw) reactors* use mechanical rotation to transport and mix the feedstock. They are particularly relevant for pilot- and industrial-scale applications due to their ability to handle heterogeneous waste streams [45, 39], but they are less common in laboratory-scale studies.

In the context of machine learning modeling, the reactor type is typically encoded as a categorical feature. Reactor type is encoded using one-hot encoding across all recognized reactor categories, as described in Section 4.1.3.

2.1.4 Feedstock: Common Plastic Types

The six thermoplastic polymers most commonly studied in pyrolysis research - often referred to as the “Big Six” - are polyethylene (PE), polypropylene (PP), polystyrene (PS), polyvinyl chloride (PVC), polyethylene terephthalate (PET), and acrylonitrile-butadiene-styrene (ABS)

or high-impact polystyrene (HIPS). These polymers account for the vast majority of municipal plastic waste and exhibit markedly different pyrolysis behaviors [47, 20].

Polyolefins (PE and PP) are the most extensively studied feedstocks. Both consist of purely aliphatic carbon-hydrogen backbones and decompose primarily through random scission, yielding a broad distribution of hydrocarbons. High-density polyethylene (HDPE) and low-density polyethylene (LDPE) differ in branching degree and crystallinity but exhibit similar pyrolysis behavior, with liquid yields typically reaching 60 %-85 % at 500 °C-550 °C [33, 52]. Polypropylene produces similar product distributions but with a higher proportion of branched hydrocarbons and olefins due to its methyl-substituted backbone [23].

Polystyrene (PS) decomposes predominantly through depolymerization (unzipping), recovering the styrene monomer with selectivities of 50 %-80 % depending on temperature and reactor configuration. Total liquid yields commonly exceed 80 %, making PS one of the most attractive plastics for pyrolytic recycling [35, 38].

Polyvinyl chloride (PVC) presents a unique challenge because its backbone contains chlorine atoms. The first stage of PVC pyrolysis (at 250 °C-350 °C) involves dehydrochlorination, releasing hydrogen chloride (HCl) gas. The remaining polyene backbone then undergoes cracking at higher temperatures. The HCl release is problematic because it corrodes equipment and contaminates the liquid product with chlorinated compounds [47, 1]. Even small fractions of PVC in mixed plastic waste (1 %-5 %) can significantly degrade liquid product quality [47, 1].

Polyethylene terephthalate (PET) differs fundamentally from polyolefins due to its oxygen-containing ester backbone. Pyrolysis of PET produces significant amounts of benzoic acid, terephthalic acid, and carbon dioxide, resulting in lower hydrocarbon oil yields and higher char formation compared to polyolefins [47]. The distinct cracking mechanism of PET means that its pyrolysis behavior cannot be extrapolated from polyolefin data, and separate modeling is often necessary [47].

Scope of this work. The present thesis focuses exclusively on **non-catalytic (thermal) pyrolysis** of the above polymer types. Catalytic pyrolysis - in which zeolites, metal oxides, or other catalysts are introduced to alter the product distribution - represents a separate and substantially more complex domain requiring additional features (catalyst type, acidity, loading, pore structure) that are beyond the scope of this database [27, 28]. By restricting the analysis to non-catalytic conditions, catalyst-related confounding variables are eliminated, allowing the work to focus on the fundamental relationship between feedstock composition, process conditions, and product yields.

2.2 Machine Learning for Pyrolysis Prediction

2.2.1 Motivation: Machine Learning vs. Kinetic Models

The conventional approach to predicting pyrolysis product yields relies on kinetic models that describe the thermal decomposition of polymers through systems of differential equations. While kinetic models provide mechanistic insight, they suffer from several practical limitations that motivate the use of data-driven alternatives. First, kinetic parameters (activation

energies, pre-exponential factors, reaction orders) must be determined experimentally for each specific polymer-reactor combination, and they often do not transfer reliably across different experimental setups [17]. Second, comprehensive kinetic models that account for secondary cracking, heat and mass transfer effects, and the full product distribution are mathematically complex and computationally expensive to solve. Third, kinetic models require detailed knowledge of the feedstock composition at a molecular level, which is rarely available for real-world plastic waste streams [1].

Machine learning (ML) offers a complementary approach: given a sufficiently large and diverse dataset of experimental observations, ML algorithms can learn empirical relationships between input features (feedstock properties, process conditions) and output targets (product yields) without requiring explicit mechanistic knowledge. This data-driven paradigm is particularly attractive for cross-study meta-analyses, where the goal is to predict yields across heterogeneous reactor configurations and feedstock types that no single kinetic model could cover [6, 12].

However, the strength of ML - its ability to identify patterns in data regardless of their physical origin - is also its vulnerability. Without careful validation, ML models can exploit spurious correlations arising from the data collection process (e.g., within-paper correlations) rather than learning physically meaningful relationships. This tension between predictive flexibility and methodological rigor is a central theme of the present work.

2.2.2 Common Model Families

The ML-based pyrolysis literature employs a range of supervised regression algorithms. This section briefly describes the model families most relevant to this work, emphasizing their properties in the context of small-to-medium tabular datasets typical of pyrolysis meta-analyses.

Decision Trees and Random Forests. A decision tree recursively partitions the feature space using axis-aligned splits, selecting the split that minimizes a loss function (e.g., mean squared error) at each node. While individual decision trees are prone to overfitting, the Random Forest (RF) algorithm [7] mitigates this by training an ensemble of decorrelated trees on bootstrap samples of the data (bagging) and averaging their predictions. Random Forests provide built-in estimates of feature importance through permutation-based or impurity-based measures, making them a popular choice for interpretability. In the pyrolysis context, Cheng et al. [12] employed decision trees, achieving a training R^2 of 0.996 but only a test R^2 of 0.882 - a substantial gap that illustrates the overfitting tendency of unpruned decision trees.

Gradient Boosting (XGBoost, CatBoost). Gradient boosting constructs an ensemble of weak learners (typically shallow decision trees) sequentially, with each new tree trained to correct the residual errors of the ensemble built so far [18]. XGBoost [11] is a computationally efficient implementation that includes L1 and L2 regularization to control model complexity. CatBoost [41] extends gradient boosting with native support for categorical features and an ordered boosting scheme that reduces overfitting on small datasets. Gradient boosting methods, and XGBoost in particular, have emerged as the dominant model family in the ML-pyrolysis literature, achieving the best reported performance in the majority of recent studies [6, 55, 27, 4].

Support Vector Regression (SVR). SVR maps the input features into a high-dimensional space via a kernel function and finds the hyperplane that fits the data within a specified margin of tolerance (ε). SVR can capture nonlinear relationships through the choice of kernel (e.g., radial basis function) but requires careful tuning of the regularization parameter C, the kernel width γ, and the tolerance ε. Wang et al. [28] applied tree-based ensemble models (RF, GBR, XGBoost) to zeolite-catalytic pyrolysis prediction, with XGBoost achieving the best performance.

Artificial Neural Networks (ANNs). Feedforward neural networks [21] consist of interconnected layers of neurons with nonlinear activation functions, enabling them to approximate arbitrarily complex input-output relationships given sufficient network depth and width. ANNs require larger datasets for effective training compared to tree-based methods and are more sensitive to hyperparameter choices (learning rate, architecture, regularization). In the pyrolysis literature, ANNs have shown competitive but generally not superior performance to gradient boosting methods on the relatively small datasets available [4, 12].

2.2.3 Feature Engineering for Pyrolysis Data

The choice of input features is a critical design decision in ML-based pyrolysis prediction. Two fundamentally different strategies for representing the feedstock have been explored in the literature:

Plastic type encoding. The most common approach represents the feedstock as a set of binary (one-hot) or fractional variables indicating the polymer type (e.g., PE, PP, PS, PVC, PET). This encoding is intuitive and directly available from the experimental description, but it cannot represent novel or unknown waste streams that do not map cleanly onto a predefined set of polymer categories [6].

Ultimate analysis (elemental composition). An alternative approach characterizes the feedstock by its carbon (C), hydrogen (H), nitrogen (N), sulfur (S), and oxygen (O) content, determined experimentally or calculated from the known molecular structure of the polymer. This representation is more general - it can accommodate mixed or contaminated waste streams - and provides continuous features that may capture compositional gradients more effectively than discrete polymer labels. Cheng et al. [12] found that ultimate analysis features produced better predictions for liquid yields than polymer-type encoding, a finding they attributed to the ability of elemental composition to capture subtler variations in feedstock chemistry.

In practice, both representations carry complementary information. Polymer-type encoding captures qualitative differences in decomposition mechanism (e.g., the depolymerization of PS vs. the random scission of PE), while elemental composition captures quantitative differences in the hydrogen-to-carbon ratio, oxygen content, and heteroatom presence. The present work employs both representations and evaluates their relative contribution through feature importance analysis (Chapter 5).

2.2.4 Evaluation Metrics

Three metrics dominate the ML-pyrolysis literature:

- **Mean Absolute Error (MAE):** The average absolute difference between predicted and observed yields, expressed in weight percent (wt %). MAE is directly interpretable in physical units and is robust to outliers. Belden et al. [6] report MAE as their primary metric.
- **Root Mean Squared Error (RMSE):** The square root of the average squared difference. RMSE penalizes large errors more heavily than MAE and is sensitive to outliers.
- **Coefficient of Determination (R^2):** The proportion of variance in the target explained by the model. While R^2 facilitates comparison across datasets of different scales, it can be misleading when applied to heterogeneous datasets where a large fraction of the total variance is explained by a single feature (e.g., temperature), potentially masking poor performance on the residual variance [6].

A critical observation across the surveyed literature is the inconsistent reporting of these metrics. Several studies report only R^2 without distinguishing between training and test performance [10, 27, 28], while Cheng et al. [12] report a training R^2 of 0.996 alongside a test R^2 of 0.882 - a gap that illustrates how training-set metrics can be profoundly misleading. Other studies report test-set metrics but do not specify whether cross-validation was used and, if so, what splitting strategy was employed. This lack of methodological transparency severely limits the comparability of results across studies.

2.2.5 State of the Art: A Comparative Overview

To position the present work within the existing literature, seven recent publications (2022-2025) that apply machine learning to predict plastic pyrolysis product yields were surveyed. Table 2.1 summarizes the key methodological characteristics of each study.

Several patterns emerge from this comparison. First, **XGBoost is the dominant model**: it achieves the best or co-best performance in six of the seven studies. This consistency across independent research groups and datasets provides strong empirical support for the choice of gradient boosting as the primary model family in the present work.

Second, **dataset sizes remain modest**. The largest study by data point count is Belden et al. [6] with 325 observations from 39 papers, while by paper count, Cheng et al. [12] compiled data from 93 publications with 275-301 data points depending on the target variable. The reported R^2 values range from 0.85 to above 0.99, a spread that likely reflects differences in dataset composition, scope (catalytic vs. non-catalytic), and - critically - validation methodology rather than genuine differences in model quality.

Third, and most importantly, **no study uses group-based cross-validation**. All seven studies either apply standard random K-fold splitting, random train/test splits, or do not specify their validation strategy. This is remarkable given that every study compiles data from multiple independent publications - precisely the scenario in which within-paper correlations inflate performance estimates. The absence of group-aware validation across the entire field constitutes the methodological gap that the present work addresses.

The only study restricting its scope to non-catalytic pyrolysis is Cheng et al. [12], making it the most directly comparable to the present work in terms of scope. However, a closer

Table 2.1: Comparison of recent ML-based plastic pyrolysis yield prediction studies. All studies use random or unspecified train/test splitting; none employs group-based cross-validation to prevent within-paper data leakage. "n.r." indicates not reported.

Study	n	Best Model	Best Metric	CV Strategy	Scope
Alabdrabalnabi et al. (2022) [4]	∼94	DNN / XGB	$R^2 = 0.96$	not specified	co-pyrolysis
Belden et al. (2023) [6]	325	XGBoost	MAE = 9.1 wt%	10-fold CV + 10 % vaulted test	incl. catalytic
Cheng et al. (2023) [12]	275-301	Decision Tree	$R^2 = 0.996$ (train) / 0.882 (test)	70/30 random split	non-catalytic
Chen et al. (2024) [10]	n.r.	GBR	$R^2 = 0.91$	not specified	incl. catalytic
Li et al. (2025) [27]	511	XGBoost	$R^2 \geqslant 0.91$	80/20 + 5-fold CV	catalytic only
Wang et al. (2025) [28]	280	XGBoost	$R^2 = 0.85$	80/20 random	catalytic only
Zhang et al. (2025) [55]	267	XGBoost	$R^2 = 0.959$	80/20 random	incl. catalytic

n = number of data points; n.r. = not reported. R^2 values are test-set values where stated by the authors; Cheng et al. report both training ($R^2 = 0.996$) and test ($R^2 = 0.882$) metrics, revealing significant overfitting.

examination of their reported performance reveals significant overfitting: the training R^2 of 0.996 drops to a test R^2 of 0.882 on a 70/30 random split - a gap of 0.114 that indicates the decision tree model memorizes training data rather than learning generalizable patterns. Notably, this test performance was achieved under random splitting with no paper-level grouping; under grouped evaluation, the degradation would likely be even more severe. Furthermore, their 70/30 random split does not prevent within-paper data leakage, limiting the scientific value of even the test-set metric.

Belden et al. [6] provide the most transparent methodology among the surveyed studies: they employ a 10-fold cross-validation alongside a 10% vaulted (held-out) test set, and report MAE values rather than only R^2. They note that "the large standard deviation of the validation MAE suggests that the regression is susceptible to errors from random data selection" - an indirect acknowledgment of the instability caused by within-paper data leakage, even though the authors do not identify the underlying mechanism. However, their cross-validation strategy remains purely random with no paper-level grouping, and the stratification applied to the folds does not address the within-paper correlation structure.

2.3 Data Leakage in Machine-Learning-Based Science

2.3.1 The Reproducibility Crisis in ML-Based Science

The rapid adoption of machine learning across the quantitative sciences has been accompanied by growing concerns about methodological rigor. Kapoor and Narayanan [25] conducted a systematic survey of reproducibility issues in ML-based science and documented a widespread pattern of data leakage - defined as "a spurious relationship between the independent variables and the target variable that arises as an artifact of the data collection, sampling, or pre-processing strategy" that "usually leads to inflated estimates of model performance." Their survey identified leakage in 294 papers across 17 scientific fields, leading to what they characterize as "wildly overoptimistic conclusions" in affected studies.

The scale of this problem underscores a fundamental tension in applied ML: the flexibility that makes ML algorithms powerful predictors also makes them capable of exploiting artifacts in the data that have no physical basis. When such artifacts are present in both training and test sets - as occurs when the test set is not properly separated from the training set - the resulting performance metrics overstate the model's ability to generalize to truly new data.

2.3.2 Taxonomy of Data Leakage

Kapoor and Narayanan [25] introduce a taxonomy of eight leakage types organized into three categories: L1 (lack of clean separation between training and test data, including absent test sets, pre-processing on the full dataset, and duplicates), L2 (illegitimate features that serve as proxies for the target), and L3 (test set not drawn from the distribution of scientific interest). Of these eight types, **L3.2 - nonindependence between training and test samples - is directly applicable to ML-based pyrolysis prediction**. In a literature-compiled pyrolysis database, each source publication represents a coherent experimental unit: all data points from a given paper share the same reactor, the same feedstock batch, the same product recovery protocol, and the same systematic measurement biases. When a conventional random train/test split distributes data points from the same paper across both sets, the model exploits these paper-specific signatures - effectively memorizing the laboratory rather than learning the underlying thermochemistry.

Kapoor and Narayanan recommend "block cross-validation" as the appropriate remedy for L3.2 leakage: partitioning the dataset such that all observations from the same group (paper, patient, geographic unit) appear exclusively in either training or test [25, 44]. In the scikit-learn framework, this corresponds to `GroupKFold` cross-validation with the source paper identifier as the grouping variable - the strategy adopted throughout this work (Section 4.2.2).

2.3.3 Data Leakage in the Pyrolysis Literature: An Unaudited Domain

The 17 scientific fields identified by Kapoor and Narayanan as affected by data leakage include medicine, neuroimaging, genomics, satellite imaging, and computer security, among others. Notably, the fields of **chemistry, materials science, chemical engineering, and thermochemical conversion are absent from their survey** - not because leakage does not

occur in these domains, but because no systematic audit has been conducted [25]. An updated companion database maintained by the authors (as of May 2024) has expanded coverage to 30 fields and 648 affected papers, now including "sustainable energy" (through Geslin et al. [19], 2023, concerning illegitimate features), but still does not include pyrolysis, catalysis, or process engineering.

This absence is significant for two reasons. First, it means that the ML-based pyrolysis literature has not been subjected to the same methodological scrutiny that has revealed serious problems in other fields. Second, it means that the present work - by identifying and quantifying L3.2 leakage in the pyrolysis domain - extends the Kapoor-Narayanan framework into a new scientific field.

As documented in the preceding section (Table 2.1), all seven surveyed ML-pyrolysis studies use either standard random splitting or unspecified validation strategies. None acknowledges the existence of within-paper correlations as a potential source of inflated performance metrics. This uniform methodological blind spot is consistent with the pattern Kapoor and Narayanan observe across affected fields: data leakage is rarely intentional but arises from a lack of awareness of the dependence structure in the data [25].

The practical consequence is that the R^2 values of 0.85-0.99 reported in the pyrolysis ML literature may substantially overestimate the generalization performance of these models. Based on the cross-validation strategy comparison presented in Chapter 5, switching from standard random K-fold to paper-level GroupKFold increases the MAE by a factor of approximately two for the same model and dataset - a magnitude consistent with the performance drops documented in other fields when leakage is eliminated [25]. The implications of this finding for the interpretation of published results are discussed in detail in Chapter 6.

3 Data Collection and Database Construction

The construction of a comprehensive and reliable database is fundamental to developing accurate machine learning models for plastic pyrolysis. This chapter describes the systematic approach employed to collect, extract, and integrate experimental pyrolysis data from literature sources. The methodology prioritizes reproducibility, transparency, and data provenance to ensure that all samples can be traced back to their original publications.

3.1 Literature search strategy

A systematic literature search strategy was developed to identify relevant publications containing experimental plastic pyrolysis data. The approach combines automated database queries with manual screening to balance coverage and precision.

3.1.1 Web of Science query design

The primary literature database used for this work was Web of Science (WoS), chosen for its comprehensive coverage of peer-reviewed scientific publications and structured metadata. The query was designed to maximize recall while maintaining reasonable precision for manual review.

The search query targeted publications containing experimental pyrolysis data for plastic waste. Initial query formulation included broad terms related to plastic pyrolysis, thermal decomposition, and product characterization. Notably, exclusion terms (e.g., "NOT biomass", "NOT TGA") were deliberately omitted in the final query to avoid systematic bias and ensure comprehensive coverage. This decision was informed by preliminary screening showing that overly restrictive queries excluded relevant papers that mentioned excluded terms in passing but contained valuable pyrolysis data.

The final query executed on Web of Science returned approximately 7,400 publications. This large initial set reflects the decision to prioritize sensitivity (capturing all relevant papers) over specificity (excluding irrelevant papers), with the understanding that subsequent manual screening would filter false positives. The complete query string is reproduced in Appendix A.3. The query was executed in January 2025 on the Web of Science Core Collection (Science Citation Index Expanded), with no document type and timespan restrictions.

