

**Data-Driven Prediction of Liquid Accumulation in
Gas Wells Using Machine Learning Models**

by

Mukhtar Sarsenbay

Submitted to the School of Engineering and Digital Sciences
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

April 2026

© Nazarbayev University 2026. All rights reserved.

Author
School of Engineering and Digital Sciences
April 11, 2026

Certified by
Iliyas Tursynbek
Supervisor
Thesis Supervisor

Accepted by
Chair
Adnan Yazici

Data-Driven Prediction of Liquid Accumulation in Gas Wells

Using Machine Learning Models

by

Mukhtar Sarsenbay

Submitted to the School of Engineering and Digital Sciences
on April 11, 2026, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Liquid loading (liquid accumulation) is a major contributor to production decline and flow instability in mature gas wells. Physics-based critical-velocity correlations provide useful engineering guidance but are limited for operational surveillance because they require uncertain inputs and do not explicitly model the gradual temporal buildup preceding a loading episode. The thesis reframes liquid loading prediction as a *time-to-event* problem and develops a discrete-time survival (hazard) modeling pipeline using daily operational telemetry aligned with engineer-provided loading annotations. Using a public multi-well dataset of 30 wells, we construct a time-to-first-event cohort with 5,463 daily observations, including 27 first loading events and 3 right-censored wells. Models are evaluated under a leakage-resistant protocol with a well-held-out split (24 wells for training, 6 unseen wells for testing), and operating thresholds are selected using out-of-fold predictions across training wells. Gradient-boosted decision trees provide the strongest early-warning signal on unseen wells (ROC-AUC = 0.929, PR-AUC = 0.145). At an out-of-fold tuned operating point, the hazard model achieves precision = 0.116, recall = 0.833, and F1 = 0.204, corresponding to 5 of 6 events detected with 38 false alarms across 415 test days. The horizon-specific classification algorithm to predict whether loading happens in the coming $H \in \{7, 14, 30\}$ days shows time-varying AUC scores of 0.458, 0.530, and 0.566, respectively, suggesting some level of predictive power but with increasing accuracy at longer horizons. In addition, the classification of loading severity (mild, moderate, and severe) produces a macro F1 score of 0.48 ± 0.27 when using well-balanced cross-validation. In summary, the thesis findings suggest that the discrete-time approach to survival analysis serves as an alerting system that can measure loading risk rankings; however, the method cannot be used for autonomous alerting until better alarm rules are developed.

Keywords. gas wells, liquid loading, liquid accumulation, time-to-event, discrete-time survival, hazard modeling, rare-event detection, well-level generalization, PR-AUC, time-dependent AUC, gradient boosting, severity classification

Thesis Supervisor: Iliyas Tursynbek
Title: Supervisor

Contents

Abbreviations	7
1 Introduction	9
1.1 Background and Motivation	9
1.2 Problem Statement	10
1.3 Research Objectives and Contributions	11
1.3.1 Scope of the thesis	12
1.4 Thesis Organization	13
2 Literature Review	15
2.1 From Physics-Based Diagnosis to Data-Driven Prediction	15
2.2 Recent Studies Relevant to This Thesis	16
2.3 How the Literature Informs the Present Thesis	17
3 Theoretical Background and Methodology	20
3.1 Theoretical Background	20
3.1.1 Multiphase Flow Fundamentals	20
3.1.2 Machine Learning Theory	24
3.2 Dataset Description and Exploratory Data Analysis	26
3.2.1 Dataset Source	26
3.2.2 Feature Description	28
3.2.3 Target Variable	29
3.2.4 Data Distribution and Exploratory Data Analysis	30

3.3	Methodology	33
3.3.1	Data Preprocessing	34
3.3.2	Pipeline Architecture	38
3.3.3	Feature Engineering	39
3.3.4	Train–Test Split and Validation	42
3.3.5	Evaluation Metrics	43
3.3.6	Summary	44
3.4	Model Development and Validation	44
3.4.1	Models Implemented	45
3.4.2	Hyperparameter Tuning	47
3.4.3	Model Complexity Analysis and Diagnostics	48
3.4.4	Validation Summary and Model Selection Criteria	50
4	Results	54
4.1	Results	54
4.1.1	Hazard prediction on unseen wells	54
4.1.2	Operating point on unseen wells (OOF-tuned threshold) . . .	54
4.1.3	Secondary task: severity classification	57
5	Discussion	61
5.1	Interpretation of the Results	61
5.2	Limitations	63
5.3	Ethical and Safety Considerations	65
6	Conclusion	67
	Bibliography	70

Abbreviations

ANN	Artificial Neural Network
AUC	Area Under the Curve
BCE	Binary Cross-Entropy
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
EDA	Exploratory Data Analysis
FN	False Negative
FP	False Positive
GBDT	Gradient-Boosted Decision Tree
GRU	Gated Recurrent Unit
IQR	Interquartile Range
ML	Machine Learning
OOF	Out-of-Fold
PR-AUC	Area Under the Precision-Recall Curve
RF	Random Forest
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
SCADA	Supervisory Control and Data Acquisition
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine
tdAUC	Time-Dependent Area Under the Curve
TN	True Negative
TP	True Positive