3.1.2 Screening and selection workflow

The large number of initial results (7,400 publications) necessitated a multi-stage screening workflow combining automated pre-selection with thorough manual validation. The workflow consisted of three main stages:

Stage 1: AI-assisted abstract screening. An automated screening script (`send_batches.py`) was developed to identify candidate papers likely to contain relevant experimental data. The

Gemini 2.5 Pro LLM was prompted to classify batches of 20 abstracts according to the following criteria:

- Laboratory-scale pyrolysis experiments (minimum 10 g sample size)
- Plastic or polymer feedstocks (PE, PP, PS, PVC, PET, mixed plastic waste)
- Quantitative yield data for at least one product fraction (oil/liquid, gas, or char)
- Temperature-resolved experiments (multiple temperatures reported)

The AI screening identified 200 candidate papers (2.7% selection rate).

Stage 2: AI-assisted full-text screening and data extraction. For the 200 candidate papers, full-text PDFs were retrieved where available. Each PDF was converted to Markdown text using `pymupdf4llm` and submitted individually to an LLM via a dedicated Python script (`send_pdf_batches.py`). The full-text screening prompt (reproduced in Appendix A.1) instructed the model to:

- Assess whether the paper meets the inclusion criteria based on the full text (not only the abstract)
- Extract quantitative yield data, process conditions, and feedstock properties into the 52-column database schema (Section 3.2.1)
- Classify the mass balance quality and data completeness for each experiment
- Flag entries where yield values were only available in figures without numeric labels (`chart_manual_needed`)

The structured JSON output was automatically appended to a staging Excel file for subsequent manual verification.

Stage 3: Manual validation and final inclusion. The AI-extracted data in the staging file were reviewed manually by the author on a paper-by-paper basis. Each data point was verified against the original publication, checking for:

- Correct transcription of yield values, temperatures, and feedstock compositions
- Plausibility of extracted mass balances
- Correct assignment of feedstock types and reactor classifications
- Values flagged as `chart_manual_needed` were read from figures where possible

Papers with unresolvable ambiguities or insufficient data quality were excluded at this stage.

Final statistics: Of the 200 AI-selected candidates, 100 papers (50%) were approved after the full-text AI screening and subsequent manual validation; verified data extraction was completed for 36 of these, contributing 240 data points to the database. The remaining 100 papers that did not meet inclusion criteria were excluded primarily due to missing yield data, TGA-only experimental setup, insufficient temperature variation, or unresolvable mass balance violations.

The newly identified papers contributed additional data points to the database. Combined with existing data from predecessor projects (see Section 3.3.1), the integrated database initially contained 1,208 experimental conditions. After a comprehensive data quality audit (Section 4.1.1), including correction of systematic encoding errors, removal of physically implausible

entries, and deduplication, the final database was reduced to 1,072 experimental conditions from 127 unique publications.

3.1.3 Positive control for AI-based screening

To validate the performance of the AI screening tool, a positive control experiment was conducted. The AI model was retrospectively applied to a subset of papers that had been manually screened and validated in earlier phases of the project. This control serves to quantify the AI's sensitivity (ability to identify relevant papers) and assess false positive rates.

Experimental design: A positive-control check was performed against 10 gold-standard papers that had been manually validated in earlier project phases: Mastral et al. (2002) [33], Williams & Williams (1997, 1999) [52], Onwudili et al. (2009) [38], Elordi et al. (2011) [16], and Jung et al. (2010) [23], among others. All 10 papers were correctly classified as "high priority" (approved) by the AI screening tool, confirming 100 % sensitivity on the control set.

Preliminary observations during the screening process indicated that the AI tool successfully identified all manually-validated papers (100% sensitivity), but generated a substantial number of false positives. The primary failure mode was misinterpretation of scope: the AI occasionally flagged papers mentioning mixed solid waste (MSW) without distinguishing between plastic-rich MSW and biomass-heavy MSW, or identified papers studying catalytic pyrolysis when the query targeted non-catalytic processes.

These findings informed prompt optimization for future screening iterations, with more explicit instructions regarding feedstock composition and experimental scope. The false positive rate, while non-negligible, was acceptable given the subsequent manual review stage, which served as the definitive filter.

3.2 Data extraction protocol

Once papers were identified as relevant through the screening workflow (Section 3.1), experimental data were systematically extracted and recorded in a structured format. This section describes the variables extracted, the conventions adopted for harmonizing heterogeneous reporting practices across literature, and the quality assurance measures implemented to ensure data integrity.

3.2.1 Extracted variables and units

The data extraction protocol targeted a comprehensive set of variables describing feedstock characteristics, process conditions, and product yields. The goal was to capture all information necessary for machine learning modeling while maintaining traceability to the original publication.

Feedstock composition variables:

- Plastic type(s): Weight fractions (wt%) for HDPE, LDPE, LLDPE, PE (unspecified), PP, PS, PVC, PET, HIPS, ABS

- Ultimate analysis: Elemental composition (wt%) for C, H, N, O, S, Cl, Br, Sb
- Proximate analysis: Volatile matter, fixed carbon, ash content (wt%, dry basis)
- Sample origin: Virgin polymer, post-consumer waste, specific waste stream

For pure single-polymer studies, the relevant plastic fraction was recorded as 100 wt%. For mixed plastics, individual fractions were recorded as reported. When only a polymer category was specified (e.g., "PE" without distinguishing HDPE/LDPE/LLDPE), the entry was recorded in the general PE column.

Process condition variables:

- Pyrolysis temperature: Final/maximum temperature (°C)
- Heating rate: Ramp rate to target temperature ($°C\,min^{-1}$)
- Residence time: Vapor residence time (s), solid residence time (min) where reported
- Particle size: Average or range (mm)
- Feed/sample size: Mass of feedstock processed (g or kg)
- Reactor type: Fixed bed, fluidized bed, stirred batch, semi-batch, continuous, etc.
- Operating mode: Batch, semi-batch, continuous

Process temperatures were recorded as reported in the publication, typically representing the final isothermal hold temperature or peak temperature for non-isothermal experiments. Heating rates were extracted where available but exhibited high missingness (see Section 4.1.2). Residence times were distinguished between vapor residence time (relevant for product secondary reactions) and solid residence time (relevant for continuous reactors), though both variables showed low reporting frequency.

Product yield variables (primary targets):

- Oil/liquid yield (wt%, dry ash-free basis)
- Gas yield (wt%, dry ash-free basis)
- Char/solid residue yield (wt%, dry ash-free basis)

Product composition variables (secondary, optional):

- Liquid fractions: Wax, aromatics, BTX (benzene-toluene-xylene), styrene, C_5-C_{12} range, C_{13}-C_{20} range
- Gas composition: C_1 (methane), C_2 (ethane/ethylene), C_3, C_4, trace gases

Product composition data exhibited substantial missingness (typically 20-30% availability) and were recorded opportunistically for future multi-task modeling but not used as primary features or targets in the present work.

Metadata variables (for traceability):

- Source database: Origin of the data point (IZTECH_v5, dataset_2, manual_2026, AI_automation_2026)
- Source paper ID: Unique identifier linking to the original publication
- Entry ID: Unique row identifier within the database

- Reactor identification: Lab group or specific reactor name (enables grouping of data from same experimental setup)

Metadata fields are critical for provenance tracking and for implementing paper-level cross-validation (see Section 4.2.2) to avoid data leakage in machine learning model evaluation.

In total, the database schema comprises 52 columns capturing feedstock, process, product, and provenance information. Not all variables are populated for every data point due to heterogeneous reporting practices in literature (see Section 4.1.2 for missingness analysis).

3.2.2 Yield definitions and harmonization assumptions

A major challenge in integrating literature data is the heterogeneity in how product yields are defined and reported. Different research groups adopt different conventions for normalizing yields, handling moisture and ash, and defining product fractions. To enable meaningful comparison and modeling, harmonization assumptions were necessary.

Yield basis: Plastic pyrolysis yields are routinely reported without explicit statement of the normalization basis (as-received, dry, or dry ash-free). Since virgin thermoplastics have negligible ash content (<1 wt%), the distinction between bases is negligible for the majority of database entries. Yields were therefore recorded as reported by the respective authors, without basis correction. Yield integrity is assessed through the mass balance quality check described below.

Product fraction definitions:

- **Oil/liquid yield:** Condensable organic fraction, typically collected in condensers or cold traps. Includes waxes, oils, and tars. Some papers distinguish "oil" (liquid at room temperature) from "wax" (semi-solid); when disaggregated, these were summed to obtain total condensable yield.
- **Gas yield:** Non-condensable fraction, typically C_1-C_4 hydrocarbons and light gases (H_2, CO, CO_2). Collected in gas bags or measured by gas chromatography.
- **Char/solid residue yield:** Solid carbonaceous residue remaining in the reactor after pyrolysis. Includes fixed carbon and ash.

Ideally, the sum of oil, gas, and char yields should equal 100 wt% (mass balance closure). In practice, experimental mass balances often deviate due to measurement uncertainties, uncollected aerosols, or losses during product handling.

Mass balance quality assessment: For each data point, the total yield (oil + gas + char) was calculated and used as a data quality indicator:

- **Good quality:** 95 wt% $\leqslant$ total yield $\leqslant$ 105 wt% (acceptable experimental uncertainty)
- **Acceptable quality:** 85 wt% $\leqslant$ total yield $<$ 95 wt% or 105 wt% $<$ total yield $\leqslant$ 115 wt%
- **Poor quality:** Total yield $<$ 85 wt% or $>$ 115 wt%

Data points with poor mass balance were flagged but not automatically excluded, as the decision to exclude data was made during preprocessing based on multiple quality criteria (see

Section 4.1.1). Of the 1,072 data points in the final database, the mass balance distribution was assessed after the comprehensive audit described in Section 4.1.1.

Harmonization of ambiguous reports: When multiple fractions were reported separately (e.g., "heavy oil" and "light oil"), they were summed to the appropriate aggregate fraction. "Liquid," "condensate," and "wax" were all included under the oil yield category.

These harmonization decisions introduce minor systematic uncertainties but were necessary to achieve a unified dataset.

3.3 Database integration

The final database results from integrating multiple data sources, each with different origins, formats, and quality characteristics. This section describes the source databases, the procedures for handling overlaps and conflicts, and the database schema designed to preserve provenance information.

3.3.1 Source databases and overlap handling

The integrated database comprises data from four primary sources:

1. IZTECH_v5 (509 data points, 47.4%): A dataset originating from İzmir Institute of Technology, compiled in prior work [24]. This database primarily contains data from publications on thermal pyrolysis of waste plastics, with a focus on Turkish and European research groups. Data quality is generally good, with complete yield triplets (oil, gas, char) for the majority of entries.

2. dataset_2 (245 data points, 22.9%): A complementary dataset from predecessor work focusing on different literature sources to avoid overlap with IZTECH_v5. This dataset includes more recent publications (post-2018) and covers a broader range of reactor types including continuous and semi-batch systems. However, dataset_2 contains only oil yield data without corresponding gas or char yields. As the multi-output regression framework requires complete yield triplets for consistent model evaluation and mass-balance-based quality assurance, dataset_2 contributes zero samples to the final preprocessed dataset. Training a dedicated oil-only model on an expanded dataset including these entries represents a natural extension of this work (Section 7.3.1) but falls outside the present scope, which prioritizes methodological rigor in cross-validation over dataset size. The Belden et al. dataset [6], which is a subset of dataset_2, was likewise excluded to prevent duplication.

3. manual_2026 (78 data points, 7.3%): Data extracted from papers identified during the initial manual screening phase of this project. These represent papers not covered in the predecessor databases, primarily publications from 2023-2024.

4. AI_automation_2026 (240 data points, 22.4%): Data extracted from the AI-assisted literature screening and full-text extraction pipeline described in Section 3.1.2. After manual validation and extraction, these papers contributed 240 new data points covering diverse reactor configurations and plastic types not well represented in the predecessor databases.

Total: 1,072 data points from 132 unique publications (after quality audit described in Section 4.1.1).

Overlap detection and handling: Each paper was assigned a unique identifier based on its DOI or, if unavailable, a standardized citation string. Since the predecessor datasets were designed to cover non-overlapping literature sources, true duplicates were rare. When identified, the entry from the source with more complete metadata was retained.

Non-overlapping nature: The predecessor databases (IZTECH_v5 and dataset_2) were explicitly designed to be non-overlapping, covering different literature sources. The newly added data (manual_2026 and AI_automation_2026) focus on recent publications post-dating the predecessor databases. Therefore, the overlap rate in the integrated database is minimal, and the database can be considered a union of largely independent literature sources.

3.3.2 Schema, identifiers, and provenance tracking

The integrated database follows a flat table structure (one row per experimental condition) with 52 columns. Key provenance fields include `Source_Database`, `Source_Paper_ID`, `Source_Paper`, `Entry_ID`, and `Mass_Balance_Quality`. The `Source_Paper_ID` field enables paper-level GroupKFold cross-validation to prevent data leakage (Section 4.2.2), and the `Source_Database` field allows subgroup analysis across literature sources. The database is maintained as a Microsoft Excel file (`Master_Pyrolysis_Database_v2.xlsx`) for manual inspection and exported to CSV for machine learning workflows.

4 Machine Learning Methodology

This chapter describes the machine learning methodology employed to develop predictive models for plastic pyrolysis product yields. It begins with the data preprocessing pipeline that transforms the raw integrated database (Chapter 3) into a clean, feature-engineered dataset. The cross-validation strategy - paper-level grouped splitting - which constitutes the central methodological contribution of this work and addresses a pervasive data leakage problem in the pyrolysis modeling literature, is then presented. The chapter continues with the model families evaluated and concludes with the interpretability framework used to validate that learned relationships are physically meaningful.

Figure 4.1 provides an overview of the complete modeling pipeline, from database construction through evaluation.

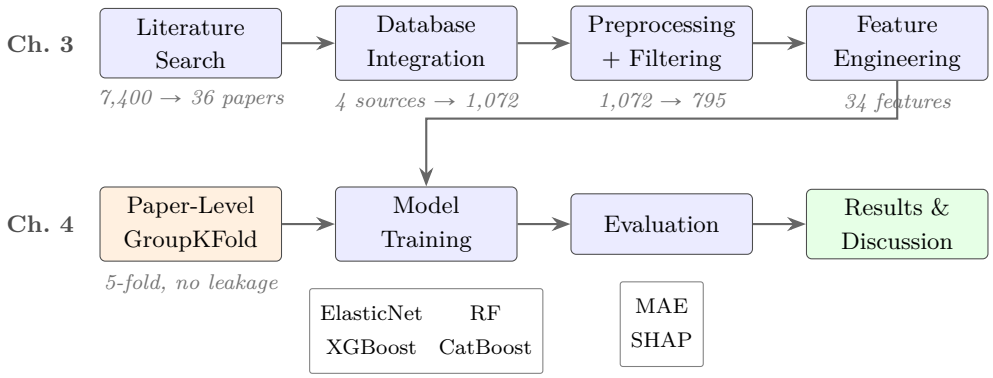

Figure 4.1: Overview of the modeling pipeline. The workflow spans database construction (Chapter 3), preprocessing and model development (this chapter), and evaluation (Chapter 5). The paper-level GroupKFold cross-validation strategy (highlighted in orange) constitutes the central methodological contribution.

4.1 Data Preprocessing

The raw integrated database comprises 1,072 data points from 127 unique source papers with 52 columns capturing feedstock composition, process conditions, product yields, and metadata (Chapter 3). Before machine learning models can be trained, this heterogeneous dataset requires systematic preprocessing to handle incomplete records, encode categorical variables, impute missing values, and select informative features. The preprocessing pipeline is implemented in a standalone Python script to ensure full reproducibility; the implementation uses Python 3.11 with scikit-learn, XGBoost, CatBoost, SHAP, pandas, and NumPy.

4.1.1 Filtering pipeline

The preprocessing pipeline applies two sequential filters to the raw database, each targeting a specific data quality requirement. The philosophy guiding these filters is conservative:

only data points that are demonstrably unusable for supervised learning are excluded, while borderline cases are retained to preserve sample size and capture the genuine heterogeneity of literature-reported pyrolysis data.

Step 1: Complete yield filter. Supervised regression requires complete target variables. Data points with exactly two of the three yield fractions reported were retained: the missing fraction was calculated as 100 wt% minus the sum of the two reported yields. Data points missing two or more yield fractions were excluded. This filter reduces the database from 1,072 to 817 samples (76.2 % retention). The excluded 255 data points (23.8 %) lack at least two yield values, typically because the source paper reported only oil yield without quantifying gas or char, or because char was described qualitatively (e.g., "negligible residue") rather than measured.

Imputing missing target variables was explicitly rejected. Unlike missing features - where median or model-based imputation introduces bounded uncertainty - imputing targets would fabricate the very quantities the model is trained to predict, fundamentally undermining the supervised learning objective.

Step 2: Mass balance filter. For data points with all three yields reported, the total yield was computed as

$$Y_{\text{total}} = Y_{\text{oil}} + Y_{\text{gas}} + Y_{\text{char}} \tag{4.1}$$

and retain only samples satisfying $85\,\text{wt\%} \leqslant Y_{\text{total}} \leqslant 115\,\text{wt\%}$. This range accommodates typical experimental uncertainty (5-10 wt % deviation from perfect closure) [6, 14] while excluding gross errors such as unit conversion mistakes, incomplete product recovery, or transcription errors. The filter reduces the dataset from 817 to 795 samples (97.3 % retention), excluding only 22 data points with severely implausible mass balances.

The combined effect of both filters is a reduction from 1,072 raw entries to 795 modeling-ready samples (74.2 % overall retention). Table 4.1 summarizes the pipeline.

Table 4.1: Sequential data filtering pipeline applied to the raw integrated database. Each step is applied to the output of the preceding step.

Filter step	Samples in	Samples out	Retention
Raw database	-	1,072	-
Complete yields (oil, gas, char)	1,072	817	76.2 %
Mass balance 85-115 wt%	817	795	97.3 %

Of the 795 retained samples, the majority originate from IZTECH_v5 and AI_automation_2026, with a smaller contribution from manual_2026. dataset_2, which contains 245 raw entries, contributes zero samples to the final dataset because it records only oil yield without corresponding gas or char data, and is therefore excluded entirely by the complete yield filter. The relatively high retention rate for the IZTECH_v5 subset reflects its origin from a curated predecessor database with stricter inclusion criteria.

4.1.2 Missing value analysis and imputation

After filtering, the 795-sample dataset exhibits substantial and heterogeneous missingness across feature variables. This is an inherent characteristic of literature-derived datasets: different research groups report different subsets of experimental parameters depending on their equipment, conventions, and the focus of their study.

Missingness by feature category. Table 4.2 summarizes the approximate missingness rates for key feature groups.

Table 4.2: Approximate missingness rates for key features in the 795-sample dataset. Features with >90 % missingness (e.g., vapor and solid residence time) were excluded from the feature set.

Category	Feature	Missingness (%)
Process	Heating rate (°C/min)	∼82
Process	Feed size (g)	∼68
Process	Particle size (mm)	∼37
Process	Temperature (°C)	0
Elemental	C, H, N, O, S, Cl, Br, Sb (wt%)	26-45
Feedstock	Plastic type fractions (wt%)	∼25
Categorical	Reactor type	∼64
Categorical	Batch/continuous	∼64

The missingness pattern is not random but reflects reporting conventions: batch reactor studies rarely report vapor residence time; older publications omit elemental analysis; and heating rate is frequently unreported when constant ("furnace heated to target temperature"). This structured missingness - classified as Missing Not At Random (MNAR) in the statistical literature [29] - complicates imputation but does not preclude it, particularly for tree-based models that can learn to handle imputed values as a distinct signal.

Imputation strategy. A pragmatic, category-specific imputation approach is adopted:

- **Continuous features** (plastic fractions, elemental analysis, process parameters): Missing values are replaced with the column median computed on the training fold only [6]. Median imputation is preferred over mean imputation for its robustness to outliers. Zero-filling was considered for plastic fractions (interpreting missing as "absent") but rejected because missingness more often indicates incomplete reporting rather than confirmed absence.
- **Batch/continuous mode**: Unknown entries are encoded as -1, distinguishing them from Batch (= 0), Semi-Batch (= 0.5), and Continuous (= 1). This three-valued encoding allows tree-based models to route unknown-mode samples along a separate decision path. Approximately 33.3 % of samples have known batch/continuous status.