Chapter 1

Introduction

One of the principal reasons for deliverability reduction and well operations instability in mature gas wells is liquid loading. When reservoir pressure declines and gas velocity reduces, the capability of the well to move produced fluids to the surface is impaired, causing liquid accumulation in the wellbore, increased hydrostatic backpressure, rate cycling/slugging, and even leading to intermittent or shut-in well operations in the worst-case scenario. Modern summarizing literature in petroleum engineering indicates liquid loading as one of the most critical problems in production from mature gas wells [1]. Traditional approaches based on critical-velocity and critical-flow-rate criteria are recognized by engineers as a fundamental reference on the subject matter [2, 3]. Meanwhile, practical experience in recent applications demonstrates that their implementation heavily relies on the chosen flow regime, accurate estimation of fluid properties, and availability of reliable pressure/temperature data [4, 5]. Liquid loading, therefore, becomes both a flow assurance problem and an operational surveillance task: it is necessary to detect deterioration in time to act before well production stops completely.

1.1 Background and Motivation

Modern studies demonstrate the trend from mere thresholding towards predictive and hybrid models. The physics-oriented line of research advances the knowledge of liquid

film and mist flow formation and film rupture when the initiation of liquid loading depends on the transition of multiphase state parameters rather than some static threshold value [6, 7]. Recent studies that deal with the liquid film dynamics and droplet entrainment also draw similar conclusions when they investigate horizontal wells with more complex fluid flows [8, 9]. Meanwhile, the emergence of machine learning methods brings a shift towards viewing liquid loading as an inference task using available telemetry data from the production process. Some of the latest studies are devoted to the onset ranking and abnormal pattern identification in relation to the onset problem [10, 11]. Some studies consider regression approaches, while others concentrate on the height prediction and temporally based predictions [12, 13, 14]. Hybrid approaches attempt to integrate engineering knowledge or physical meaningful intermediate values to improve the practical use of the solution [15, 16]. In other words, what matters to the operators is not a binary answer indicating whether a certain well is loaded or not; it is the warning about the future loading.

This thesis pragmatic relevance lies precisely in addressing this discrepancy between engineering diagnostics and decision-making processes in operations. A daily risk estimation system would assist with prioritizing wells for further investigation, avoid delays in interventions, and facilitate planning for manually induced deliquification processes. Yet, at the same time, the existing literature reveals an ample opportunity for hype in this regard. Indeed, many of the papers in question often cover only a few oil wells, feature inconsistent task specifications, or fine-tune their threshold values according to impractical scenarios [10, 15, 17]. It is evident that there is a need for a rigorous baseline that utilizes reasonable validation and explicitly acknowledges its limits.

1.2 Problem Statement

The main problem with conventional correlations for liquid loading is not their physical irrelevance but their use as snapshot diagnostics. They are helpful in confirming that the present gas rate might be inadequate under certain assumptions; however,

they do not consider the temporal precursors that are common preceding the identified loading events. The key predictors could be a gradual decline in the gas rate, shifts in the behavior of the wellhead pressure, higher variability, and operational intervention effects. Recent machine learning literature indicates that the precursors can be trained using real data; however, the methods used vary widely depending on the target variable. Some papers rank the likelihood of an imminent onset [10] or identify anomalies [11], while others use regression models for the liquid height or event categorization [12, 13]. Temporal deep learning papers move closer to predicting early warning signs using sequential features [14, 16]. This variety is good science, but it complicates deciding on the best method for daily monitoring.

The issue is resolved in this thesis through the framing of liquid loading as a time-to-event task with respect to discrete time points. The core model does not make a prediction about an arbitrary future label but rather calculates the hazard of the first occurrence of an observable loading event for each well on a particular day, given that no previous event has taken place. The formulation is superior to a binary classification approach since it generates a risk profile per day and clearly delineates an early warning window for action. Furthermore, it allows for a leakage-free assessment procedure where wells themselves are reserved as test sets and threshold values are derived solely from out-of-fold predictions.

1.3 Research Objectives and Contributions

The primary goal of the thesis is the creation of a reproducible pipeline for turning the publicly-available multi-well data from Wang et al. into tabular forms for both survival-style analysis of the first event, as well as analysis of event-day severity levels [18]. The second goal is to measure the performance of robust tabular classifiers on such a heavily-imbalanced task with well-level generalization constraints. The third goal is to perform evaluation based on measures which retain meaning under such heavy skew, like ROC-AUC, PR-AUC, alarm rates, and horizon-specific time-dependent discrimination ability. Finally, the fourth secondary goal is to determine

whether there is sufficient signal within the existing labels for performing coarse severity classification.

The thesis thus represents a significant advance in method, as opposed to simply an increase in numbers. The thesis makes no pretense of being able to displace established engineering practice. Instead, the contribution lies in providing a structured and leakage-aware framework for testing the utility of operational telemetry for early warnings.

1.3.1 Scope of the thesis

The scope of the current thesis paper is established in terms of four major aspects: database, computing platform, modeling scope, and target application.