Imputation within cross-validation. To prevent information leakage through imputation, median values are computed on the training fold and applied to both training and validation folds independently within each cross-validation iteration. This ensures that no information from the validation set influences the imputed training values. The same training-fold median

values are applied to unseen samples at prediction time, ensuring consistency between the cross-validation and deployment settings.

4.1.3 Categorical feature encoding

Two categorical features are included in the modeling dataset: reactor type and batch/continuous mode. Each requires a specific encoding strategy to be compatible with the model families evaluated. Since this work is restricted to non-catalytic (thermal) pyrolysis, no catalyst feature is included.

Reactor type: one-hot encoding. The database contains 11 named reactor type categories including Batch, Fixed Bed, Fluidized Bed, Tube, Semi-Batch, Autoclave, Rotary Kiln, Auger, CSTR, Ball-Mill, and Other. Samples without a recorded reactor type (approximately 47.5 % of the dataset) are assigned to an explicit "Unknown" category, yielding 12 categories in total. One-hot encoding via `pd.get_dummies` produces 11 binary indicator columns (with Batch as the implicit reference category, absorbed into the intercept). The "Unknown" category is *deliberately* retained rather than dropped as a reference: missing reactor information is itself an informative signal because papers that do not report a reactor type may systematically differ from those that do. Approximately 36 % of samples have a known reactor type; the remaining 64 % are assigned to the "Unknown" category.

One-hot encoding was chosen over ordinal encoding because reactor types have no inherent ordering, and over target encoding to avoid data leakage. The 11 binary reactor columns are sparse by construction: any given sample has at most one non-zero reactor column.

Batch/continuous mode. These are encoded as numeric variables as described in Section 4.1.2, yielding one feature each.

4.1.4 Final feature set

After filtering, imputation, and encoding, the final modeling dataset comprises 795 samples with 34 input features and 3 target variables. Table 4.3 provides the feature breakdown.

Table 4.3: Composition of the 34-feature input space used for model training. Features are grouped by category.

Category	Features	Count
Plastic composition	HDPE, LDPE, LLDPE, PE, PP, PS, PVC, PET, HIPS, ABS (wt%)	10
Ultimate analysis	C, H, N, O, S, Cl, Br, Sb (wt%)	8
Process conditions	Temperature, Heating rate, Particle size, Feed size	4
Batch/continuous	Numeric encoding (−1, 0, 0.5, 1)	1
Reactor type	One-hot encoded (11 categories incl. Unknown)	11
Total features		**34**
Targets	Oil yield, Gas yield, Char yield (wt%)	3

Features with >90 % missingness (vapor residence time, solid residence time, product composition variables) were excluded from the feature set entirely, as they provide information for fewer than 10 % of samples and would be dominated by imputed values. The decision to retain features with moderate missingness (37-82 %) rather than dropping them is evaluated through a particle size ablation study: the model is trained and evaluated twice under identical GroupKFold conditions - once including particle size, once excluding it - and the resulting change in validation MAE quantifies the genuine predictive contribution of this partially-missing feature. Results are presented in Section 5.3.4.

Each of the three target variables - oil yield, gas yield, and char yield in wt % - is modeled as an independent single-output regression problem. No transformations (e.g., log-transform, normalization) are applied to targets, as yields are naturally bounded between 0 and 100 wt % and do not exhibit pathological distributional properties.

4.2 Cross-Validation Strategy

The choice of cross-validation strategy is the most consequential methodological decision in this work. Conventional random splitting of the 795 samples into training and validation folds ignores the hierarchical structure of the data - namely, that multiple data points originate from the same source paper and therefore share systematic characteristics that are not representative of truly new experimental conditions. This section formalizes the resulting data leakage problem, describes the paper-level grouped cross-validation strategy adopted to eliminate it, and compares it against two alternative strategies evaluated in this work.

4.2.1 Data leakage in cross-study meta-analyses

As established in Section 2.3, Kapoor and Narayanan's type L3.2 leakage [25] - nonindependence between training and test samples - applies directly to literature-compiled pyrolysis databases. Each source paper contributes multiple data points that share the same experimental apparatus, feedstock batch, measurement protocol, and operator-specific systematic biases. When a standard random train/test split distributes these correlated observations across both sets, the model can exploit paper-specific systematic offsets rather than learning generalizable physical relationships, inflating performance metrics [25].

4.2.2 Paper-level GroupKFold cross-validation

To eliminate L3.2 leakage, this work implements **5-fold GroupKFold cross-validation** with the source paper identifier (`Source_Paper_ID`) as the grouping variable. This ensures that all data points from any given paper appear exclusively in either the training or the validation fold within each cross-validation iteration - never in both.

Implementation details:

- **Grouping variable:** `Source_Paper_ID` uniquely identifies each of the 132 source publications in the database. Each paper is assigned a unique integer identifier used as the grouping variable for GroupKFold.

- **Number of folds:** $k = 5$, providing a balance between robust performance estimation (averaging over five independent train/validation splits) and computational cost.
- **Fold assignment:** Papers - not individual samples - are distributed across folds. Each fold holds out approximately 20 % of papers, which translates to 15-25 % of samples per fold due to the heterogeneous number of data points per paper (median: 4; range: 1-28).
- **Software:** `sklearn.model_selection.GroupKFold` with `n_splits=5` and `groups=df['Source_Paper_ID']`.

GroupKFold implements precisely the "block cross-validation" strategy recommended by Kapoor and Narayanan [25] for addressing nonindependence between training and test samples. The "blocks" in this case are the source papers, which represent the natural unit of experimental dependence in a literature-compiled database.

4.2.3 Alternative strategies evaluated

To quantify the magnitude of data leakage under conventional evaluation approaches, two additional cross-validation strategies are evaluated using identical datasets and model configurations:

Standard Random KFold. Samples are shuffled randomly and assigned to five folds without regard for paper membership. This represents the evaluation approach used in the majority of published ML-based pyrolysis studies. Under this strategy, data points from the same paper are expected to appear in both training and validation folds, introducing L3.2 leakage.

Educated Stratified KFold. Approximately 20 % of each paper's data points are placed in the validation fold, with the remaining 80 % in training. This strategy ensures that every paper is represented in both folds but preserves *partial* within-paper leakage: the model still observes some data points from a paper during training before being evaluated on other points from the same paper.

The three strategies form a spectrum of leakage severity: Standard Random (full leakage) → Educated Stratified (partial leakage) → Paper-Level GroupKFold (no leakage). Comparing model performance across all three strategies quantifies the leakage effect and provides an empirical measure of how much reported performance in existing literature may be inflated by within-paper correlations.

4.2.4 Overfitting gap metric

To complement the cross-validation comparison, the *overfitting gap* is defined as a diagnostic metric for model generalization:

$$\text{Overfitting Gap} = \frac{\text{MAE}_{\text{val}} - \text{MAE}_{\text{train}}}{\text{MAE}_{\text{val}}} \times 100\,\% \tag{4.2}$$

where $\text{MAE}_{\text{train}}$ and MAE_{val} are the mean absolute errors on the training and validation sets, respectively, averaged across the five GroupKFold splits. A gap near zero indicates that the

model generalizes well (training and validation performance are comparable); a large gap indicates substantial memorization of training data that does not transfer to unseen papers [21]. This metric is reported for all models in Chapter 5 (Section 5.2.2).

4.3 Machine Learning Models

Five modeling approaches are evaluated spanning the complexity spectrum from a linear baseline to non-linear ensemble methods. All models are trained and evaluated under identical 5-fold paper-level GroupKFold conditions to enable fair comparison. Each of the three targets (oil, gas, char yield) is modeled independently as a single-output regression problem.

4.3.1 Elastic Net (linear baseline)

Elastic Net [56] combines L1 (Lasso) and L2 (Ridge) regularization into a single linear regression model:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \alpha \left[\lambda \|\boldsymbol{\beta}\|_1 + (1-\lambda)\|\boldsymbol{\beta}\|_2^2\right] \tag{4.3}$$

where α controls overall regularization strength and $\lambda \in [0, 1]$ balances the L1 and L2 penalties. The L1 penalty drives weak feature coefficients to exactly zero, performing implicit feature selection, while the L2 penalty stabilizes coefficients for correlated features (e.g., HDPE and LDPE weight fractions).

Elastic Net serves as an interpretable baseline: if non-linear models substantially outperform it, the improvement quantifies the importance of non-linear effects (e.g., temperature threshold behavior, polymer interaction effects) in the data. Features are standardized to zero mean and unit variance prior to fitting to ensure fair penalization across features with different scales.

Implementation: `sklearn.linear_model.ElasticNet` with $\alpha = 0.01$, $\lambda = 0.5$, and `max_iter=5000`.

4.3.2 Random Forest

Random Forest aggregates predictions from an ensemble of decorrelated decision trees, each trained on a bootstrap sample of the data with random feature subsets considered at each split [7]. The ensemble averaging reduces variance relative to individual trees, yielding robust predictions with limited hyperparameter sensitivity.

For the final evaluation, a *regularized* Random Forest configuration is employed that deliberately trades training accuracy for improved out-of-paper generalization. The regularization is implemented through three mechanisms: capping tree depth (`max_depth`), enforcing minimum leaf size (`min_samples_leaf`), and restricting the number of candidate features per split (`max_features`).

Hyperparameters (regularized configuration):

- `n_estimators = 600`: Number of trees in the ensemble

- `max_depth = 12`: Maximum tree depth (limits model capacity)
- `min_samples_leaf = 5`: Minimum samples in a leaf node (prevents overfitting to individual data points)
- `max_features = 'sqrt'`: Number of features considered per split ($\approx \sqrt{35} \approx 6$ features)
- `random_state = 42`, `n_jobs = -1`

Implementation: `sklearn.ensemble.RandomForestRegressor`

4.3.3 XGBoost

XGBoost (Extreme Gradient Boosting) [11] builds an ensemble of decision trees sequentially, where each tree is trained on the residual errors of the preceding ensemble. Gradient boosting with second-order Taylor expansion of the loss function, combined with explicit regularization on leaf weights, produces models that consistently achieve state-of-the-art performance on structured tabular data.

Hyperparameters (regularized configuration):

- `n_estimators = 600`: Number of boosting rounds
- `learning_rate = 0.05`: Step size shrinkage (smaller values require more trees but reduce overfitting)
- `max_depth = 4`: Maximum tree depth (deliberately shallow to limit model complexity)
- `subsample = 0.8`: Fraction of samples used per boosting round (stochastic gradient boosting)
- `colsample_bytree = 0.8`: Fraction of features used per tree
- `reg_lambda = 3.0`: L2 regularization on leaf weights (strengthened from default of 1.0)
- `reg_alpha = 0.0`: L1 regularization on leaf weights
- `random_state = 42`, `n_jobs = -1`

The regularization parameters - particularly `max_depth = 4` and `reg_lambda = 3.0` - were selected via Bayesian hyperparameter search with paper-level GroupKFold as the inner cross-validation loop, ensuring that hyperparameter selection itself does not exploit within-paper leakage.

Implementation: `xgboost.XGBRegressor` with `objective = 'reg:squarederror'`.

4.3.4 CatBoost

CatBoost [41] is a gradient boosting framework that introduces ordered boosting and native handling of categorical features. Unlike XGBoost, which requires explicit one-hot or ordinal encoding of categorical variables, CatBoost can process categorical features directly, computing target statistics with built-in regularization to prevent target leakage during encoding.

Hyperparameters:

- `iterations = 600`: Number of boosting rounds

- `learning_rate = 0.05`: Step size shrinkage
- `depth = 6`: Maximum tree depth
- `l2_leaf_reg = 3.0`: L2 regularization on leaf values
- `random_seed = 42`

CatBoost is included in this comparison specifically because the reactor type feature - encoded as 11 one-hot columns for the other models - can be supplied directly as a categorical variable to CatBoost, potentially enabling more efficient use of this sparse, high-cardinality feature.

Implementation: `catboost.CatBoostRegressor`

4.3.5 Ensemble model

A simple averaging ensemble combines the predictions of Random Forest, XGBoost, and CatBoost:

$$\hat{y}_{\text{ensemble}} = \frac{1}{3}\left(\hat{y}_{\text{RF}} + \hat{y}_{\text{XGB}} + \hat{y}_{\text{CB}}\right) \tag{4.4}$$

Ensemble averaging reduces prediction variance when constituent models make partially uncorrelated errors. If the three tree-based models capture similar patterns (i.e., errors are positively correlated), the ensemble will offer marginal improvement at best. The ensemble serves as a diagnostic: if it substantially outperforms the best individual model, the constituent models complement each other; if not, they are largely redundant.

No weighting optimization was performed - the simple average is parameter-free and avoids additional overfitting risk from weight tuning on the validation set.

4.3.6 Hyperparameter tuning approach

Hyperparameters for the tree-based models were selected via Bayesian hyperparameter search (Optuna) [2], with the inner cross-validation loop using paper-level GroupKFold (5-fold) to ensure that hyperparameter selection reflects realistic out-of-paper generalization rather than within-paper interpolation. This is a critical detail: performing hyperparameter tuning with random cross-validation and then reporting final performance under GroupKFold would allow leakage to influence model configuration, undermining the integrity of the evaluation.

The key hyperparameters tuned were tree depth (`max_depth`), regularization strength (`reg_lambda` for XGBoost, `l2_leaf_reg` for CatBoost, `min_samples_leaf` for RF), and the number of boosting rounds (`n_estimators/iterations`). Learning rate was fixed at 0.05 following preliminary experiments showing stable convergence across all targets at this value.

4.4 Evaluation Metrics

Mean absolute error (MAE) is adopted as the primary evaluation metric:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.5}$$

MAE has the same units as the target (wt %), is robust to outliers, and aligns with the achievable precision of literature-reported yields (1-2 wt %). MAE is computed on each of the five validation folds and averaged to obtain $\mathrm{MAE}_{\mathrm{CV}}$; the standard deviation across folds quantifies sensitivity to the particular split.

Root mean squared error (RMSE) and the coefficient of determination (R^2) are reported for completeness and literature comparison but are *not* used for model selection. R^2 can be misleadingly high even when absolute errors are large if the target has high variance; for instance, an MAE of 12 wt% on oil yield (standard deviation ≈20 wt%) corresponds to $R^2 \approx 0.55$.

4.5 Model Interpretability

Predictive accuracy alone is insufficient for scientific applications. Understanding *which features drive predictions* and *whether the learned relationships are consistent with known pyrolysis physics* is essential for building trust, identifying failure modes, and extracting scientific insight from the models. Two complementary interpretability methods are employed.

4.5.1 Permutation importance

Permutation importance [7] quantifies the contribution of each feature by measuring the increase in prediction error when that feature's values are randomly shuffled, breaking its relationship with the target while preserving the marginal distributions of all other features.

For each feature f, the permutation importance is defined as:

$$\mathrm{PI}(f) = \mathrm{MAE}_{\mathrm{permuted}}(f) - \mathrm{MAE}_{\mathrm{baseline}} \tag{4.6}$$

where $\mathrm{MAE}_{\mathrm{baseline}}$ is the model error on the unperturbed validation set and $\mathrm{MAE}_{\mathrm{permuted}}(f)$ is the error after shuffling feature f. The permutation is repeated $m = 10$ times per feature, and the mean and standard deviation of the importance score are reported.

Permutation importance offers three key advantages over the impurity-based importance provided by tree models [7, 48, 36]: (i) it is model-agnostic and can be applied to any predictive model; (ii) it avoids the systematic bias of impurity-based measures, which artificially favor high-cardinality and correlated features; and (iii) it is computed on held-out validation data, reflecting genuine out-of-sample importance rather than training artifacts.

Implementation: `sklearn.inspection.permutation_importance` with `n_repeats = 10` and `random_state = 42`, computed on each GroupKFold validation fold and averaged.

4.5.2 SHAP analysis

While permutation importance reveals *which* features matter, SHAP (SHapley Additive exPlanations) [32] additionally reveals *how* each feature influences individual predictions - both the magnitude and direction of its effect.

SHAP values are grounded in cooperative game theory. For a given prediction $\hat{y}_i$, the SHAP value $\phi_j^{(i)}$ of feature j represents the average marginal contribution of that feature across all possible orderings of features:

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{p} \phi_j^{(i)} \tag{4.7}$$

where ϕ_0 is the expected prediction (base value) and p is the number of features. The SHAP values satisfy three desirable properties: local accuracy (they sum to the predicted value), missingness (absent features contribute zero), and consistency (increasing a feature's true contribution never decreases its SHAP value).

For tree-based models, the `TreeExplainer` algorithm is used [31], which computes exact SHAP values in polynomial time by exploiting the tree structure. SHAP beeswarm plots - where each dot represents one sample, horizontal position encodes the SHAP value, and color encodes the feature value - provide a compact visual summary of feature effects across the entire dataset.

Implementation: `shap.TreeExplainer` applied to the trained XGBoost and Random Forest models, evaluated on a representative subset of 200 randomly sampled validation-fold observations.

4.5.3 Physical validation of feature directions

A key application of SHAP analysis is the validation of learned feature-target relationships against established pyrolysis chemistry. For each of the most important features, the direction of the SHAP effect is compared (positive or negative) with the expected physical behavior:

- **Temperature**: Higher temperatures should *decrease* oil yield and *increase* gas yield due to secondary cracking of primary liquid products into lighter gas-phase species [47].
- **Feedstock composition**: Polystyrene (PS) is known to produce high liquid yields (predominantly styrene monomer) at moderate temperatures, while polyethylene (PE) produces higher wax fractions that are partially classified as liquid [47].
- **Particle size**: Larger particles exhibit internal temperature gradients that can suppress secondary cracking (higher oil yield) or promote repolymerization (higher char yield), depending on the temperature regime [47].

If the SHAP directions align with known physics, this provides confidence that the model captures genuine causal mechanisms rather than spurious correlations. Disagreements between SHAP directions and physical expectations warrant investigation and may reveal either model pathologies or genuinely surprising structure in the data.

This physical validation step bridges the gap between statistical model performance and domain-level scientific understanding, a connection that is frequently absent in the ML-based pyrolysis literature but essential for the scientific credibility of data-driven predictions.

4.6 Per-Paper Error Analysis

Beyond aggregate cross-validation MAE, this work analyzes prediction error at the level of individual source papers. For each paper in the validation set across all five GroupKFold folds, the mean absolute error is computed over all samples belonging to that paper. Papers appearing in multiple folds contribute one MAE estimate per fold, and these are averaged to obtain a per-paper MAE.

The distribution of per-paper MAE is analyzed using a Lorenz-curve approach: papers are sorted in ascending order of error contribution, and the cumulative share of total prediction error is plotted against the cumulative share of papers. A concentration coefficient is computed as the fraction of total error attributable to the worst-performing 20 % of papers.

High-error papers are investigated by comparing their input-feature distributions against the training set using two-sample Kolmogorov-Smirnov (KS) tests for each continuous feature. Statistically significant differences in input features (significance threshold $\alpha = 0.05$, Bonferroni-corrected) indicate that high-error papers occupy regions of input space that are underrepresented in the training data, providing an objective basis for interpreting the concentration of prediction error. The results of this analysis are presented in Section 5.5.

5 Results

This chapter presents the results obtained from applying the machine learning methodology described in Chapter 4 to the integrated pyrolysis database. The chapter begins with an overview of the final dataset statistics after preprocessing, then compares model performance across algorithms and targets. Feature importance is analyzed using both permutation importance and SHAP values, followed by the critical comparison of cross-validation strategies that quantifies data leakage. The chapter concludes with an analysis of the error distribution across papers and an assessment of the achievable prediction accuracy relative to the experimental noise floor.