Dataset. This thesis is carried out only using the multi-well gas production data set introduced by Wang et al [18]. The data include daily operating records for 30 gas wells operated on land between December 2009 and April 2024. For the thesis, the entire data set is used for time-to-first-event modeling analysis purposes, resulting in a total of 5,463 well-day observations for 30 wells. Out of 30, 27 have a documented loading first-time event, while 3 are regarded as right-censored. There are no proprietary data used for the analysis, and hence, all findings

Computational platform. The vital steps for the project, like data preprocessing, feature construction, model training, threshold selection, and evaluation, are completed by Python 3 and open source libraries. For the modelling `scikit-learn`, logistic regression and `HistGradientBoostingClassifier`, along with `pandas` and `numpy` used for data processing. The final choices are summarized in Table 3.2 and Table 3.3.

Modeling boundaries. The research aims at predicting three different but interconnected problems: (i) time-to-event hazard modeling of first-event occurrences, (ii) horizon-specific early warning performance for horizon windows $H = 7, 14, 30$ days, and (iii) broad severity classification on event days. The project does not attempt any modeling of continuous liquid-level regression, reconstruction of downhole pressure, physics-based simulation of multiphase flow processes, or automated intervention optimization. The discussion about some physics-based correlations (e.g., Turner criti-

cal velocity criterion) is presented to illustrate the engineering challenge but without embedding them into the pipeline model.

Intended practical deployment. The models developed in this thesis will serve as a *decision support layer* for well monitoring as opposed to being an automatic alarm or control system. The practical application scenario will involve a daily batch-scoring routine where every well that is currently active will be given a hazard score that can help the engineers determine if the well should be inspected, diagnosed, or deliquified. Automated action based on this decision is not within the scope of this thesis, especially since the model produces false alarms on the hold-out wells, as seen in Section 5.3.

1.4 Thesis Organization

In chapter 2, the literature review is covered, focusing particularly on the definition of the prediction target, the creation of validation methods, and the performance reported in more recent papers. Chapter 3 covers the theoretical framework, data collection, data preprocessing, feature extraction and modeling strategy used, as well as the methodology adopted to perform the evaluation. In chapter 4, the results are presented. The discussion on the results, including their limitations, ethical and safety aspects, is covered in chapter 5. Chapter 6 finalizes the thesis and summarizes directions for future work.

Chapter 2

Literature Review

2.1 From Physics-Based Diagnosis to Data-Driven Prediction

In this context, the available literature on liquid loading phenomena can be better categorized depending on the questions addressed by the respective work. At an early stage, there was interest in diagnosis: Is the current gas velocity sufficiently high to entrain the produced liquid, given the current reservoir conditions? A classical approach to answering this question has been proposed in various forms of critical-velocity models of the Turner type and further refined works that are still broadly used due to their simplicity, physical clarity, and applicability in field operations [2, 3, 4]. However, as it was found from further investigations, loading starts at different regimes, depends on the liquid film characteristics, well geometry, and dynamic changes of thermodynamic parameters. Therefore, deterioration cannot be described only by a threshold, but it occurs gradually [5, 6, 7]. Modern research in the area of liquid film theory, its further development, and the corresponding mechanical approach follows this trend of research as well [8, 9].

This constraint is also why contemporary research has taken up the challenge of data-based prediction. Rather than merely determining if the well is at a critical flow rate at present, such efforts seek to determine if loading is impending, forming, or

if intervention is necessary due to an advanced state. There are large discrepancies between the definitions of these tasks. For example, Jia et al. formulated an approach to the liquid loading onset prediction problem as one of ranking model candidates rather than optimizing performance on a particular criterion alone [10]. Chen et al. addressed liquid loading prediction as an adaptive anomaly detection problem in shale gas wells by acknowledging abnormal precursor behaviors prior to loading [11]. More recently, Wang et al. released an open benchmark of a multi-well dataset for liquid-level and natural gas production, enabling reproducibility not always seen in past proprietary work [18]. Overall, it can be seen that the field has made significant progress towards forecasting problems.

2.2 Recent Studies Relevant to This Thesis

This area of recent literature is especially helpful based on the prediction target. The first group of literature includes papers dealing with quantifying the severity or liquid height estimates. Xia et al. adopted a novel XGBoost prediction workflow using optimization techniques, which is able to predict the accumulation height in the tubing of the wells, indicating the capability of the boosted tree models to utilize non-linear correlations in tabulated production data when the quantified target exists [12]. While this paper is helpful in terms of going one step beyond the binary loaded-unloaded scenario, it relies on the availability of a target which is more difficult and rich in nature compared to just an event annotation. In addition to quantification methods, there have been other works which considered the operational status of the wells as a classification task based on severity. For example, Peng et al. adopted a semi-supervised approach for shale gas well severity classification, whereas Zeng et al. recently employed contrastive learning for better feature discrimination in the severity classification problem [13, 16].

Another area of research focuses directly on the topic of early warning. Chen et al. proposed a method based on deep regression for forecasting gas-well liquid loading, stressing the importance of the temporal modeling aspect due to the slow

process of development of an onset phase [14]. Eita et al. used the combination of physical insights and machine learning in order to forecast liquid loading for a German gas field, which is relevant due to the possible higher acceptance from operators who would be able to interpret the model results in an engineering manner [15]. The mentioned approach seems quite promising for practical application; however, the problem of generalization persists, since it may happen that a successful model requires some field-specific conditions related to the range of operations, treatment methods, and completion techniques. Salisu et al. also discuss the role of good-quality data, adequate features, and realistic validation scheme while describing different machine-learning techniques for liquid loading prediction of gas-condensate wells [17].

As reviewed jointly, it is clear that none of the recent techniques stands out as being superior across all contexts. Rather, it becomes evident that the superiority of each technique depends on the context defined by the target, the type of dataset, and the way it is evaluated. Tree-based approaches work well when the input data are structured and the task is classification or regression [12, 15]. Deep learning approaches for sequential datasets fare well when they include richer sequence structure and the task takes the form of precursor detection [14]. Semi-supervised or contrastive learning approaches hold promise when there are few labels or they are noisy, particularly for severity grading, but they come at a cost of increased model complexity and operability [13, 16].

2.3 How the Literature Informs the Present Thesis

Three topics in the literature affect the way this thesis will be structured. The first one is the definition of the target of prediction. If a machine learning algorithm is trained on event severity, then its task is different from an algorithm which predicts the likelihood of a first-event occurrence, or another algorithm which calculates the liquid height. Comparing models trained for these different tasks without controlling the target will lead to confusing results. The second topic to consider is validation. There have been many successful projects, however, several of them suffer from small

sample sizes, specialized data, or even a biased validation strategy [10, 15, 17]. Finally, performance metrics matter. When dealing with rare events, accuracy alone is not enough because the trade-off between false negatives and false positives should be considered, together with the choice of the optimal threshold.

In light of the above discussion, the proposed thesis does not set out to establish an up-to-date methodology for any and all kinds of liquid loading predictions. Rather, this project aims to narrow its focus to an interesting problem, whose relevance is justified through a careful survey of recent publications: can a leakage-robust, discrete-time survival baseline generate early signals for the first loading events of wells that were unseen during training? This approach relies less on novelty and more on building upon existing research: from deep learning for temporal predictions, to the concern for practical explainability from hybrid modeling papers, and to quantitatively rigorous analysis typical of the new table-benchmarking papers.

Chapter 3

Theoretical Background and Methodology

3.1 Theoretical Background

In this section, an overview of the underlying physical principles for the liquid loading problem in gas wells is discussed, along with some basics on machine learning that will be used in the thesis. Whenever an equation from existing literature related to engineering and/or machine learning is borrowed, the source is referenced.

3.1.1 Multiphase Flow Fundamentals

Critical velocity and the onset of liquid loading. The phenomenon of liquid loading may be thought of as one of transport limitation. For any particular well geometry and fluids, the upflow of the gas stream needs to have enough momentum to transport the liquid droplets and films upwards to the surface. In case the gas velocity drops below a certain threshold, liquid starts accumulating, the hydrostatic pressure heads increase, and the deliverability of the gas stream is adversely affected. The classical critical velocity equation adopted in this thesis work is based on Turner et al. [2].

$$v_c = C \left(\frac{\sigma(\rho_L - \rho_G)}{\rho_G^2} \right)^{\frac{1}{4}}, \quad (3.1)$$

with σ being the surface tension, ρ_L and ρ_G being the fluid density, and C being an empirically-derived constant that takes into account the droplet shape and drag considerations. In reality, Eq. (3.1) is transformed into a critical gas rate by multiplying by the tube cross-sectional area A :

$$q_{g,c} = v_c A. \quad (3.2)$$

In essence, if q_g becomes lower than $q_{g,c}$ for the particular pressure and temperature conditions, the well may be said to be prone to loading. While the criterion is still very popular, empirical data analyses have shown that the selection of correlations and their calibration are important, particularly in case the operating regime differs significantly from those assumed during derivation [5, 15]. Regression analyses also revealed that correlation-based estimates alone were insufficient in cases where quantification was required, not just classification [12].

Gas–liquid flow regimes and transport mechanisms. With the decreasing efficiency of the gas wells, the flow regime changes as well. The flow regime starts as an annular mist flow regime and develops into the churn or slugging flow and intermittent flow regimes as the velocity of the gas becomes lower while the liquid flow increases. The reason why this process is significant lies in the fact that when the flow regime changes, so does the technique for liquid removal as well. When the velocity of the gas is relatively high, the liquid droplets flow with it; however, once it lowers, the liquid films can start reversing, slugs will develop, and flow will turn into an intermittent one. This causes nonlinearity in pressure and flow rates which explains why data-driven models outperform the threshold ones.