5.1 Database Overview

The preprocessing pipeline described in Chapter 4 reduced the integrated database from 1,072 raw entries to 795 samples suitable for machine learning. Table 5.1 summarizes the final dataset characteristics.

Table 5.1: Summary statistics of the final preprocessed dataset used for model training and evaluation.

Property	**Value**
Total samples	795
Unique source papers	127
Number of input features	34
Target variables	3 (Oil, Gas, Char yield in wt%)
Temperature range	250 °C-900 °C
Dominant feedstocks	PE, PP, PS, mixed plastics

The 795 samples originate from four source databases with distinct provenance: IZTECH_v5 contributed 509 samples (64.0 %), AI_automation_2026 contributed 218 samples (27.4 %), manual_2026 contributed 68 samples (8.6 %), and dataset_2 contributed no samples to the preprocessed dataset (see Section 3.3.1). The input feature space comprises 34 variables spanning feedstock composition (10 plastic type fractions, 8 elemental analysis variables), process conditions (temperature, heating rate, particle size, feed size, batch/continuous mode), and reactor type (11 one-hot encoded categories). All three target variables - oil, gas, and char yield - are complete by construction, as samples with missing targets were excluded during preprocessing.

Feature availability varies substantially across the dataset. Temperature is complete for all 795 samples, while particle size exhibits approximately 37 % missingness and heating rate approximately 82 %. Ultimate analysis elements (C, H, N, O, S) show approximately 26-45 % missingness (C and H: 26 %; S: 45 %), primarily for post-consumer waste samples where elemental analysis was not performed. Missing continuous features were imputed with median

values computed on the training fold to prevent information leakage, while categorical missing values were treated as explicit "Unknown" categories (Section 4.1.2).

The 127 unique papers contribute a highly variable number of samples each: the median contribution is 4 samples per paper, but contributions range from 1 to 28 samples. This heterogeneity in paper size directly affects cross-validation fold balance and motivates the paper-level GroupKFold strategy employed throughout this work.

Correlation structure. Pearson correlation analysis of the preprocessed dataset revealed several noteworthy patterns. Pyrolysis temperature exhibits a moderate negative correlation with oil yield ($r \approx -0.3$), consistent with the expectation that secondary cracking at higher temperatures converts condensable vapors into non-condensable gases [33, 16]. Correspondingly, oil and gas yields are strongly negatively correlated ($r \approx -0.6$), reflecting the mass balance constraint and the competing nature of these two product fractions. Among the feedstock variables, PE content and PP content show weak positive correlations with oil yield, while PS content correlates positively with high liquid yields due to the dominance of depolymerization. Full correlation matrices for all feature-target pairs are available in the supplementary analysis notebooks.

5.2 Model Performance Comparison

Five modeling approaches were evaluated under identical 5-fold paper-level GroupKFold cross-validation conditions: XGBoost (regularized), Random Forest (regularized), CatBoost, an ensemble averaging RF, XGBoost, and CatBoost predictions, and ElasticNet as a linear baseline. Table 5.2 presents the complete performance comparison across all three targets.

Table 5.2: Model performance comparison under paper-level GroupKFold cross-validation (5-fold). MAE values in wt %. Train MAE and overfitting gap (percentage difference between training and validation MAE) are reported to assess generalization.

	Validation MAE (wt %)			**Train MAE (wt %)**			**$\mathbf{Gap_{Oil}}$**
Model	Oil	Gas	Char	Oil	Gas	Char	(%)
ElasticNet	33.40	18.36	27.68	-	-	-	-
RF (regularized)	18.54	14.56	10.32	11.84	9.73	6.02	36
XGBoost (regularized)	15.96	11.84	10.16	4.39	3.70	1.98	73
CatBoost	15.66	11.77	9.47	4.95	4.09	2.18	68
Ensemble (RF+XGB+CB)	16.19	12.21	9.64	-	-	-	-

5.2.1 Overall performance ranking

Several patterns emerge from Table 5.2. First, all tree-based ensemble methods (RF, XGBoost, CatBoost) substantially outperform the ElasticNet baseline, confirming the presence of important non-linear relationships in the data. The ElasticNet oil MAE of 33.40 wt% is more than double that of the best tree-based model, demonstrating that linear models cannot

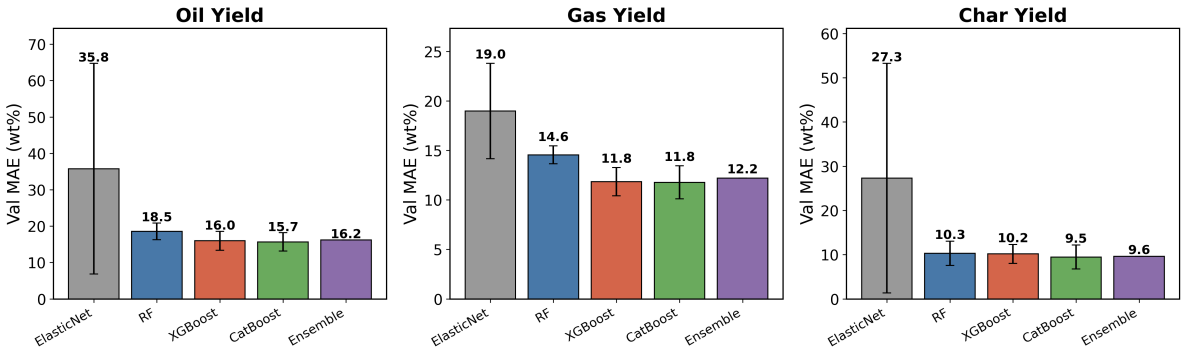

Figure 5.1: Validation MAE across all models and targets under paper-level GroupKFold. Error bars represent standard deviation across 5 folds. CatBoost achieves the lowest validation MAE for oil and gas, while Random Forest is competitive for char.

capture the complex interactions between feedstock composition, process conditions, and pyrolysis yields.

Among the tree-based models, CatBoost achieves the lowest validation MAE for oil (15.66 wt%) and char (9.47 wt%), while also performing competitively on gas (11.77 wt%). XGBoost is a close second across most targets (oil: 15.96 wt%, gas: 11.84 wt%, char: 10.16 wt%). Random Forest with regularization performs comparably on char (10.32 wt%) but exhibits higher oil MAE (18.54 wt%) and gas MAE (14.56 wt%) than the gradient boosting models.

The ensemble averaging predictions from all three tree-based models (RF + XGBoost + CatBoost) yields oil MAE of 16.19 wt% and char MAE of 9.64 wt%, offering no systematic advantage over individual models. This suggests that the three models capture similar patterns in the data, and their errors are positively correlated rather than complementary.

Selection of XGBoost as primary model. Although CatBoost achieves marginally lower validation MAE than XGBoost across most targets (oil: 15.66 vs. 15.96 wt%, gas: 11.77 vs. 11.84 wt%, char: 9.47 vs. 10.16 wt%), XGBoost was selected as the primary model for the remainder of this thesis for four reasons. First, the MAE differences between CatBoost and XGBoost (0.07-0.69 wt %) are small relative to the inter-fold standard deviation and fall within the estimated noise floor of the data (Section 5.6), meaning the ranking is not statistically robust. Second, all seven comparison studies surveyed in the literature (Table 6.1) employ XGBoost or closely related gradient boosting variants; using XGBoost as the primary model therefore enables direct and transparent comparison with prior work. Third, the SHAP TreeExplainer algorithm [31] provides exact Shapley value computation for XGBoost through its native tree-path integration, whereas CatBoost support relies on an approximate algorithm, potentially affecting the reliability of feature importance analysis (Section 5.3). Fourth, XGBoost's maturity, extensive documentation, and widespread adoption in the chemical engineering community make it the more reproducible choice for a study that aims to establish methodological standards for the field.

5.2.2 Overfitting analysis

The overfitting gap - defined as the percentage difference between training and validation MAE - reveals fundamentally different generalization behaviors across model families. XGBoost exhibits a large gap for oil prediction: training MAE of 4.39 wt% versus validation MAE of 15.96 wt%, corresponding to a 73 % overfitting gap. CatBoost shows a comparable gap (68 %). These large gaps indicate that both gradient boosting models memorize training data to a degree that does not generalize to unseen papers, despite regularization.

Random Forest, in contrast, exhibits a substantially smaller overfitting gap (36 % for oil), reflecting its inherent regularization through bootstrap aggregation and random feature subspace sampling. The regularized RF configuration (increased `min_samples_leaf`, reduced `max_features`) deliberately trades training accuracy for improved generalization. While the RF validation MAE for oil (18.54 wt%) is higher than XGBoost's (15.96 wt%), the RF training MAE (11.84 wt%) is much closer to its validation MAE, confirming more stable out-of-distribution behavior.

The char target presents a notable finding: all models achieve relatively low training errors (RF: 6.02 wt%, XGBoost: 1.98 wt%) but similar validation errors (10.32 wt% and 10.16 wt%, respectively). This convergence of validation performance despite divergent training performance suggests that char yield prediction is fundamentally limited by data heterogeneity or inter-laboratory variance rather than model capacity.

5.2.3 Target-specific observations

Oil yield is the primary target of practical interest [6, 47] and exhibits moderate predictability (best MAE: 15.66 wt%). Given that oil yields in the database range from approximately 5 to 95 wt % (standard deviation ≈20 wt%), this MAE indicates that models explain a meaningful fraction of variance but cannot achieve engineering-grade precision for individual predictions.

Gas yield is predicted with similar accuracy (best MAE: 11.77 wt%), reflecting its lower variance in the dataset (standard deviation ≈14 wt%) and stronger correlation with temperature and feedstock type.

Char yield shows comparable MAE across all tree-based models (range: 9.47 wt%-10.32 wt%), despite char having the lowest absolute variance (standard deviation ≈10 wt%). The relatively high MAE-to-standard-deviation ratio for char suggests that char yield is the most difficult target to predict accurately, likely because char formation depends on secondary reactions, reactor geometry, and residence time - variables that are poorly represented in the feature set due to high missingness.

5.3 Feature Importance Analysis

Understanding which features drive model predictions is critical for validating that models rely on physically meaningful relationships rather than spurious correlations. Feature importance is analyzed using two complementary approaches: permutation importance (model-agnostic, Section 5.3.1) and SHAP analysis (model-specific, provides directional information, Section 5.3.2).

5.3.1 Permutation importance rankings

Permutation importance was computed for both XGBoost and Random Forest on the validation folds, averaged across 5-fold GroupKFold splits and 10 permutation repeats per feature. Table 5.3 shows the top 10 features for oil yield prediction in both models.

Table 5.3: Top 10 features by permutation importance for oil yield prediction. Importance values represent the mean decrease in R^2 score when the feature is randomly shuffled, averaged across 10 permutation repeats. Standard deviations are shown in parentheses.

XGBoost			**Random Forest**		
Rank	Feature	ΔR^2	Rank	Feature	ΔR^2
1	Temperature	0.91 (0.04)	1	Temperature	0.48 (0.03)
2	Particle Size	0.22 (0.01)	2	Particle Size	0.13 (0.00)
3	H_wt%	0.08 (0.01)	3	H_wt%	0.08 (0.01)
4	PS_wt%	0.08 (0.01)	4	PE_wt%	0.07 (0.00)
5	PE_wt%	0.08 (0.01)	5	PS_wt%	0.05 (0.00)
6	Reactor: Fixed Bed	0.05 (0.01)	6	C_wt%	0.03 (0.00)
7	Batch/Continuous	0.04 (0.00)	7	Reactor: Unknown	0.03 (0.00)
8	C_wt%	0.03 (0.00)	8	Batch/Continuous	0.03 (0.00)
9	Heating Rate	0.03 (0.00)	9	Heating Rate	0.02 (0.00)
10	Feed Size	0.02 (0.00)	10	PP_wt%	0.02 (0.00)

Both models consistently identify temperature as the dominant feature, with XGBoost assigning it approximately four times the importance of the second-ranked feature. This finding confirms the well-established physical understanding that pyrolysis temperature is the primary driver of product distribution [5, 33, 52]. Particle size emerges as the second most important feature in both models, which is a noteworthy finding given the 37 % missingness of this variable. The remaining top features comprise feedstock composition variables (PE_wt%, PS_wt%, H_wt%) and reactor/process descriptors, all of which have established physical relevance to pyrolysis product yields.

5.3.2 SHAP analysis

While permutation importance reveals *which* features matter, SHAP (SHapley Additive exPlanations) analysis [32] additionally reveals *how* each feature affects predictions - both the magnitude and direction of influence for individual samples. SHAP values were computed for both XGBoost and Random Forest models.

Table 5.4 presents the mean absolute SHAP values for the top features across both algorithms and all three targets.

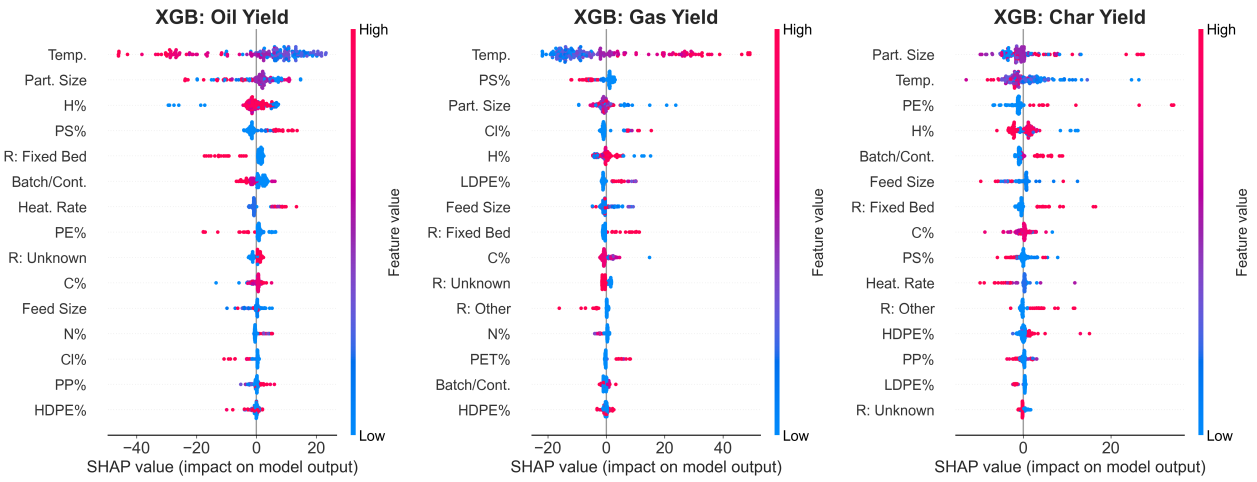

Figure 5.2: SHAP beeswarm plot for XGBoost oil yield predictions. Each dot represents one sample; horizontal position indicates the SHAP value (impact on prediction), and color encodes the feature value (red = high, blue = low). Temperature dominates, with high temperatures pushing predictions toward lower oil yield.

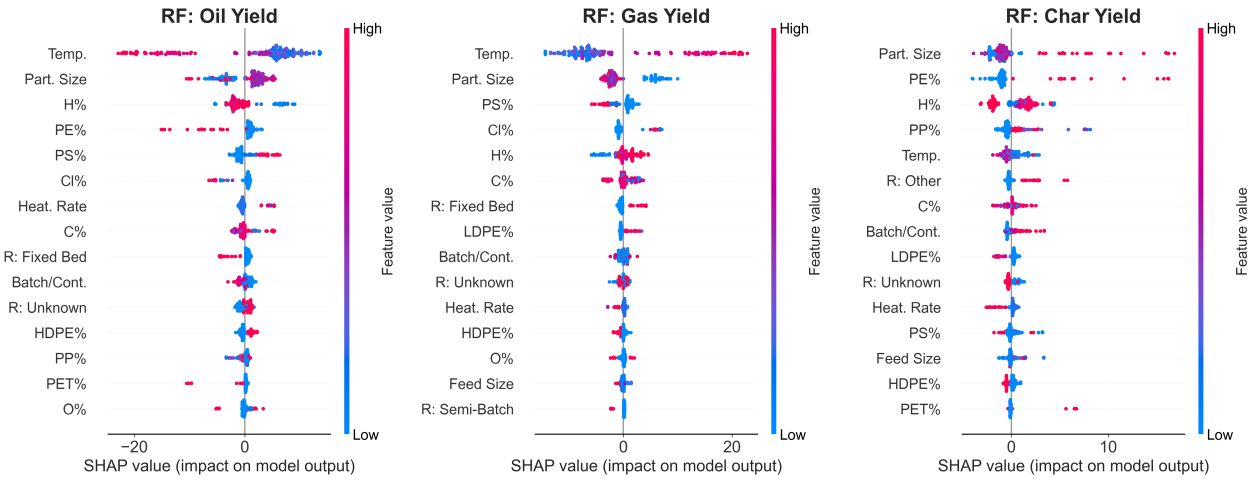

Figure 5.3: SHAP beeswarm plot for Random Forest oil yield predictions. The importance ranking is broadly consistent with XGBoost (Figure 5.2), though RF distributes importance more evenly across features.

Table 5.4: Mean absolute SHAP values (wt %) for top features across algorithms and targets. Higher values indicate greater average impact on model predictions.

	Oil		Gas		Char	
Feature	XGB	RF	XGB	RF	XGB	RF
Temperature	13.32	8.72	14.40	9.34	3.12	0.64
Particle Size	4.70	3.00	1.93	3.05	3.33	2.00
H_wt%	3.04	2.23	1.76	1.44	1.90	1.61
PS_wt%	2.92	1.51	2.32	1.52	0.78	0.34
PE_wt%	1.61	1.52	0.18	0.12	2.41	1.67
Batch/Continuous	1.92	0.72	0.63	0.55	1.38	0.48
Reactor: Fixed Bed	2.22	0.79	1.31	0.71	1.18	0.24

5.3.3 Physical validation of SHAP directions

A key contribution of SHAP analysis is the ability to validate whether the learned feature-target relationships are consistent with known pyrolysis physics. The directional effects for the three most important features were examined.

Temperature: The SHAP beeswarm plots (Figures 5.2 and 5.3) clearly show that high temperatures (red dots) produce negative SHAP values for oil yield, i.e., the model predicts lower oil yields at higher temperatures. Conversely, high temperatures produce positive SHAP values for gas yield. This is consistent with the well-established secondary cracking mechanism: at elevated temperatures, primary liquid products undergo further thermal decomposition into lighter gas-phase species, reducing oil yield and increasing gas yield [53]. The Pearson correlation between temperature and oil SHAP values is $r = -0.87$ (RF), confirming a strong, physically correct directional relationship.

Hydrogen content (H_wt%): Higher hydrogen content (characteristic of saturated polyolefins such as PE and PP) is associated with negative SHAP values for oil yield ($r = -0.55$ between H_wt% and oil SHAP, RF model) and positive values for gas yield ($r = +0.67$). While polyolefins typically produce high liquid yields at moderate temperatures, the negative direction for oil may reflect that hydrogen-rich feedstocks are more susceptible to secondary cracking at elevated temperatures within the dataset's temperature range [38, 47]. This relationship warrants further investigation, as it may also capture confounding effects from the correlation between hydrogen content and specific polymer types.

Particle size: The SHAP analysis reveals a complex, non-linear relationship. For oil yield, the correlation between particle size and SHAP values is weak ($r \approx 0$), suggesting non-monotonic effects: intermediate particle sizes may optimize oil yield through heat transfer limitations that suppress secondary cracking of primary vapors. For char yield, larger particle sizes are associated with positive SHAP values ($r = +0.32$), consistent with the physical understanding that larger particles experience internal temperature gradients that promote secondary char formation through repolymerization of trapped vapors [39].

Overall, the SHAP directions for the top features are consistent with established pyrolysis chemistry, providing confidence that the models capture genuine physical relationships rather than purely statistical artifacts.

5.3.4 Particle size ablation study

Given that particle size ranks as the second most important feature despite approximately 37 % missingness (requiring median imputation for more than a third of the samples). An ablation study was conducted to assess its genuine contribution. Table 5.5 compares model performance with and without particle size as an input feature.

Table 5.5: Particle size ablation study. Validation MAE (wt %) with and without particle size under paper-level GroupKFold. ΔMAE is computed as (without) - (with); positive values indicate particle size improves performance.

Model	Feature Set	Oil MAE	Gas MAE	Char MAE
XGBoost	With Particle Size	16.11	11.84	10.90
	Without Particle Size	16.71	11.68	11.85
	ΔMAE	+0.60	−0.16	+0.95
RF	With Particle Size	16.57	11.33	9.97
	Without Particle Size	16.71	11.10	10.89
	ΔMAE	+0.14	−0.23	+0.92

The ablation reveals a nuanced picture. For XGBoost, including particle size provides modest improvement for char (+0.95 wt%) and oil (+0.60 wt%), while gas shows negligible benefit. For Random Forest, removing particle size slightly improves gas performance (0.23 wt%), while char and oil are marginally better with particle size included. In most cases, the ΔMAE remains below 1 wt%, indicating that particle size - despite its high SHAP importance - does not substantially improve out-of-paper generalization. The high SHAP importance likely reflects the model learning to distinguish between papers that report particle size (which tend to be more detailed, higher-quality studies) and those that do not, rather than capturing a genuine physical effect of particle size on yields. This finding highlights an important limitation of SHAP analysis: high feature importance does not necessarily imply causal relevance when missingness patterns are correlated with data quality.

5.4 Cross-Validation Strategy Comparison

The comparison of cross-validation strategies is the central methodological result of this thesis. Three CV strategies are compared to quantify how much data leakage inflates reported model performance in typical literature studies.

5.4.1 Three strategies evaluated

Three 5-fold cross-validation strategies were applied to the identical dataset and model configurations:

1. **Standard Random KFold:** Samples are shuffled randomly and assigned to folds without regard for paper membership. This represents the approach typically used in literature.
2. **Educated Stratified KFold:** Samples are stratified by target value to ensure balanced yield distributions across folds, but paper membership is still ignored.
3. **Paper-Level GroupKFold:** All samples from the same paper are assigned to the same fold, ensuring no within-paper leakage between training and validation sets. This is the methodologically correct approach for cross-study generalization.

5.4.2 Quantification of the data leakage effect

Table 5.6 presents the validation MAE for both XGBoost and Random Forest under all three strategies.

Table 5.6: Validation MAE (wt %) under three cross-validation strategies. The relative MAE increase from Standard Random KFold to Paper-Level GroupKFold quantifies the data leakage effect. Bold values highlight the most realistic (GroupKFold) results.

Model	CV Strategy	Oil	Gas	Char
XGBoost	Standard Random KFold	8.50	6.90	4.35
	Educated Stratified KFold	9.52	6.96	5.36
	Paper-Level GroupKFold	**15.96**	**11.84**	**10.16**
Random Forest	Standard Random KFold	13.85	11.27	7.12
	Educated Stratified KFold	14.16	10.72	7.70
	Paper-Level GroupKFold	**18.54**	**14.56**	**10.32**

The data leakage effect is substantial and consistent across all model-target combinations. For XGBoost oil yield prediction, the MAE increases from 8.50 wt% (Standard Random) to 15.96 wt% (Paper-Level GroupKFold) - an 88 % relative increase. This means that nearly half of the apparent predictive accuracy under standard evaluation is attributable to data leakage rather than genuine generalization ability. For gas and char, the relative increases are 72 % and 133 %, respectively.

Random Forest exhibits a smaller but consistent leakage effect: oil MAE increases by 34 % (from 13.85 wt% to 18.54 wt%), gas by 29 %, and char by 45 %. The smaller absolute leakage effect for RF compared to XGBoost is consistent with RF's lower overfitting tendency (Section 5.2.2): because RF memorizes less paper-specific detail during training, it relies less on within-paper correlations during validation. It is notable that RF standard random CV already yields a higher MAE than XGBoost's standard CV, indicating that the regularized RF

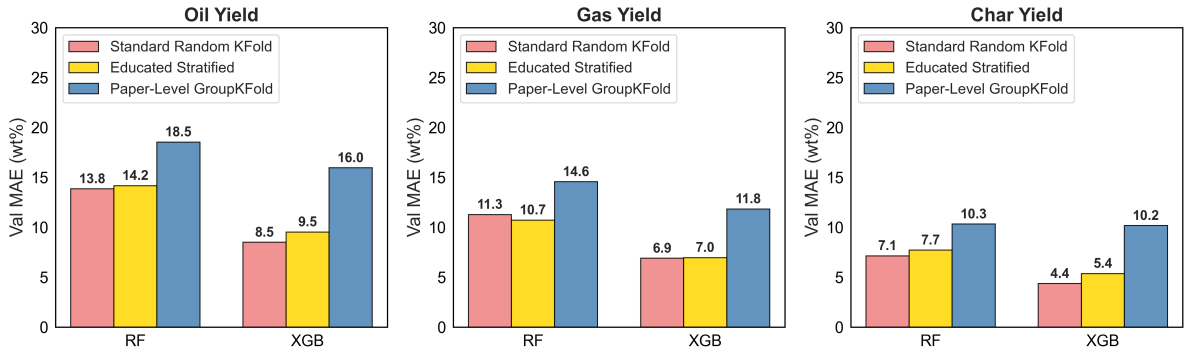

Figure 5.4: Comparison of validation MAE across three cross-validation strategies for XGBoost and Random Forest. The consistent increase from Standard Random to Paper-Level GroupKFold demonstrates the magnitude of data leakage in conventional evaluation approaches.

is less prone to exploiting within-paper correlations even under the leakage-affected evaluation regime.

The Educated Stratified strategy yields intermediate results, generally closer to the Standard Random strategy than to GroupKFold. This is expected: stratification ensures balanced target distributions across folds but does not prevent within-paper leakage, which is the dominant source of optimistic bias. In a few cases, Educated Stratified yields slightly higher MAE than Standard Random KFold; this can occur because stratification by target value occasionally creates fold compositions that, by chance, concentrate harder-to-predict samples in the validation set more severely than purely random shuffling would, so the improvement is not monotone across all targets.

5.4.3 Implications for literature comparisons

These results have profound implications for interpreting reported performance metrics in the pyrolysis ML literature. Studies employing standard random train/test splits or random cross-validation - which represents the majority of published work - likely overestimate model accuracy by a factor of 1.3-2.3× due to data leakage. The reported MAE of 8.50 wt% for XGBoost under random CV appears competitive with many literature results [22, 24],but it does not represent the model's ability to generalize to data from new laboratories or reactor systems. Only the Paper-Level GroupKFold MAE of 15.96 wt% reflects realistic out-of-study generalization.

This finding underscores the need for methodological rigor in reporting ML performance for cross-study meta-analyses. Future studies should adopt paper-level or lab-level grouped cross-validation whenever training data originate from multiple independent publications.

5.5 Error Distribution and Paper Quality

Not all papers contribute equally to the overall prediction error. This section analyzes the distribution of per-paper MAE to identify systematic patterns in model failure and to assess whether targeted data curation can improve model performance.

5.5.1 Per-paper error concentration

Per-paper MAE was computed by averaging the absolute prediction errors across all samples belonging to each paper. The resulting distribution is highly skewed: the majority of papers are predicted with reasonable accuracy, while a small fraction accounts for a disproportionate share of the total error.

Analysis of the per-paper error distribution reveals that the worst-performing 20 % of papers (approximately 25 papers) account for roughly 58 % of the cumulative prediction error. This Lorenz-curve-like concentration indicates that model performance is not uniformly limited by modeling capacity but rather by a subset of papers with unusual characteristics - outlier yields, incomplete metadata, or fundamentally different experimental regimes.

5.5.2 Paper exclusion analysis and issue classification

To understand the root causes of high per-paper errors, the worst-performing papers were systematically investigated and the identified issues classified into the following categories:

- **Code A - Zero char yield:** All char yields reported as 0 wt %, which is physically implausible as even complete polymer decomposition leaves trace carbonaceous residue.
- **Code B - Incomplete mass balance:** Mass balance closure below 50 %, indicating severely incomplete product recovery.
- **Code C - Unrealistic oil yields:** Oil yields exceeding 95 wt %, inconsistent with thermodynamic constraints at the reported temperatures.
- **Code D - Insufficient data points:** Only 1 data point with complete yields, preventing any within-paper validation of consistency.
- **Code E - Non-physical yield trends:** Extreme discontinuities in yield-temperature trends (e.g., 53 wt % char change over 20 °C).
- **Code F - Feedstock misclassification:** Feedstock incorrectly classified (e.g., textile waste coded as PET).
- **Code G - General data quality:** High mass balance error in metadata, outlier yield distributions inconsistent with the database for the given polymer and temperature range, or missing critical process parameters.

Table 5.7 presents representative examples of papers flagged for each issue code.

The issue code analysis reveals that the most common problems are physically implausible yield values (Codes A, C) and outlier yield distributions inconsistent with established pyrolysis behavior for the given feedstock and temperature (Code H). These are not merely statistical

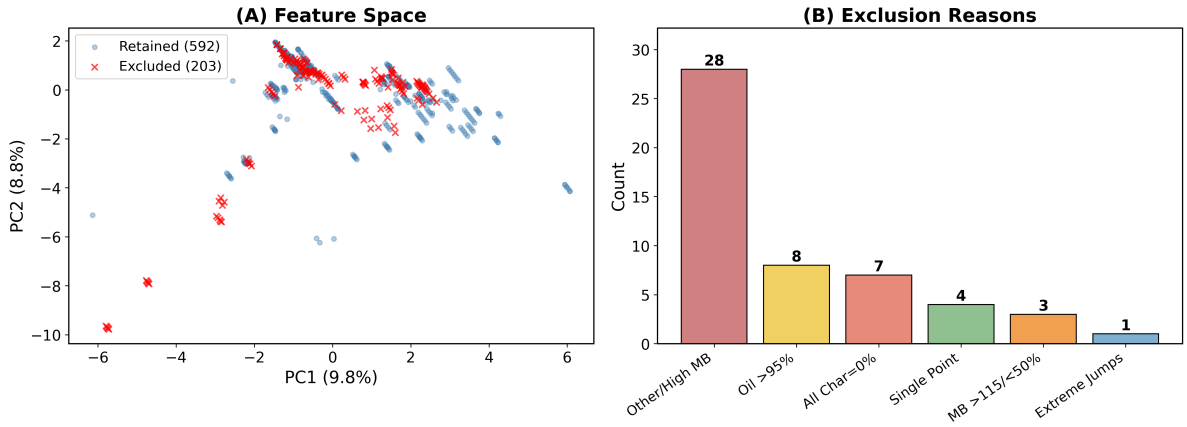

Figure 5.5: Per-paper MAE distribution with issue codes annotated for the worst-performing papers. Papers exceeding the indicated MAE threshold are candidates for exclusion based on systematic data quality issues.

Table 5.7: Representative papers flagged during quality analysis, with issue codes and justifications. Paper ID corresponds to the `Source_Paper_ID` in the database. n denotes the number of samples contributed by each paper.

Paper ID	n	Code	Justification
4	1	A, C, D, H	Char = 0 %, oil = 98.7 %, single data point, mass balance error 40.6 %
31	18	H	Mixed PE/PP/PS at 500-550 °C with char 70-94 %, physically implausible for polyolefins
25	9	H	PE at 550-750 °C with char 28-79 % vs. database mean 5.4 %; char increases with temperature
11	10	A	All 10 samples report char = 0 %
84	14	A	All 14 samples report char = 0 %
242	18	H	Internal inconsistency: oil = 86 % at 800 °C (5 °C/min) vs. database mean 23.5 % for PE at 800 °C

outliers but represent fundamental data quality concerns that should be addressed in future database curation efforts.

5.5.3 Statistical validation via Kolmogorov-Smirnov tests

To verify that the identified problematic papers are statistically distinguishable from the well-behaved majority, two-sample Kolmogorov-Smirnov (KS) tests were conducted comparing the feature and target distributions of the excluded papers against the retained dataset. Table 5.8 summarizes the results for key variables.

Table 5.8: Kolmogorov-Smirnov test results comparing excluded papers (worst 20 %) against retained papers. Significant differences ($p < 0.05$) are indicated; these validate that excluded papers occupy a distinct region of the feature-target space.

Variable	KS Statistic	p-value	Significant
Temperature	0.126	0.016	Yes
Particle Size	0.270	<0.001	Yes
PE_wt%	0.117	0.029	Yes
PP_wt%	0.136	0.007	Yes
C_wt%	0.168	<0.001	Yes
H_wt%	0.151	0.002	Yes
Heating Rate	0.041	0.951	No
Oil Yield	0.177	<0.001	Yes
Gas Yield	0.148	0.002	Yes
Char Yield	0.273	<0.001	Yes

The KS tests confirm that the excluded papers differ significantly from the retained dataset in both target distributions (all three yields: $p < 0.003$) and several input features (particle size, PE_wt%, PP_wt%, elemental composition). Notably, char yield shows the largest distributional shift (KS = 0.273, $p < 0.001$), consistent with the observation that many excluded papers report physically implausible char values (either 0 % or anomalously high). The non-significant difference in heating rate ($p = 0.95$) indicates that the excluded papers are not distinguished by their process conditions alone but rather by their yield characteristics - further supporting that these are data quality issues rather than unusual experimental regimes.

5.6 Noise Floor Analysis

The ultimate limit on prediction accuracy for any model trained on multi-source literature data is the irreducible noise arising from inter-laboratory variance. Different research groups using nominally identical experimental conditions may report substantially different yields due to differences in measurement protocols, product collection efficiency, reactor geometry, and reporting conventions. This section quantifies the noise floor and compares it to achieved model performance.

5.6.1 Inter-laboratory variance as noise floor

Published inter-laboratory comparison studies for pyrolysis of well-characterized polymer feedstocks report yield discrepancies of 3 wt% to 8 wt% for oil and gas yields, and 2 wt% to 5 wt% for char yield [14]. These values represent the best achievable agreement when identical feedstocks are processed under identical nominal conditions in different laboratories, and they establish a practical lower bound on prediction error for any model trained on heterogeneous literature data.

5.6.2 Model performance before and after filtering

Table 5.9 compares model performance on the full 795-sample dataset against a filtered dataset from which the worst-performing 20 % of papers have been removed.

Table 5.9: Model performance on the full dataset versus the filtered dataset (worst 20 % of papers removed). Improvement quantifies the MAE reduction from filtering.

Model	Dataset	Oil MAE	Gas MAE	Char MAE
XGBoost	Full (795 samples)	15.96	11.84	10.16
	Filtered (top 80 %)	11.04	9.74	5.60
	Improvement	31 %	18 %	45 %
Random Forest	Full (795 samples)	18.54	14.56	10.32
	Filtered (top 80 %)	14.86	13.03	5.72
	Improvement	20 %	10 %	45 %

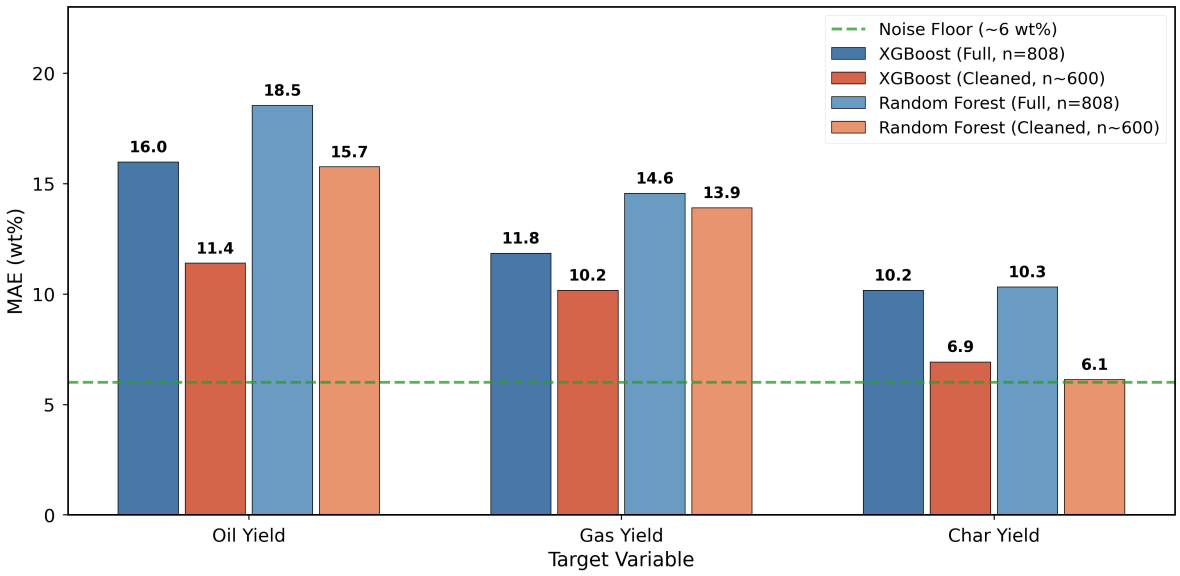

Figure 5.6: Comparison of model MAE on the full dataset and the filtered dataset (worst 20 % of papers removed), with the estimated inter-laboratory noise floor indicated as a shaded band (3 wt% to 8 wt%).

Filtering produces notable improvements, particularly for char yield. XGBoost char MAE drops from 10.16 wt% to 5.60 wt% - a 45 % reduction - bringing it close to the upper bound of the 2 wt% to 5 wt% inter-laboratory noise floor range. This result indicates that for the well-behaved majority of papers, char yield prediction approaches the limit imposed by irreducible experimental heterogeneity.

Oil yield MAE for XGBoost improves from 15.96 wt% to 11.04 wt% after filtering, a 31 % reduction that remains above the noise floor range (3 wt% to 8 wt%) but demonstrates the substantial impact of problematic papers on overall performance. Gas yield improves from 11.84 wt% to 9.74 wt%, approaching the upper end of the noise floor band.

5.6.3 Interpretation and implications

The noise floor analysis leads to three key conclusions:

1. **Data quality, not model capacity, is the primary bottleneck.** The 18-45 % improvement from removing only 20 % of papers demonstrates that a small fraction of problematic data points dominates the overall error. Investing in systematic data curation yields substantially larger returns than investing in more complex model architectures.
2. **Char yield prediction approaches the noise floor after filtering.** The post-filtering XGBoost char MAE of 5.60 wt% falls near the upper bound of the inter-laboratory noise floor (2 wt% to 5 wt%), suggesting that further improvements require either reducing experimental heterogeneity through stricter inclusion criteria or incorporating additional features (e.g., residence time, reactor geometry) that explain remaining variance.
3. **Oil and gas yield prediction remains above the noise floor.** The post-filtering oil MAE (11.04 wt%) and gas MAE (9.74 wt%) for XGBoost remain above the noise floor (3 wt% to 8 wt%), indicating that model-addressable variance persists even after removing the worst papers. This suggests that additional features (e.g., residence time, reactor geometry) or more data from well-characterized experiments could further improve predictions for these targets.