Pressure drop relationships in multiphase flow. A useful engineering decomposition of the pressure gradient along the wellbore, standard in multiphase-flow analysis, is:

$$\left(\frac{dP}{dz}\right)_{\text{total}} = \left(\frac{dP}{dz}\right)_{\text{hyd}} + \left(\frac{dP}{dz}\right)_{\text{fr}} + \left(\frac{dP}{dz}\right)_{\text{acc}}, \quad (3.3)$$

The total gradient in most production systems usually comprises hydrostatic, fric-

tional, and acceleration components, with hydrostatic and frictional components normally being dominant. The expression for frictional losses in a single-phase system can be given as:

$$\Delta P_{\text{fr}} = f \frac{L}{D} \frac{\rho v^2}{2}, \quad (3.4)$$

In a system with friction factor f , length of the pipe L , diameter D , density ρ , and velocity of flow v , the story becomes more complicated when one has to deal with multiphase flow. Here, the density of the fluid and the friction losses are not constant; they vary with the percentage of the fluid present (the holdup of the fluid) and the flow, which depends on the rates of gas and liquid being produced. The important aspect of the system, from an operational point of view, is that as the holdup of the fluid increases, the density of the fluid, i.e., ρ_{mix} , also increases, which results in an increase in the pressure drop. This, in turn, decreases the bottomhole flowing pressure, which then decreases the flow of the gas, hence reducing the system's ability to carry fluid.

Basic mass balance and temporal evolution of liquid holdup. In general terms, once accumulation has started, it is all about balancing the amount of incoming fluid with the amount that can be taken away. We can think of $H_L(t)$ as the "effective liquid holdup," i.e., the amount of liquid currently held in the wellbore's control volume. The fundamental dynamic relation can be expressed in a simple balancing equation:

$$\frac{dH_L}{dt} = q_{L,\text{in}}(t) - q_{L,\text{out}}(t), \quad (3.5)$$

In such a system, $q_{L,\text{in}}$, in represents the amount of produced liquid that enters the wellbore, and $q_{L,\text{out}}$ represents the amount that actually makes it to the surface. When $q_{L,\text{out}}$ is limited by the lack of gas velocity, $dH_L/dt > 0$ becomes positive, and we begin to see liquid build-up in the well. Although $H_L(t)$ is not directly observable, we may indirectly gather clues from the operation, such as the pressure trends and the unsteady production rates, which are related to the hidden state. This brings us to favor learning the signatures related to $q_{L,\text{out}}$ rather than trying to estimate $H_L(t)$ from the sparse observations.

Connecting physics to machine learning targets. Physics-based concepts motivate two common data-driven targets:

- **Binary onset classification:** Make predictions about whether accumulation is expected to occur within a certain horizon (or whether accumulation is already present). In simulation-based work, this can take the form of a binary label such as “liquid loading occurs,” which can be used for supervised learning [11].
- **Severity regression:** Alternatively, predict a continuous value such as liquid height or another accumulation proxy. Recent research suggests that boosted trees can be used to accurately predict accumulation height, with techniques to prevent overfitting [12].

The primary goal of operational early warning is typically focused on classification and forecasting, i.e., to produce a calibrated probability that a well will collect liquids within a certain time window in the future, rather than simply diagnosing events that have already occurred [15].

Machine learning loss functions used in this thesis. The choice of loss function depends on whether the task is classification or regression. The standard forms below follow the notation used in Bishop and Hastie et al. [19, 20].

Binary classification (cross-entropy). For a sample i with label $y_i \in \{0, 1\}$ and predicted probability $\hat{p}_i$, the binary cross-entropy loss is:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]. \quad (3.6)$$

Under severe imbalance, class-weighted variants are common:

$$\mathcal{L}_{\text{WBCE}} = -\frac{1}{N} \sum_{i=1}^N [w_1 y_i \log(\hat{p}_i) + w_0 (1 - y_i) \log(1 - \hat{p}_i)], \quad (3.7)$$

where $w_1 > w_0$ increases the penalty for misclassifying rare positive events.

Regression (mean squared error). For continuous targets $y_i \in \mathbb{R}$ and predic-

tions $\hat{y}_i$:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2. \quad (3.8)$$

In liquid loading, regression targets can be either the level of the liquid or its equivalent, whereas classification targets are the occurrence or the probability of its occurrence [12].

3.1.2 Machine Learning Theory

Supervised learning for classification and regression. In supervised learning, the use of labeled data $(\mathbf{x}_i, y_i)$ is employed to learn a function $f(\mathbf{x})$ that may extend beyond the data it is trained on. In the context of the current thesis, $\mathbf{x}_i$ represents the measurements obtained during operation, such as pressures, temperatures, daily production, total production, and engineered time features, while y_i may be a binary value indicating the occurrence or non-occurrence of accumulation or a continuous value indicating the severity [14]. Current liquid loading modeling makes use of both approaches, using deep temporal models for the prediction of the onset of early warnings and boosted trees with regression for the prediction of the liquid level [12].

Bias–variance tradeoff. Model performance depends on the balance between:

- **Bias:** underfitting.
- **Variance:** overfitting.

In other words, linear algorithms like logistic regression have high bias but low variance. However, deep neural nets and properly configured ensemble models have low bias but high variance risk. For multiphase well data, variance risk is even higher because of non-stationary data, interventions, and measurement noise. This is why the general rule is to start with sound and solid baseline models and only complicate the model if there is clear improvement.

Cross-validation and leakage-resistant evaluation. Cross-validation attempts to estimate the generalization capabilities of a model through the process of training

and testing on different subsets of the training dataset. In the normal k-fold validation process, it is assumed that all the cases are independently and identically distributed. Time series violates this assumption since the observations at times close to one another are correlated and cannot be used in training since they are from the future. There is leakage whenever a model is trained using future values and tested using previous values, or vice versa.