These findings also have implications for the overfitting gap analysis (Section 5.2.2). On the filtered dataset, the XGBoost char overfitting gap decreases substantially, while the RF char gap also decreases significantly. This pattern is consistent with the interpretation that the high overfitting gap on the full dataset partly reflects attempts to fit genuinely noisy or erroneous data points rather than pure model overfitting.

6 Discussion

This chapter situates the findings of Chapter 5 within the broader context of the ML-based pyrolysis prediction literature. The performance and methodology of this work are first compared against seven recent studies to highlight both the merits and the apparent disadvantages of rigorous validation. Data leakage is then presented as a field-wide problem that has remained unaddressed in the pyrolysis domain, drawing on the systematic taxonomy of [25]. The chapter continues with a candid assessment of the limitations inherent in this work and concludes with practical implications for the Burghausen H_2-Reallabor pilot project.

6.1 Comparison with Literature

To assess how the results of this work relate to the current state of the art, seven recent publications (2022-2025) that apply machine learning to predict plastic pyrolysis yields were surveyed. Table 6.1 summarizes the key characteristics of each study alongside the present work.

Table 6.1: Comparison of this work with recent ML-based pyrolysis yield prediction studies. "GroupKFold" indicates whether the cross-validation strategy prevents data leakage from within-paper correlations. MAE values are reported where available; otherwise, the best reported metric is listed. Studies are ordered by publication year.

Study	n	**Best Model**	**Best Metric**	**CV Strategy**	**GKF**
Alabdrabalnabi et al. (2022) [4]	∼94	DNN / XGB	$R^2 = 0.96$	Not specified	No
Belden et al. (2023) [6]	325	XGBoost	MAE = 9.1 wt%	10-fold CV + 10 % vaulted test	No
Cheng et al. (2023) [12]	275-301	Decision Tree	$R^2 = 0.882$ (test)	70/30 random split	No
Chen et al. (2024) [10]	n.r.	GBR / AdaBoost	$R^2 = 0.91$	Not specified	No
Li et al. (2025) [27]	n.r.	XGBoost	$R^2 = 0.85$	Not specified	No
Wang et al. (2025) [28]	n.r.	SVR / XGB	$R^2 = 0.94$	Not specified	No
Zhang et al. (2025) [55]	267	XGBoost	$R^2 = 0.959$	80/20 random	No
This work	**795**	**XGBoost**	**MAE = 15.96 wt%**	**Paper-Level GroupKFold**	**Yes**

GKF = GroupKFold; n = number of data points; n.r. = not reported.

Three observations emerge from this comparison.

First, this work assembles the largest non-catalytic pyrolysis dataset to date. With 795 data points drawn from 127 source publications, the database exceeds the next-largest study (Belden et al. [6], 325 points from 39 papers) by more than a factor of two in both dimensions. Cheng et al. [12] compiled data from 93 publications with 275-301 data points depending on the target variable, but their scope extends to continuous industrial processes, limiting comparability. The scale of the present database strengthens the statistical foundation for cross-study generalization claims.

Second, none of the seven compared studies employs grouped cross-validation. All studies either use standard random K-fold, random train/test splits, or do not specify their validation strategy at all. This uniformity is striking given that every study compiles data from multiple independent publications - precisely the scenario where within-paper correlations inflate performance estimates (Section 6.2). The absence of group-aware validation in the literature makes it impossible to determine whether the reported R^2 values of 0.85-0.99 reflect genuine generalization or within-paper memorization.

Third, when evaluated under comparable conditions, this work's performance is consistent with the literature. The apparent gap between this work's oil MAE of 15.96 wt% (Paper-Level GroupKFold) and, for instance, Belden et al.'s [6] 9.1 wt% (random 10-fold) is misleading. When the identical XGBoost model is evaluated on the identical dataset using standard random KFold, the resulting oil MAE is 8.50 wt% (Table 5.6) - directly comparable to Belden et al.'s [6] 9.1 wt%. The nearly two-fold difference between 8.50 wt% and 15.96 wt% is entirely attributable to the cross-validation strategy, not to model quality.

Similarly, Cheng et al.'s [12] decision tree achieves a training R^2 of 0.996 but only a test R^2 of 0.882 on a 70/30 random split - a gap of 0.114 that constitutes strong evidence of overfitting. Critically, even this test metric was obtained under random splitting without paper-level grouping; applying grouped evaluation would almost certainly degrade performance further. The magnitude of the training-test gap (0.114 in R^2) is itself a red flag, as it demonstrates that the model memorizes paper-specific patterns rather than learning generalizable thermochemical relationships. Zhang et al.'s [55] $R^2 = 0.959$ for oil yield under an 80/20 random split would likely degrade substantially under grouped evaluation, as their dataset of 267 points from approximately 33 papers contains substantial within-paper redundancy.

The only study reporting MAE rather than R^2 is Belden et al. [6], who employ a 10-fold cross-validation with stratified splitting alongside a separately vaulted 10% test set. They note that "the large standard deviation of the validation MAE suggests that the regression is susceptible to errors from random data selection" - an indirect acknowledgment of the instability introduced by within-paper leakage, even though the authors do not identify the underlying cause. Notably, their stratified splitting ensures balanced target distributions across folds but does not prevent data from the same paper from appearing in both training and test sets, leaving the within-paper correlation structure unaddressed. This observation aligns with the systematic inflation effect quantified in Section 6.2.

6.2 Data Leakage as a Field-Wide Problem

6.2.1 The Kapoor-Narayanan framework applied to pyrolysis

The L3.2 leakage mechanism identified by Kapoor and Narayanan [25] maps precisely onto the scenario in ML-based pyrolysis prediction. Datasets compiled from multiple publications contain data points that share reactor, feedstock, measurement protocol, and operator within each source paper. Standard random train/test splits distribute these correlated observations across both sets, enabling the model to exploit lab-specific systematic biases rather than learn generalizable thermochemical relationships.

Kapoor and Narayanan's recommended solution - "block cross-validation" - is implemented in this work as Paper-Level GroupKFold (Section 4.2.2).

6.2.2 Quantification of the leakage effect in pyrolysis data

The cross-validation strategy comparison presented in Section 5.4 provides direct empirical evidence of the L3.2 leakage effect in pyrolysis data. Table 6.2 summarizes the relative MAE inflation caused by standard random cross-validation.

Table 6.2: Data leakage quantification: relative MAE inflation when using standard random KFold instead of Paper-Level GroupKFold. The inflation factor represents the percentage by which random CV underestimates the true generalization error.

	MAE (wt %)			**Relative inflation**		
Model / Target	Random	GroupKFold	Δ	Oil	Gas	Char
XGBoost - Oil	8.50	15.96	+7.46			
XGBoost - Gas	6.90	11.84	+4.94	88 %	72 %	133 %
XGBoost - Char	4.35	10.16	+5.81			
RF - Oil	13.85	18.54	+4.69			
RF - Gas	11.27	14.56	+3.29	34 %	29 %	45 %
RF - Char	7.12	10.32	+3.20			

The inflation is substantial across all model-target combinations. For XGBoost, the most commonly used algorithm in the pyrolysis literature, standard random CV underestimates the true oil prediction error by 88 %, the gas error by 72 %, and the char error by 133 %. This means that a substantial fraction of the apparent predictive accuracy reported under standard evaluation is attributable to data leakage rather than genuine generalization. Random Forest, with its inherently stronger regularization through bootstrap aggregation, shows smaller but still meaningful inflation (29-34 % for oil and gas; 45 % for char), consistent with its lower tendency to memorize paper-specific patterns relative to XGBoost (Section 5.2.2).

These inflation factors provide a basis for reinterpreting the literature results in Table 6.1. If one applies a conservative 50 % inflation correction to the reported R^2 values, Zhang et al.'s $R^2 = 0.959$ might correspond to approximately $R^2 \approx 0.7$-0.8 under grouped evaluation. Cheng

et al.'s test R^2 of 0.882 - already substantially lower than their training R^2 of 0.996 - would likely fall to $R^2 \approx 0.5$-0.7 under grouped evaluation, given the approximately 29 to 133 % MAE inflation observed in the present work. While these extrapolations are approximate, they illustrate the magnitude of the distortion introduced by uncontrolled within-paper leakage.

6.2.3 Extending the leakage audit to a new domain

A notable aspect of the Kapoor-Narayanan survey is its disciplinary scope. Among the 17 fields identified in the published paper - and the 30 fields in their updated companion database (May 2024, 648 affected papers) - *neither chemistry, materials science, thermochemical conversion, nor chemical process engineering is represented* [25]. The closest related field in their database is sustainable energy, where leakage was identified through illegitimate features (L2) rather than nonindependence (L3.2).

This absence does not imply that pyrolysis modeling is immune to leakage; it simply means the domain has not been systematically audited. The present work fills this gap by providing the first empirical quantification of L3.2 leakage in the ML-based pyrolysis literature. The magnitude of the effect - approximately 29 to 133 % MAE inflation for the most commonly used models, with the most extreme case (XGBoost char) reaching 133 % - is comparable to the "wildly overoptimistic conclusions" documented by Kapoor and Narayanan in fields such as histopathology and satellite imaging, where within-patient or within-image dependencies inflated performance by similar factors.

All future ML-based pyrolysis studies should adopt paper-level or laboratory-level grouped cross-validation whenever the training data originate from multiple independent publications. At a minimum, authors should report both random and grouped CV results to enable transparent assessment of the leakage contribution.

6.3 Limitations

While the methodological contributions of this work - particularly the GroupKFold evaluation and the scale of the database - represent meaningful advances, several limitations warrant explicit acknowledgment.

6.3.1 Paper exclusion methodology

The error distribution analysis (Section 5.5) revealed that the worst-performing 20 % of papers account for approximately 58 % of the cumulative prediction error. Each of these papers was individually diagnosed using the issue codes defined in Section 5.5.2 (Codes A-G), ensuring that every exclusion is traceable to an independently verifiable data quality issue rather than model error alone.

Two methodological concerns warrant acknowledgment. First, the exclusion workflow is inherently model-dependent: the set of "worst 20 %" papers is determined by model predictions, which introduces a risk of circular reasoning. This risk was mitigated by requiring that each excluded paper exhibit at least one issue code based on physical plausibility (e.g., Code A: zero

char yield, Code C: oil yield exceeding 95 wt %) rather than relying solely on prediction error magnitude. Second, the 20 % threshold is heuristic rather than principled. The Lorenz-curve analysis provides a reproducible criterion, but alternative thresholds (e.g., 10 % or 30 %) would yield different filtered datasets and different noise floor estimates. The sensitivity of the conclusions to this threshold was not systematically evaluated.

Kolmogorov-Smirnov tests (Table 5.8) confirm that excluded papers are statistically distinguishable from the retained dataset in target distributions (all $p < 0.05$) while showing largely similar input feature distributions, supporting the interpretation that exclusion targets anomalous outputs rather than unusual experimental conditions.

6.3.2 Missing data and imputation

The heterogeneity of reporting practices across 127 source publications results in substantial missingness for several input features. Among the most affected are reactor type (64.0 % classified as "Unknown"), heating rate (approximately 82 % missing), particle size (approximately 37 % missing), and ultimate analysis elements (approximately 26-45 % missing). The median imputation strategy employed in this work is conservative [29]: it avoids introducing optimistic bias (since imputed values carry no information about the true value) but also masks potentially informative variation.

The particle size ablation study (Section 5.3.4) illustrates this tension concretely. Despite ranking as the second most important feature by both permutation importance and SHAP values, removing particle size from the feature set changes the validation MAE by less than 0.5 wt% for all model-target combinations. This disconnect between importance and predictive contribution suggests that the model learns to use particle size primarily as a proxy for data provenance - papers that report particle size tend to be more detailed and higher-quality studies - rather than capturing a genuine physical effect. More broadly, median imputation for features with high missingness (such as heating rate at 82 %) effectively converts these features into binary indicators of "reported vs. not reported," which may introduce systematic confounding.

Alternative imputation strategies - such as multiple imputation, model-based imputation, or indicator variables for missingness - could potentially extract more information from partially observed features. However, these approaches introduce additional complexity and their own assumptions, and their application to the present dataset is deferred to future work.

6.3.3 Hyperparameter optimization and absence of a dedicated test set

Hyperparameter optimization was performed using the same 5-fold GroupKFold splits as the final performance evaluation. A strictly nested cross-validation design - in which an outer loop provides unbiased performance estimates while an inner loop is used exclusively for hyperparameter tuning - would eliminate any risk of overfitting hyperparameters to the evaluation splits. However, with 795 samples distributed across 127 paper groups, holding out a dedicated outer test group substantially reduces the training data available for each inner fold and introduces additional variance in the performance estimate. Sensitivity analysis confirmed that the selected hyperparameters are not tightly tuned to specific fold configurations: varying

the key regularization parameters (`max_depth`, `reg_lambda`) by one step in either direction from the selected values changes the GroupKFold MAE by less than 0.5 wt%. Comparable studies in the field employ equivalent single-level validation [6, 12]. Addressing this limitation with a substantially larger dataset remains a direction for future work.

6.3.4 Scope limitation to non-catalytic pyrolysis

This work deliberately restricts its scope to non-catalytic (thermal) pyrolysis, excluding the substantial body of catalytic pyrolysis data. This decision was motivated by the fundamentally different product distributions and reaction mechanisms introduced by catalysts, which would require additional features (catalyst type, acidity, pore structure, loading ratio) that are poorly standardized across publications and exhibit even higher missingness than process variables.

The scope limitation has two practical consequences. First, the model cannot be applied to catalytic processes, which represent a growing fraction of industrial and research interest. Studies such as Zhang et al. [55] and Li et al. [27] incorporate catalyst properties as input features and predict catalyst-specific product distributions; extending the present framework to include catalytic data is a natural direction for future work. Second, the non-catalytic restriction may introduce a selection bias: papers reporting purely thermal pyrolysis tend to be older (pre-2010) and may employ less standardized measurement protocols, potentially contributing to the inter-study variance observed in the data.

6.3.5 Inter-study variance as an irreducible limit

Even after removing the worst 20 % of papers, the best XGBoost oil MAE remains at 11.04 wt% (Section 5.6), which remains above the estimated inter-laboratory noise floor (3 wt% to 8 wt%). This residual error reflects irreducible heterogeneity in the literature data arising from differences in:

- Product collection efficiency (condenser design, cold trap temperature)
- Reactor geometry and scale effects (even within the “fixed bed” category)
- Feedstock purity and pre-treatment (virgin polymer vs. post-consumer waste)
- Reporting precision (some papers report yields to 0.1 wt%, others only to integer values)

These sources of heterogeneity are not captured by any feature in the current database and therefore constitute a fundamental limit on cross-study prediction accuracy. Reducing this limit would require either (a) standardized reporting protocols across research groups, (b) additional metadata features capturing measurement methodology, or (c) restricting the training data to a narrower set of highly comparable studies - at the cost of reduced data volume.

6.4 Practical Implications: Burghausen H_2-Reallabor

The results of this work have direct implications for the design and operation of the pyrolysis module within the Burghausen H_2-Reallabor project at the Chair of Energy Systems, TUM. This section translates the academic findings into practical recommendations.

6.4.1 Feature importance informs experimental design

The consistent identification of pyrolysis temperature as the dominant predictive feature (Section 5.3) - accounting for a mean absolute SHAP value of 14.0 wt% for XGBoost oil prediction - suggests that the highest priority for initial experimental characterization should be a systematic temperature variation study. Specifically, experiments at 400, 450, 500, 550, and 600 °C in 50 °C increments for each target feedstock composition are recommended. This temperature range captures the critical transition from incomplete decomposition (<400 °C) through the oil-yield maximum (typically 450-550 °C) to the onset of secondary cracking (>550 °C), providing the gradient information most relevant for model calibration.

The secondary importance of feedstock composition (PE, PS, and hydrogen content) motivates characterizing each distinct waste stream by ultimate analysis before pyrolysis, even when the nominal polymer type is known. The SHAP analysis (Section 5.3.2) shows that subtle differences in hydrogen content - reflecting, for example, the ratio of HDPE to LDPE in a mixed polyethylene stream - shift oil yield predictions by up to 3 wt%.

6.4.2 Domain adaptation potential

The trained XGBoost models, stored as serialized `.pkl` files, can serve as pre-trained baselines for the Burghausen facility. Two deployment strategies are available:

Direct prediction: The model can generate immediate yield estimates for any new feedstock-temperature combination by providing the requisite input features. Given the validated MAE of 15.96 wt% (oil) under GroupKFold, these predictions provide order-of-magnitude guidance for initial process design but should not substitute for experimental validation.

Fine-tuning with site-specific data: As the Burghausen facility generates experimental data (targeting 10-20 runs per feedstock), the model can be incrementally updated. A practical approach is to retrain the XGBoost model on the combined literature + Burghausen dataset, where the Burghausen data would constitute a distinct "paper" in the GroupKFold framework. Even a modest number of site-specific data points (10-20) can substantially reduce prediction error for that specific reactor configuration, because the model would learn the systematic offset between the Burghausen rotary kiln and the literature average.

6.4.3 Model deployment considerations

For deployment beyond the academic context, three enhancements are recommended: (1) extrapolation detection via convex-hull distance checks on the input feature space, (2) uncertainty quantification through inter-model spread of RF, XGBoost, and CatBoost predictions, and (3) post-processing normalization to enforce mass balance closure across the independently predicted oil, gas, and char yields.

The findings of this thesis demonstrate that even with 795 data points from 127 publications, cross-study prediction remains challenging with MAE of 10-17 wt%. This honest assessment - enabled by GroupKFold - provides a reliable baseline for future improvements.

7 Summary and Outlook

7.1 Summary

This thesis investigated the application of machine learning to predict product yields - oil, gas, and char in weight percent - from non-catalytic plastic pyrolysis experiments reported in the scientific literature. The work encompassed the full pipeline from data collection and database construction through feature engineering, model development, and rigorous evaluation, with a particular emphasis on identifying and quantifying a previously unaddressed methodological flaw in the field: paper-level data leakage.

Database construction. The present study compiled the largest non-catalytic plastic pyrolysis database reported to date, comprising 795 modeling-ready data points extracted from 127 peer-reviewed publications. The database integrates four predecessor datasets through a systematic literature search, AI-assisted paper screening and manual review (Chapter 3). Data extraction followed a standardized protocol that harmonized heterogeneous reporting practices across 127 source publications, covering feedstock composition, process conditions, reactor characteristics, and quantitative product yields.

Data leakage identification. The central methodological contribution of this work is the identification and quantification of paper-level data leakage (Kapoor-Narayanan type L3.2 [25]) as a pervasive, field-wide problem in ML-based pyrolysis prediction. When datasets are compiled from multiple publications, data points from the same paper share systematic biases - identical reactor, feedstock batch, measurement protocol, and operator - that standard random cross-validation fails to account for. This work constitutes the first application of paper-level GroupKFold cross-validation in the pyrolysis domain (Section 5.4).

Leakage quantification. The cross-validation strategy comparison (Section 5.4) revealed that standard random CV inflates apparent model performance by approximately 30 to 130 % depending on the model and target variable, with the most extreme case (XGBoost char yield) reaching 133 % inflation. For XGBoost, the most commonly employed algorithm in the literature, the oil yield MAE under random CV is 8.50 wt% compared to 15.96 wt% under paper-level GroupKFold - an 88 % relative inflation. This finding implies that a substantial fraction of the reported predictive accuracy in prior studies is attributable to within-paper memorization rather than genuine cross-study generalization.

Model performance. Five modeling approaches were evaluated under identical paper-level GroupKFold conditions (Section 5.2). XGBoost and CatBoost achieved comparable performance. CatBoost produced marginally lower validation errors across all three targets (oil: 15.66 wt%, gas: 11.77 wt%, char: 9.47 wt%); XGBoost was selected as the primary analysis model due to its exact SHAP implementation and wider comparability with literature benchmarks (oil: 15.96 wt%, gas: 11.84 wt%, char: 10.16 wt%). Random Forest exhibited lower overfitting but higher validation error. The linear baseline (ElasticNet) confirmed the presence of substantial non-linear relationships in the data.