Time series models require that the chronological order be maintained for each well when developing the model, but not when evaluating the final model at the well level due to concerns about leakage. The primary issue of concern for multiwell modeling is generalizing to new wells [17].

Overfitting in rare-event prediction. Rare event prediction tends to be even more susceptible to overfitting. This can happen due to the model becoming too focused on a small number of positives or a unique operational pattern and subsequently failing in forward-looking testing on new wells.

- **Model regularization:** Consider L2/L1 regularization, maximum depth of trees, and shrinkage, dropout, and early stopping when necessary.
- **Data strategies:** Implement class weights, set threshold values carefully, and visualize results through precision recall graphs instead of accuracy rates.
- **Calibration:** Employ the outputs from the model itself to get accurate probability estimates to prevent misleading conclusions.

Recent research on deep learning models has shown that recall, F1, and threshold metrics are crucial in loading prediction, showing the expense of missing an event [14]. In boosted tree research, the efficiency of using regularization to enhance boosting has been emphasized [12].

Summary: implications for model design in this thesis. These ideas will inform the key choices made in the process of building a thesis like this one.

1. Feature choice and interpretation have been driven by physical principles, with

velocity and rate being particularly significant, alongside feedback in the form of pressure loss.

2. Liquid loading has been considered from the angle of a supervised early warning, with an emphasis on time splits and metrics indicating rare events.
3. Baseline approaches, which are based on the availability of few labeled data, are compared with deep learning models trained on time-dependent datasets, capable of capturing slow changes.
4. Preventing overfitting and assessing generalization have been given great importance in thesis.

3.2 Dataset Description and Exploratory Data Analysis

This section reviews publicly available datasets that could, in principle, support data-driven modeling of liquid loading and related production instabilities. The most relevant candidates are the *Scientific Data* dataset associated [18], Equinor’s Volve archive [21], and the Petrobras 3W dataset [22]. Of these, the first is the best domain match because it provides multi-well gas-production telemetry together with liquid-level and loading-related annotations. Volve is broader but centered on offshore oil-field operations, while 3W is a useful methodologically for rare-event detection yet does not map directly to gas-well liquid loading. For that reason, the present thesis uses the dataset from [18].

3.2.1 Dataset Source

Primary dataset (Scientific Data / SciDB). The thesis revolves around a multi-well dataset that was introduced in [18]. The dataset is designed to facilitate predictions of natural gas production, along with associated fluid buildup behavior (liquid level and loading), by utilizing time-series data collected during operations. The

dataset contains a number of years of daily resolution data and has per-well identifiers to facilitate validation that accounts for each individual well.

Data used in this thesis. The source archive contains 30 wells with 61,060 daily records (2009-12-15 to 2024-04-18). However, the primary modeling dataset in this thesis is constructed directly from the per-well spreadsheet files by aligning daily measurements with engineer annotations by date and building a *time-to-first-event* cohort. For wells that experience loading, we keep the daily trajectory from start-up until (and including) the first loading day; wells with no observed loading are treated as right-censored. This produces a discrete-time survival dataset (`survival_dataset.csv`) with:

- **Number of wells:** 30.
- **Cohort size:** 5,463 daily observations (time-to-first-event trajectories).
- **First events:** 27 first loading events; **censored wells:** 3.

The full daily archive remains useful for descriptive statistics and exploratory analysis, while the cohort is used for hazard modeling and early-warning evaluation. The file contains the following properties, which match a real-field scenario with long-horizon surveillance and support time-forward and cross-well evaluation.

Alternative datasets considered. Volve (Equinor). The Equinor Volve data set is special in its comprehensiveness as it includes data about production, drilling and completion, seismics, and well logs/models [21]. Yet, as the name suggests, it is still an offshore oil field, and operational processes like water injection make it irrelevant to the current problem statement. It would take considerable effort to normalize the data set in order to use it for benchmarking purposes.

3W dataset (Petrobras). The 3W dataset was specifically designed to detect rare adverse events in offshore oil wells. It is widely used as a benchmark for imbalanced event detection [22]. While the dataset provides significant methodological insights for the application of rare event ML, its event taxonomy does not correspond to gas well liquid accumulation. Therefore, the dataset does not match the domain application for predicting liquid loading.

Rationale for selecting the primary dataset. The dataset in [18] is the best fit for the present thesis because it provides multi-well daily operational time series together with engineer annotations that can be converted into both first-event survival targets and event-day severity labels. This makes end-to-end supervised evaluation possible even under strong class imbalance.

3.2.2 Feature Description

The combined dataset is centered around a limited set of key operational variables and total production signals. To clarify the columns, the headers have been standardized for the purpose of the current text (note that the original document uses a non-English language for the headers). Let $\mathbf{x}_t$ be the feature vector for a well day measurement taken at time t . The main variables used are:

- **Wellhead pressure (MPa):** Instantaneous surface pressure proxy. This value has a physical basis related to the losses due to the presence of phases and the effects of backpressures, especially when the system is under a load.
- **Daily gas rate:** Production rate in units (volume/day) proxy, as found in the source data set. It makes a substantial contribution towards the carrying capacity, since reduction in gas flow rates results in reductions in liquids lifting capacity.
- **Cumulative gas (in 10^4 m^3):** Cumulative production proxy, considering the phase of depletion and maturity.
- **Cumulative water (m^3):** Proxy measure of the water production volume, along with time series of water production.
- **Well identifier:** Categorical index of wells which allows well-based grouping, leakage resistant splits.
- **Date:** Timestamp for chronological ordering and forward validation.

Constant features. Two variables (*inlet pressure* and *inlet temperature*) in the provided consolidated data file are constant across all observations. Variables that have constant values do not provide any predictive power and will be omitted from correlation and modeling. These are regarded as characteristics of the data set rather than attributes of the well itself.

Domain-motivated feature examples. In [18], which is an article about the underlying data, there is a larger list of operational parameters that might be useful for predicting liquid loading problems. These include tubing pressure, casing pressure, temperature, and injection parameters. They may enhance the performance of the algorithm as they will make the model more sensitive to regime transitions and interventions. If such data were not provided in the data table used in modeling, this thesis uses only the key signals that have been collected and leaves the rest to be considered in the future.

3.2.3 Target Variable

Primary task: time-to-event (survival) modeling. Liquid loading is represented as an *event* that occurs on a specific day for each well. Let T_w denote the day index of the first observed loading event for well w , and let $\delta_w \in \{0, 1\}$ indicate whether an event is observed ($\delta_w = 1$) or the trajectory is right-censored ($\delta_w = 0$). In discrete-time form, we create a daily hazard label

$$y_{w,t} = \begin{cases} 1, & t = T_w \text{ (first loading day)} \\ 0, & t < T_w \text{ (pre-event day)} \end{cases}$$

and learn a model that outputs a hazard score $h_{w,t} = P(y_{w,t} = 1 \mid y_{w,0:t-1} = 0, \mathbf{x}_{w,t})$.

Early-warning horizons. For operational evaluation, we also consider horizon-based labels indicating whether the first loading event occurs within the next H days:

$$y_{w,t}^{(H)} = \mathbb{I}[0 < (T_w - t) \leq H], \quad H \in \{7, 14, 30\}.$$

This supports “risk within the next H days” monitoring, which is more actionable than a single next-day prediction.

Secondary task: severity classification. On event days, the dataset provides categorical severity labels (e.g., mild/moderate/severe). As a secondary experiment, we predict the severity class on event days only, using well-group cross-validation due to limited event counts.

3.2.4 Data Distribution and Exploratory Data Analysis

Exploratory Data Analysis (EDA) is done to (i) quantify the degree of imbalance, (ii) examine the distribution of feature values and detect outliers, (iii) examine the relationships among the features to help select and interpret features, and (iv) validate where the missing data appear.

Class imbalance. The source archive is highly imbalanced: loading-related days are rare relative to the full history of daily observations. In the broader consolidated archive, there are 118 positive loading labels among 61,060 daily records, corresponding to a positive rate of approximately 0.193%. This confirms that liquid-loading prediction is a rare-event problem in which naive accuracy would be misleading.

For the final experiments, however, the thesis uses a *survival cohort* rather than the full consolidated archive. That cohort contains 5,463 well-days from 30 wells, with 27 observed first events and 3 censored wells. The class imbalance is therefore still severe, but the target is now better aligned with the operational question of first-event risk rather than day-by-day binary status. Because the original exploratory figures are not central to the final survival analysis, the thesis emphasizes summary statistics and evaluation metrics over descriptive graphics.

Summary statistics (feature scale and spread). Table 3.1 Summary statistics for the non-constant numerical features are provided in the report. The report explains that the wellhead pressure varies over a wide range, from 0 to a maximum of 24.44 MPa, and that the daily gas rate is strongly right-skewed, with many days at very low values. This is consistent with the process of decline in the later years and

Table 3.1: Summary statistics for key numerical features (30 wells; 61,060 daily samples).

Feature	Mean	Std	Min	P25	P50	P75	Max
Wellhead pressure (MPa)	3.6272	3.8665	0.0000	1.1700	2.8200	3.5400	24.4400
Daily gas rate	0.8176	1.3356	0.0000	0.0000	0.5642	1.0471	18.4202
Cum. gas (10^4 m ³)	12.9141	22.7585	0.0000	1.2746	6.7682	16.4445	446.3530
Cum. water (m ³)	5.9089	11.2618	0.0000	0.0000	0.0000	8.0000	204.8000

intermittent operation. The cumulative values also have long tails, which reflect the heterogeneity in the age and productivity of the wells.

Correlation structure. A correlation analysis can be used to detect redundant predictors and to gain physical intuition. Two observations are notable:

- The daily gas rate has a moderate to strong positive correlation with cumulative gas, which makes sense: wells that produce more gas will tend to accumulate that production more rapidly over time.
- Wellhead pressure has a tendency to vary inversely with daily gas rates, though not exactly. This means that, although flowing pressure is low, deliverability could also be high, though not sufficiently to account for the change in loading behavior.

Since the process of liquid loading is a nonlinear and regime-dependent process, we use correlation as a screening tool, not as a means of proof.