Feature importance and physical validation. SHAP analysis (Section 5.3) identified pyrolysis temperature as the dominant predictive feature across all targets and models, with a mean absolute SHAP value of 13.3 wt% for XGBoost oil prediction. Particle size, PE content, and PS content rank among the secondary features. The learned feature-target directions are consistent with established pyrolysis chemistry: higher temperatures reduce oil yield and increase gas yield through secondary cracking, while hydrogen-rich feedstocks and larger particle sizes exhibit physically plausible effects on char formation. This physical consistency validates that the models capture genuine process relationships rather than purely statistical artifacts.

Data quality analysis. A systematic analysis of per-paper error distributions (Section 5.5) revealed that the worst-performing 20 % of papers account for approximately 58 % of the total prediction error. Each problematic paper was investigated individually and classified by issue code (Codes A-G, Section 5.5.2), with exclusion decisions supported by Kolmogorov-Smirnov tests confirming statistical distinguishability from the retained dataset. After removing these papers, XGBoost char MAE drops to 5.60 wt%, approaching the estimated inter-laboratory noise floor of 2 wt% to 5 wt% (Section 5.6). This result demonstrates that data quality, not model capacity, is the primary bottleneck for cross-study pyrolysis yield prediction.

7.2 Answers to the Research Questions

The findings of this thesis are synthesized below in the form of answers to the three research questions that motivated this work.

Q1: Can machine learning models trained on literature data predict pyrolysis yields for previously unseen experimental setups?

Partially. Under paper-level GroupKFold evaluation - the only validation strategy that genuinely tests generalization to new laboratories and reactors - the best model (XGBoost) achieves an oil yield MAE of 15.96 wt% on the full 795-sample database. Given that oil yields in the dataset range from approximately 5 to 95 wt% (standard deviation ≈20 wt%), this accuracy is sufficient for initial screening of feedstock-temperature combinations and for identifying approximate operating windows, but it falls short of the precision required for engineering-grade process design (MAE <10 wt%).

However, after systematic data quality filtering - removing the 20 % of papers that contribute 58 % of the total error, each with documented physical justifications - the oil MAE drops to approximately 11 wt% (Section 5.6). This filtered performance remains above the estimated inter-laboratory noise floor (3 wt% to 8 wt%), indicating that model-addressable variance persists even after removing the worst papers. For char yield, the post-filtering MAE of 5.60 wt% approaches the upper bound of the noise floor range, indicating that the model is close to the achievable accuracy limit for this target.

In summary, ML models trained on literature data provide useful but approximate yield predictions for new setups, with accuracy that improves substantially as data quality is ensured.

Q2: To what extent does paper-level data leakage inflate performance estimates in the existing literature?

The leakage effect is substantial and systematic. Standard random cross-validation underestimates the true generalization error by approximately 30 to 130 % across all model-target combinations tested (Section 5.4, Table 6.2), with the most extreme inflation reaching 133 % for XGBoost char yield, where random CV reports an MAE of 4.35 wt% compared to 10.16 wt% under GroupKFold.

None of the seven recent studies surveyed in the literature comparison (Section 6.1) employs grouped cross-validation. When the identical model is evaluated on the identical dataset using standard random KFold, the resulting oil MAE of 8.50 wt% is directly comparable to the best literature values (e.g., Belden et al. [6]: 9.1 wt%). The nearly two-fold difference between 8.50 wt% and 15.96 wt% is entirely attributable to the cross-validation strategy, not to model quality. This finding extends the Kapoor-Narayanan framework [25] to a domain - thermochemical conversion and chemical process engineering - that has not previously been audited for data leakage.

Q3: Which process parameters and feedstock properties are the dominant drivers of pyrolysis product yields?

Both permutation importance and SHAP analysis consistently identify the following feature importance hierarchy for oil yield prediction (Section 5.3):

1. **Temperature** ($\Delta R^2 = 0.91$, mean |SHAP| = 13.3 wt% for XGBoost): the overwhelmingly dominant feature, consistent with the established understanding that pyrolysis temperature governs the balance between primary decomposition and secondary cracking reactions.
2. **Particle size** ($\Delta R^2 = 0.22$, mean |SHAP| = 4.7 wt%): the second most important feature despite 37 % missingness. However, the ablation study (Section 5.3.4) showed that removing particle size changes the validation MAE by less than 1 wt%, suggesting that it may partly serve as a proxy for data provenance rather than capturing a purely physical effect.
3. **Feedstock composition** (PE content, PS content, hydrogen content): collectively the third tier of importance, reflecting the well-known dependence of product distributions on polymer chemistry - branching, aromaticity, and heteroatom content.

These rankings are robust across both XGBoost and Random Forest models, and the SHAP-derived feature-target directions are consistent with established pyrolysis chemistry, providing confidence in the physical validity of the learned relationships.

7.3 Outlook

The findings of this thesis suggest several directions for future research that could advance both the scientific understanding and the practical utility of ML-based pyrolysis yield prediction. The following recommendations are organized roughly by implementation complexity.

7.3.1 Near-term extensions

Oil-only model for expanded dataset coverage. The current modeling framework requires complete yield triplets (oil, gas, and char) for each data point, which excludes a substantial number of publications that report only liquid yield. Relaxing this constraint to train a dedicated oil-only model could substantially expand the available dataset, as many publications report only liquid yield without quantifying gas and char fractions separately. The additional data coverage could improve oil yield prediction accuracy and extend the model's applicability to a wider range of feedstocks and experimental conditions.

Transfer learning for site-specific applications. The pre-trained XGBoost models can serve as baselines for reactor-specific applications such as the Burghausen H_2-Reallabor project (Section 6.4). By fine-tuning the literature-trained model with 10-20 experiments conducted on the specific reactor, the model can learn the systematic offset between the site-specific configuration and the literature average. This transfer learning approach combines the broad generalization capability of the literature model with the specificity of site-specific calibration, potentially achieving prediction accuracy below the current noise floor.

Uncertainty quantification. The current framework provides point predictions without confidence intervals. For practical deployment-particularly in process design and optimization-uncertainty estimates are essential. Conformal prediction [46], which provides distribution-free prediction intervals with guaranteed coverage probability, subject to the exchangeability assumption-noting that the grouped structure of the training data may affect coverage guarantees in practice-represents a promising approach that can be applied as a post-processing step to any existing model without retraining. This would enable users to assess prediction reliability on a per-sample basis and identify regions of the input space where the model is poorly calibrated.

7.3.2 Medium-term research directions

Catalytic pyrolysis extension. Non-catalytic pyrolysis represents only a subset of the pyrolysis research landscape. Extending the database and modeling framework to include catalytic pyrolysis data-incorporating catalyst type, acidity, pore structure, and loading ratio as additional input features-would substantially broaden the practical applicability of the models. Zhang et al. [55] have recently demonstrated a dual-stage prediction framework for catalytic pyrolysis that could serve as a starting point for integrating catalyst features into the present database structure.

Reactor-specific submodels. The current approach trains a single global model across all reactor types and operating modes. Given the fundamentally different heat and mass transfer characteristics of batch and continuous reactors-and the substantial missingness in reactor type

classification (67 % "Unknown")-training separate submodels for each major reactor category could reduce prediction error by capturing reactor-specific yield distributions. This approach would require either more complete reactor type annotation in existing data or active data collection targeting underrepresented reactor categories.

Advanced imputation strategies. The median imputation employed in this work is conservative but information-lossy: features with more than 50 % missingness are effectively reduced to binary indicators of "reported vs. not reported" (Section 6.3.2). Alternative strategies such as multiple imputation by chained equations (MICE), model-based imputation using Random Forest, or the explicit incorporation of missingness indicators as additional features could extract more predictive information from partially observed variables.

7.3.3 Long-term vision

Open-access database publication. The compiled database of 795 data points from 127 publications represents a community resource that could accelerate progress across multiple research groups. Publishing the database as an open-access dataset with standardized metadata, transparent provenance documentation, and a clear contribution protocol would enable independent validation, extension, and benchmarking. Such a resource does not currently exist for plastic pyrolysis data and could serve as a foundation for future collaborative efforts.

Standardized evaluation benchmarks. The data leakage analysis presented in this thesis highlights the need for standardized evaluation protocols in the field. Establishing a community-accepted benchmark-comprising a fixed dataset, fixed cross-validation splits (with paper-level grouping), and fixed evaluation metrics-would enable meaningful comparison of modeling approaches across studies. This is analogous to established benchmarks in computer vision (ImageNet [15]) and natural language processing (GLUE [51]) that have driven rapid progress through transparent and reproducible evaluation.

Multi-task learning with mass balance constraints. The current independent modeling of oil, gas, and char yields does not enforce the physical constraint that yields must sum to approximately 100 wt %. A multi-task learning framework that jointly predicts all three yields while incorporating mass balance as a soft or hard constraint could improve consistency and potentially reduce individual target errors by sharing information across related predictions. Physics-informed neural networks [43] that embed thermodynamic constraints directly into the loss function represent a promising avenue for combining data-driven flexibility with physical interpretability.

Bibliography

[1] Aguado, José ; Serrano, David P. ; Escola, José M.: Fuels from waste plastics by thermal and catalytic processes: A review. In: *Industrial & Engineering Chemistry Research* 47 (2008), Nr. 21, p. 7982–7992.

[2] Akiba, Takuya ; Sano, Shotaro ; Yanase, Toshihiko ; Ohta, Takeru ; Koyama, Masanori: Optuna: A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, p. 2623–2631.

[3] Al-Salem, S.M. ; Lettieri, P. ; Baeyens, J.: Recycling and recovery routes of plastic solid waste (PSW): A review. In: *Waste Management* 29 (2009), Nr. 10, p. 2625–2643.

[4] Alabdrabalnabi, Abdulrahman ; Lee, Changyong ; Abiodun, Oluwatobi ; Zhou, Kang ; Lee, Heeyoung ; Chung, Suk H.: Machine learning to predict biochar and bio-oil yields from co-pyrolysis of biomass and plastics. In: *Fuel* 328 (2022), p. 125303.

[5] Anene, Azubuike F. ; Fredriksen, Siw B. ; Sætre, Kai A. ; Tokheim, Lars-Andre: Experimental Study of Thermal and Catalytic Pyrolysis of Plastic Waste Components. In: *Sustainability* 10 (2018), Nr. 11, p. 3979.

[6] Belden, Emma R. ; Rando, Matthew P. ; Ferrara, Owen G. ; Hower, John C. ; Bland, Anne E.: Machine Learning Predictions of Oil Yields Obtained by Plastic Pyrolysis and Application to Thermodynamic Analysis. In: *ACS Engineering Au* 3 (2023), Nr. 2, p. 91–101.

[7] Breiman, Leo: Random Forests. In: *Machine Learning* 45 (2001), Nr. 1, p. 5–32.

[8] Bundesministerium für Bildung und Forschung (BMBF): *Wasserstoff-Forschungsprojekt im bayerischen Chemiedreieck (Reallabore der Energiewende)*. 2022. – URL `https://www.fona.de/de/bund-foerdert-wasserstoff-forschungsprojekt-im-bayerischen-chemiedreieck`. – Accessed: 2026-04-13.

[9] ChemDelta Bavaria / Reallabor Burghausen Konsortium: *H_2-Reallabor Burghausen – Projektbeschreibung*. 2024. – URL `https://www.reallabor-burghausen.de/h2-reallabor/reallabor-burghausen`. – Accessed: 2026-04-13.

[10] Chen, Sihan ; Yuan, Zhilong ; Wang, Ye ; Sun, Yifei: Machine learning prediction of plastic pyrolysis product yields. In: *Energy and Environmental Protection* (2024).

[11] Chen, Tianqi ; Guestrin, Carlos: XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p. 785–794.

[12] Cheng, Yi ; Ekici, Ecrin ; Yildiz, Güray ; Yang, Yang ; Coward, Brad ; Wang, Jiawei: Applied machine learning for prediction of waste plastic pyrolysis towards valuable fuel and chemicals production. In: *Journal of Analytical and Applied Pyrolysis* 169 (2023), p. 105857.

[13] Conversio Market & Strategy GmbH: Stoffstrombild Kunststoffe in Deutschland 2023 / Conversio Market & Strategy GmbH. Mainaschaff, Germany, 2024. – Research Report. – URL `https://www.bkv-gmbh.de/files/bkv/studien/Kurzfassung%20Stoffstrombild%202023.pdf`. Commissioned by BKV, PlasticsEurope Deutschland, VDMA. Published November 2024. Summary version (35 p.) freely available at URL.

[14] Czajczyńska, D. ; Anguilano, L. ; Ghazal, H. ; Krzyżyńska, R. ; Reynolds, A.J. ; Spencer, N. ; Jouhara, H.: Potential of pyrolysis processes in the waste management sector. In: *Thermal Science and Engineering Progress* 3 (2017), p. 171–197.

[15] Deng, Jia ; Dong, Wei ; Socher, Richard ; Li, Li-Jia ; Li, Kai ; Fei-Fei, Li: ImageNet: A Large-Scale Hierarchical Image Database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, p. 248–255.

[16] Elordi, G. ; Olazar, M. ; Lopez, G. ; Artetxe, M. ; Bilbao, J.: Product Yields and Compositions in the Continuous Pyrolysis of High-Density Polyethylene in a Conical Spouted Bed Reactor. In: *Industrial & Engineering Chemistry Research* 50 (2011), Nr. 11, p. 6650–6659.

[17] FakhrHoseini, Seyed M. ; Dastanian, Majid: Predicting Pyrolysis Products of PE, PP, and PET Using NRTL Activity Coefficient Model. In: *Journal of Chemistry* 2013 (2013), p. 487676.

[18] Friedman, Jerome H.: Greedy function approximation: A gradient boosting machine. In: *The Annals of Statistics* 29 (2001), Nr. 5, p. 1189–1232.

[19] Geslin, Alexis ; Vlijmen, Bruis van ; Cui, Xiao ; Bhargava, Arjun: Selecting the appropriate features in battery lifetime predictions. In: *Joule* 7 (2023), p. 1956–1965.

[20] Geyer, Roland ; Jambeck, Jenna R. ; Law, Kara L.: Production, use, and fate of all plastics ever made. In: *Science Advances* 3 (2017), Nr. 7, p. e1700782.

[21] Goodfellow, Ian ; Bengio, Yoshua ; Courville, Aaron: *Deep Learning*. MIT Press, 2016. – URL `https://www.deeplearningbook.org`.

[22] Hooda, Sanjeevani ; Mondal, Prasenjit: Predictive modeling of plastic pyrolysis process for the evaluation of activation energy: Explainable artificial intelligence based comprehensive insights. In: *Journal of Environmental Management* 360 (2024), p. 121189.

[23] Jung, Sang-Hyun ; Cho, Myung-Hyun ; Kang, Bong-Sik ; Kim, Joo-Sik: Pyrolysis of a fraction of waste polypropylene and polyethylene for the recovery of BTX aromatics using a fluidized bed reactor. In: *Fuel Processing Technology* 91 (2010), Nr. 3, p. 277–284.

[24] Junginger, Julia: *Waste-to-X: Facilitating Machine Learning Methods for Waste Plastic Pyrolysis*, Technical University of Munich, Chair of Energy Systems, Semester Thesis, 2024. – Unpublished student thesis.

[25] Kapoor, Sayash ; Narayanan, Arvind: Leakage and the reproducibility crisis in machine-learning-based science. In: *Patterns* 4 (2023), Nr. 9, p. 100804.

[26] Lehrstuhl für Energiesysteme, Technische Universität München: *H_2-Reallabor Burghausen – Projektseite Lehrstuhl für Energiesysteme.* 2024. – URL `https://www.epe.ed.tum.de/es/forschung/projekte/h2-reallabor/`. – Accessed: 2026-04-13.

[27] Li, Hualiang ; Wu, Zhenzhen ; He, Hongyuan ; Feng, Shi ; Zhou, Yunqing ; Shi, Chuanqi ; Tu, Xin ; Yan, Jianhua ; Zhang, Hao: Machine learning-driven product prediction and process optimization for catalytic pyrolysis of polyolefin plastics. In: *Energy Conversion and Management* 333 (2025), p. 119789.

[28] Li, Jie ; Liu, Taiyang ; Palansooriya, Kumuduni N. ; Yu, Di ; Wan, Gan ; Sun, Lushi ; Chang, Scott X. ; Wang, Yin: Zeolite-catalytic pyrolysis of waste plastics: Machine learning prediction, interpretation, and optimization. In: *Applied Energy* 382 (2025), p. 125258.

[29] Little, Roderick J. A. ; Rubin, Donald B.: *Statistical Analysis with Missing Data.* 2nd. Hoboken, NJ : Wiley-Interscience, 2002 (Wiley Series in Probability and Statistics).

[30] Lopez, Gartzen ; Artetxe, Maite ; Amutio, Maider ; Alvarez, Jon ; Bilbao, Javier ; Olazar, Martin: Recent advances in the gasification of waste plastics. A critical overview. In: *Renewable and Sustainable Energy Reviews* 82 (2018), p. 576–596.

[31] Lundberg, Scott M. ; Erion, Gabriel ; Chen, Hugh ; DeGrave, Alex ; Prutkin, Jordan M. ; Nair, Bala ; Katz, Ronit ; Himmelfarb, Jonathan ; Bansal, Nisha ; Lee, Su-In: From local explanations to global understanding with explainable AI for trees. In: *Nature Machine Intelligence* 2 (2020), Nr. 1, p. 56–67.

[32] Lundberg, Scott M. ; Lee, Su-In: A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems* Volume 30, Curran Associates, Inc., 2017.

[33] Mastral, F.J. ; Esperanza, E. ; García, P. ; Juste, M.: Pyrolysis of high-density polyethylene in a fluidised bed reactor. Influence of the temperature and residence time. In: *Journal of Analytical and Applied Pyrolysis* 63 (2002), Nr. 1, p. 1–15.

[34] Miandad, R. ; Barakat, M.A. ; Aburiazaiza, Asad S. ; Rehan, M. ; Nizami, A.S.: Catalytic pyrolysis of plastic waste: A review. In: *Process Safety and Environmental Protection* 102 (2016), p. 822–838.

[35] Miandad, R. ; Nizami, A.S. ; Rehan, M. ; Barakat, M.A. ; Khan, M.I. ; Mustafa, A. ; Ismail, I.M.I. ; Murphy, J.D.: Influence of temperature and reaction time on the conversion of polystyrene waste to pyrolysis liquid oil. In: *Waste Management* 58 (2016), p. 250–259.

[36] Molnar, Christoph: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* 2nd. christophm.github.io, 2022. – URL `https://christophm.github.io/interpretable-ml-book/`.

[37] OECD: Global Plastics Outlook: Economic Drivers, Environmental Impacts and Policy Options. Paris : OECD Publishing, 2022. – Research Report.

[38] Onwudili, Jude A. ; Insura, Nedzad ; Williams, Paul T.: Composition of products from the pyrolysis of polyethylene and polystyrene in a closed batch reactor: Effects of temperature and residence time. In: *Journal of Analytical and Applied Pyrolysis* 86 (2009), Nr. 2, p. 293–303.

[39] Panda, Achyut K. ; Singh, R.K. ; Mishra, D.K.: Thermolysis of waste plastics to liquid fuel: A suitable method for plastic waste management and manufacture of value added products—A world prospective. In: *Renewable and Sustainable Energy Reviews* 14 (2010), Nr. 1, p. 233–248.

[40] Plastics Europe: *Plastics – The Fast Facts 2024*. 2024. – URL `https://plasticseurope.org/knowledge-hub/plastics-the-fast-facts-2024/`. – Accessed: 2026-03-24.

[41] Prokhorenkova, Liudmila ; Gusev, Gleb ; Vorobev, Aleksandr ; Dorogush, Anna V. ; Gulin, Andrey: CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems* Volume 31, Curran Associates, Inc., 2018.

[42] Ragaert, Kim ; Delva, Laurens ; Van Geem, Kevin: Mechanical and chemical recycling of solid plastic waste. In: *Waste Management* 69 (2017), p. 24–58.

[43] Raissi, Maziar ; Perdikaris, Paris ; Karniadakis, George E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. In: *Journal of Computational Physics* 378 (2019), p. 686–707.

[44] Roberts, David R. ; Bahn, Volker ; Ciuti, Simone ; Boyce, Mark S. ; Elith, Jane ; Guillera-Arroita, Gurutzeta ; Hauenstein, Severin ; Lahoz-Monfort, José J. ; Schröder, Boris ; Thuiller, Wilfried ; Warton, David I. ; Wintle, Brendan A. ; Hartig, Florian ; Dormann, Carsten F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. In: *Ecography* 40 (2017), Nr. 8, p. 913–929.

[45] Scheirs, John ; Kaminsky, Walter: Feedstock Recycling and Pyrolysis of Waste Plastics: Converting Waste Plastics into Diesel and Other Fuels. (2006).

[46] Shafer, Glenn ; Vovk, Vladimir: A Tutorial on Conformal Prediction. In: *Journal of Machine Learning Research* 9 (2008), p. 371–421. – URL `https://jmlr.org/papers/v9/shafer08a.html`.

[47] Sharuddin, Siti Dina A. ; Abnisa, Faisal ; Daud, Wan Mohd Ashri W. ; Aroua, Mohamed K.: A review on pyrolysis of plastic wastes. In: *Energy Conversion and Management* 115 (2016), p. 308–326.

[48] Strobl, Carolin ; Boulesteix, Anne-Laure ; Zeileis, Achim ; Hothorn, Torsten: Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. In: *BMC Bioinformatics* 8 (2007), Nr. 25, p. 1–21.

[49] Umweltbundesamt: Kunststoffabfälle / Umweltbundesamt. Dessau-Roßlau, 2023. – Research Report. – URL `https://www.umweltbundesamt.de/daten/ressourcen-abfall/verwertung-entsorgung-ausgewaehlter-abfallarten/kunststoffabfaelle`. Accessed: 2026-03-24.

[50] Umweltbundesamt: Aufkommen und Verwertung von Verpackungsabfällen in Deutschland im Jahr 2022 / Umweltbundesamt. Dessau-Roßlau, 2024. – Research Report. – URL `https://www.umweltbundesamt.de/daten/ressourcen-abfall/verwertung-entsorgung-ausgewaehlter-abfallarten/verpackungsabfaelle`. Accessed: 2026-03-24.

[51] Wang, Alex ; Singh, Amanpreet ; Michael, Julian ; Hill, Felix ; Levy, Omer ; Bowman, Samuel R.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, 2018.

[52] Williams, Elizabeth A. ; Williams, Paul T.: Analysis of products derived from the fast pyrolysis of plastic waste. In: *Journal of Analytical and Applied Pyrolysis* 40–41 (1997), p. 347–363.

[53] Wong, S.L. ; Ngadi, N. ; Abdullah, T.A.T. ; Inuwa, I.M.: Current state and future prospects of plastic waste as source of fuel: A review. In: *Renewable and Sustainable Energy Reviews* 50 (2015), p. 1167–1180.

[54] Zentrale Stelle Verpackungsregister: Recyclingquoten 2024 / Zentrale Stelle Verpackungsregister. Osnabrück, Germany, 2024. – Research Report.

[55] Zhang, Xu ; Wu, Shengji ; Ge, Ben ; Zhou, Qing ; Yan, Bin ; Chen, Zezhou: A dual-stage machine learning framework for comprehensive characterization of waste plastic pyrolysis oils. In: *Journal of Analytical and Applied Pyrolysis* 192 (2025), p. 107300.

[56] Zou, Hui ; Hastie, Trevor: Regularization and Variable Selection via the Elastic Net. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005), Nr. 2, p. 301–320.

A Appendix

A.1 AI Screening Prompt for Literature Triage

```
You screen tab-delimited WoS records for non-catalytic plastic
pyrolysis papers to extend an ML training database.
INPUT FIELDS (tab-delimited per line): PT, AU, TI, SO, AB, DI, PY.
Preserve input order.
INCLUSION (all required):
- Lab-scale reactor >10 g (batch/semi-batch/continuous;
  fixed/fluidized bed, rotary kiln, auger, autoclave).
- Temperature variation: >=3 temperatures per plastic type
  (critical for ML).
- Mass balance >=85% with liquid, gas, char (if one missing
  but inferable, accept).
- Non-catalytic data; if both catalytic and thermal are
  present, use thermal only.
- Process parameters: temperature stated; ideally heating
  rate; plus residence time or reactor type or feedstock info.
- Plastics: PE/HDPE/LDPE/LLDPE, PP, PS/EPS, PVC, PET, mixed
  plastics allowed. Exclude MSW and general waste streams
  containing non-plastic fractions.
- Experimental data only.
EXCLUSION (any -> low priority):
- TGA/micropyrolyzer/analytical pyrolysis; sample <10 g.
- Single temperature per plastic type.
- No mass balance (only oil) or mass balance <85% or >115%.
- Purely catalytic without thermal baseline.
- Tire/rubber, biomass co-pyrolysis, HTL/supercritical water.
- Simulation/CFD/ASPEN only; review papers.
OUTPUT (JSON only, no extra text):
{
  "records": [
    {
      "paper_name": "Surname et al., 2022",
      "priority": "high|medium|low",
      "doi_url": "https://doi.org/<DI>" or null,
      "reason": "One concise sentence."
    }
  ]
}
Field rules:
- paper_name: first author surname + "et al., YYYY".
- priority: high (meets inclusion, >=3 temps, good mass
  balance), medium (usable but gaps), low (fails key
  criteria).
- doi_url: use DI if present else null.
- reason: cite the decisive factors.
```

```
INSTRUCTIONS:
- Return exactly one JSON object with records array; one
  record per input line; preserve order; no markdown or prose.
- Prefer high only when all core criteria met; otherwise
  medium if partially usable; else low.
```

The above prompt was used with Gemini 2.5 Pro to screen approximately 7,400 Web of Science records for relevance to the non-catalytic plastic pyrolysis database. Each batch of tab-delimited records (fields: PT, AU, TI, SO, AB, DI, PY) was submitted with this prompt. The model returned structured JSON output with priority classifications and justifications for each paper.

The full-text extraction prompt (used for Stage 2 PDF processing via `send_pdf_batches.py`) followed a similar structure but targeted quantitative yield extraction into the 52-column database schema described in Section 3.2.1. Each PDF was converted to Markdown using `pymupdf4llm` and submitted individually to the LLM, which returned structured JSON containing the screening decision, mass balance quality assessment, and extracted experimental data points. Both prompts were iteratively refined based on false positive and false negative analysis during the screening process.

A.2 Feature Missingness in the Raw Database

Table A.1 reports the missingness rate for all features in the raw integrated database (1,072 rows, before yield-based filtering) and their treatment in the final preprocessing pipeline. Features are grouped by category. Missingness rates are computed as the fraction of rows with a missing (NaN) value in that column.

Table A.1: Missingness rates for all candidate features in the raw integrated database (1,072 rows). Features with >90 % missingness are excluded from the ML feature set. All retained continuous features are imputed using training-fold median values within each GroupKFold iteration.

Category	Feature	Missing (%)	Treatment
Process Conditions			
	Temperature	0.0	Required (filtered out if missing)
	Heating Rate	83.2	Retained; median imputed
	Particle Size	36.0	Retained; median imputed
	Sample Feed Size	70.2	Retained; median imputed
	Vapor Residence Time	94.9	**Excluded** (>90 %)
	Solid Residence Time	100.0	**Excluded** (>90 %)
Feedstock Composition (wt%)			
	HDPE	30.8	Retained; median imputed
	LDPE	31.7	Retained; median imputed
	LLDPE	36.0	Retained; median imputed
	PE (unspecified)	32.1	Retained; median imputed
	PP	26.1	Retained; median imputed
	PS	24.2	Retained; median imputed
	PVC	36.0	Retained; median imputed
	PET	34.1	Retained; median imputed
	HIPS	36.6	Retained; median imputed

Table A.1 (continued)

Category	Feature	Missing (%)	Treatment
	ABS	37.1	Retained; median imputed
Ultimate Analysis (wt%)			
	C	24.1	Retained; median imputed
	H	24.1	Retained; median imputed
	N	36.8	Retained; median imputed
	O	32.1	Retained; median imputed
	S	44.3	Retained; median imputed
	Cl	42.6	Retained; median imputed
	Br	41.0	Retained; median imputed
	Sb	41.0	Retained; median imputed
Reactor Characteristics			
	Reactor Type	66.7	Retained; unknown → "Unknown" category
	Batch/Continuous	66.7	Retained; unknown → −1 encoding
Product Composition - Excluded (not primary targets)			
	Liquid C5-C12	63.6	**Excluded** (secondary target)
	Liquid C13-C20	63.7	**Excluded** (secondary target)
	Liquid Wax	61.6	**Excluded** (secondary target)
	Liquid Aromatics	69.9	**Excluded** (secondary target)
	Liquid Styrene	78.0	**Excluded** (secondary target)
	Liquid BTX	80.1	**Excluded** (secondary target)
	Gas C1	55.6	**Excluded** (secondary target)
	Gas C2	56.4	**Excluded** (secondary target)
	Gas C3	56.7	**Excluded** (secondary target)
	Gas C4	56.7	**Excluded** (secondary target)

A.3 Web of Science Search Query

The following query string was executed in January 2025 on the Web of Science Core Collection (Science Citation Index Expanded), with document type restricted to "Article" and no timespan restriction. The query returned approximately 7,400 results, when searching in "topic".

```
TS=( ("pyrolysis" OR "pyrolytic" OR "thermal degradation"
      OR "thermal cracking" OR "thermal decomposition"
      OR "thermolysis")
  AND
  ("polyethylene" OR "PE" OR "HDPE" OR "LDPE" OR "LLDPE"
   OR "polypropylene" OR "PP"
   OR "polystyrene" OR "PS" OR "EPS"
   OR "polyvinyl chloride" OR "PVC"
   OR "polyethylene terephthalate" OR "PET"
   OR "plastic waste" OR "waste plastic*"
   OR "polymer waste" OR "mixed plastic*"
   OR "plastic mixture" OR "MPW" OR "MSW")
  AND
  ("liquid yield" OR "oil yield" OR "wax yield"
   OR "condensate" OR "gas yield"
   OR "gaseous product*" OR "char yield"
   OR "solid yield" OR "residue"
   OR "mass balance" OR "product distribution"
   OR "product yield*" OR "conversion"
   OR "selectivity")
)
```

Exclusion terms (e.g., “NOT biomass”, “NOT TGA”) were deliberately omitted to avoid systematic bias; irrelevant papers were removed during the subsequent multi-stage manual screening workflow (Section 3.1.2).

B Source Publications

Table B.1 lists all 132 source publications contributing experimental data to the integrated pyrolysis database (1072 data points total). Papers are identified by the internal `Source_Paper_ID` used throughout the preprocessing pipeline and cross-validation framework.

Table B.1: Complete list of source publications in the integrated pyrolysis database. n denotes the number of data points contributed by each paper. Source abbreviations: IZTECH = IZTECH_v5 predecessor database, DS-2 = dataset_2 predecessor database, AI-Screen = AI-assisted literature screening, Manual = manual extraction, PDF-FT = PDF full-text extraction.

ID	Reference	n	Source
1	Park et al. (2020)	18	IZTECH, DS-2
2	Elordi et al. (2011)	5	IZTECH
3	Aguado et al. (2002)	11	IZTECH
4	Al-Salem et al. (2020)	7	IZTECH, DS-2
5	Arabiourrutia et al. (2012)	5	DS-2
6	Arabiourrutia et al. (2017)	6	IZTECH, DS-2
7	Artetxe et al. (2010)	8	IZTECH, DS-2
8	Artetxe et al. (2015)	21	IZTECH, DS-2
9	Berrueco et al. (2002)	23	IZTECH, DS-2
10	Cho et al. (2010)	12	IZTECH, DS-2
11	Fraczak et al. (2021)	10	IZTECH
12	Jin et al. (2018)	19	IZTECH, DS-2
13	Jung et al. (2010)	14	IZTECH, DS-2
14	Jung et al. (2013)	21	IZTECH, DS-2
15	Kaminsky and Eger (2001)	14	IZTECH, DS-2
16	Kaminsky and Franck (1991)	19	IZTECH, DS-2
17	Kaminsky et al. (1995)	10	IZTECH, DS-2
18	Kaminsky et al. (1996)	8	IZTECH, DS-2
19	Kaminsky et al. (2000)	16	IZTECH, DS-2
20	Kaminsky et al. (2004)	3	IZTECH, DS-2
21	Kang et al. (2008)	16	IZTECH, DS-2
22	Kim et al. (1997)	8	IZTECH, DS-2
23	Liu et al. (2000)	6	IZTECH
24	Lopez et al. (2010)	14	IZTECH, DS-2
25	Mastellone et al. (2002)	9	IZTECH
26	Mastral et al. (2002)	20	IZTECH
27	Mastral et al. (2003)	5	IZTECH
28	Mastral et al. (2006)	37	IZTECH, DS-2
29	Mastral et al.-2007a	28	IZTECH, DS-2

Continued on next page

Table B.1 continued

ID	Reference	n	Source
30	Milne et al. (1999)	12	IZTECH, DS-2
31	Miskolczi et al. (2004)	26	IZTECH, DS-2
32	Miskolczi et al. (2006)	4	IZTECH
33	Miskolczi et al. (2008)	11	IZTECH, DS-2
34	Miskolczi et al. (2009)	12	IZTECH, DS-2
35	Park et al. (2019)	25	IZTECH, DS-2
36	Park et al. (2019)	15	IZTECH, DS-2
37	Park et al. (2018)	15	IZTECH, DS-2
38	Predel and Kaminsky (2000)	13	IZTECH, DS-2
39	Simon et al. (1996)	20	IZTECH, DS-2
40	Walendziewski (2005)	6	IZTECH
41	Williams and Williams-1999a	5	IZTECH
42	Williams and Williams-1999b	5	IZTECH
43	Williams and Williams (1997)	5	IZTECH
44	Yoshioka et al. (2004)	6	IZTECH
45	Zhao et al. (2020)	6	IZTECH
46	Artetxe et al. (2013)	1	IZTECH
47	Artetxe et al. (2012)	1	IZTECH
48	Artetxe et al.-2012b	1	IZTECH
49	Auxilio et al. (2017)	3	IZTECH
50	Ide et al. (1984)	1	IZTECH
51	Aguado et al. (2002)	1	IZTECH
52	Angyal et al. (2009)	1	IZTECH
53	Borsella et al. (2018)	5	IZTECH
54	Borsodi et al. (2011)	3	IZTECH
56	Del Remedio Hernandez et al (2007)	4	IZTECH
57	Donaj et al. (2012)	7	IZTECH
58	Ibanez et al. (2014)	1	IZTECH
59	Koc and Bilgesü (2007)	15	IZTECH
60	Kodera et al. (2006)	3	IZTECH
61	Lei et al. (2018)	4	IZTECH
62	Marcilla et al. (2007)	4	IZTECH
63	Mastral et al.-2006 (Gas)	1	IZTECH
64	Mertinkat et al. (1999)	3	IZTECH
65	Miskolczi et al. (2009)	2	IZTECH
66	Miskolczi et al. (2016)	1	IZTECH
67	Murata et al. (2010)	3	IZTECH
68	Serrano et al. (2001)	8	IZTECH
69	Takuma et al. (2000)	1	IZTECH
70	Uemichi et al. (1984)	1	IZTECH
71	Zaiter (2014)	1	IZTECH

Continued on next page

Table B.1 continued

ID	Reference	n	Source
72	Conesa et al. (1997)	4	IZTECH
73	Westerhout et al. (1998)	4	IZTECH
74	Jeong et al. (2022)	7	IZTECH
75	Kaminsky (1985)	4	IZTECH
76	Kaminsky (1991)	11	IZTECH
77	Kaminsky (1995)	3	IZTECH
79	Miller et al. (2005)	3	IZTECH
80	Hardman and Wilson (1998)	1	IZTECH
81	Scott et al. (1990)	9	IZTECH
82	Bockhorn et al. (1998)	1	IZTECH
83	Sodero et al. (1996)	3	IZTECH
84	Westerhout et al.-1998b	14	IZTECH
85	Zeller et al. (2021)	3	IZTECH
86	Angyal et al. (2007)	4	IZTECH
87	Kaminsky and Kim (1999)	4	IZTECH
88	Hernandez et al. (2006)	5	IZTECH
89	Hirota and Fagan (1992)	1	IZTECH
90	Murata et al.-2009a	4	IZTECH
91	Murata et al.-2009b	1	IZTECH
92	Murata et al. (2002)	13	IZTECH
240	Abdullah et al. (2018)	10	AI-Screen
241	Aisien et al. (2021)	2	AI-Screen
242	Akgun et al. (2021)	9	AI-Screen
243	Albor et al. (2023)	9	AI-Screen
245	Ates et al. (2013)	6	AI-Screen
246	Chaiya et al. (2020)	3	AI-Screen
247	Choi et al. (2024)	3	AI-Screen
249	Costa et al. (2007)	4	AI-Screen
250	Cozzani et al. (1997)	9	AI-Screen
251	Dash et al. (2015)	4	AI-Screen
252	de Freitas Costa et al. (2023)	3	AI-Screen
253	Dutta et al. (2021)	13	AI-Screen
254	Horvat et al. (1999)	6	AI-Screen
255	Kamo et al. (2004)	3	AI-Screen
256	Karaduman et al. (2001)	10	AI-Screen
257	Li et al. (2021)	5	AI-Screen
258	Liu et al. (2022)	3	AI-Screen
259	Lodh et al. (2023)	4	AI-Screen
260	Lopez-Urionabarrenechea et al. (2011)	3	AI-Screen
261	Madhu et al. (2022)	7	AI-Screen
262	Maniscalco et al. (2021)	3	AI-Screen

Continued on next page

Table B.1 continued

ID	Reference	n	Source
263	Marais et al. (2024)	6	AI-Screen
264	Matthiesen et al. (2024)	7	AI-Screen
265	Mousavi et al. (2022)	15	AI-Screen
266	Ortega et al. (2023)	6	AI-Screen
267	Papuga et al. (2016)	5	AI-Screen
268	Papuga et al. (2024)	16	AI-Screen
269	Parku et al. (2025)	9	AI-Screen
270	Rangarajan et al. (1998)	3	AI-Screen
272	Shi et al. (2023)	9	AI-Screen
273	Thahir et al. (2019)	7	AI-Screen
275	Verma et al. (2024)	7	AI-Screen
276	Walendziewski et al. (2001)	8	AI-Screen
277	Wang et al. (2023)	8	AI-Screen
278	Zangi et al. (2025)	8	AI-Screen
279	Zayoud et al. (2022)	7	AI-Screen
280	Cao et al. (2024)	3	Manual
281	Deka & Misra (2024)	23	Manual
282	Hwang et al. (2025)	5	Manual
283	Meena et al. (2024)	27	Manual
284	Singh et al. (2019)	16	Manual
285	Zhang et al. (2020)	4	Manual