Distribution shapes and scaling considerations. Distributions of feature values are also notable for their heavy tails and skew, particularly for rates and cumulative volumes. Log-transformed views of representative features show strong right-skewness, especially for gas-rate-like variables. These shapes motivate:

- Discussion of the scaling choices for the rate-like variables when using distance-based or linear methods.
- As opposed to the tree-based methods (random forest and gradient boosting) that are not significantly affected by monotonic transformations.

Outlier detection. Outliers may be caused by sensor issues, interruptions in operations, shut-ins, and intervention. The IQR method indicates that there are non-negligible fractions of outliers for important variables such as unusually high pressure rates when considering the overall distribution. For this thesis, by default, no removal of outliers is performed because it is possible that there are important signals (including those close to loading conditions). Three strategies are used to provide robustness: (i) robust models that handle outliers (such as tree ensembles), (ii) robust scaling for linear baselines, and (iii) time-forward validation to make sure that the discovered patterns are generalizable and not merely reacting to spikes.

Missing values. The consolidated data set used within this thesis contains no missing values within its numeric columns (missing rate = 0 for all used variables). This is less problematic than it would be for field data sets, where missing data and sensor outages are often major challenges. For reproducibility purposes, constant columns are removed, and other columns are standardized as required by the used model family.

Implications for modeling. The EDA also points out some modeling constraints and opportunities: For example, the EDA recommends precision-recall, F1, and average precision metrics for rare events, training with class imbalance, applying log or robust transformations to some baselines, and tree-based ensemble methods. The EDA also recommends cross-well validation, if possible. These findings directly influence the writing of the Methodology section and Experimental Results section.

Figure 3-1 shows how the final survival cohort is distributed across wells. Most wells contribute one first-event trajectory, while three wells remain censored throughout observation. The cohort is therefore small not only in terms of event count, but also in the number of independent well trajectories. Figure 3-2 provides an illustration of the trend of the standardized features averaged over the 30-day period leading up to the first occurrence of the labeled event. The curves are not perfectly deterministic, and this can account for the difficulty in discrimination despite high ROC-AUC scores even with the small hold-out dataset.

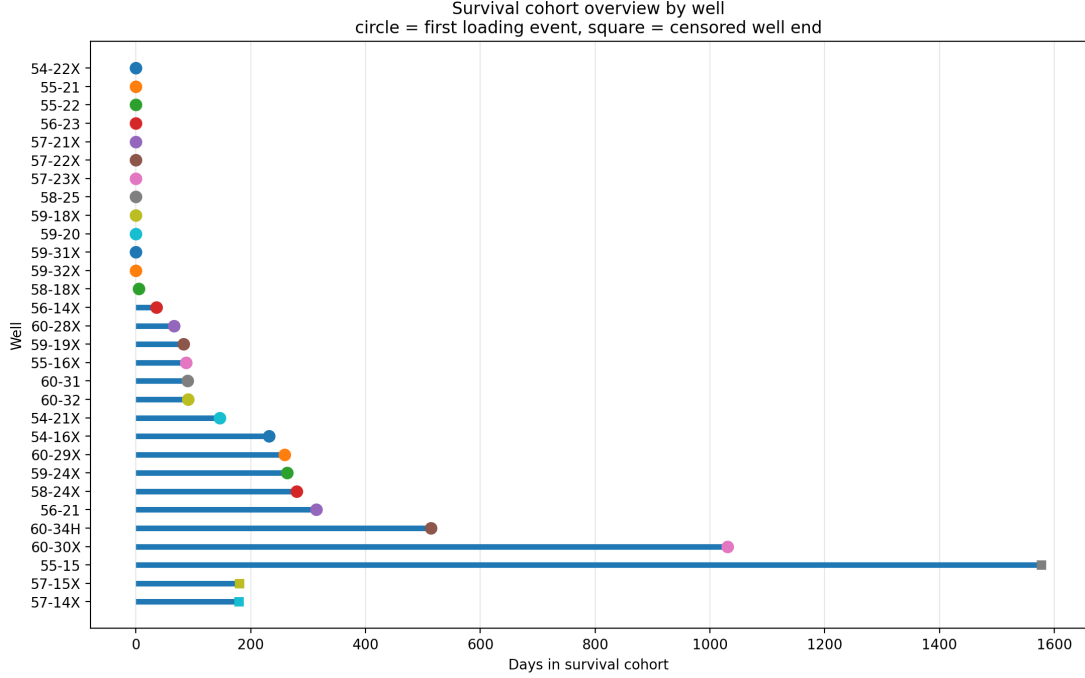


Figure 3-1: Well-based final survival cohort: bars represent actual duration of wells; circles represent the first loading event of wells; squares represent right-censored survival endpoints.

3.3 Methodology

This section describes the end-to-end method for modeling liquid loading as a *time-to-event* problem. The workflow comprises: (i) preprocessing and consistency checks on daily operational telemetry; (ii) feature engineering to capture physically meaningful relationships; (iii) construction of a time-to-first-event cohort from aligned daily measurements and engineer annotations; (iv) a leakage-resistant validation protocol based on holding out entire wells; and (v) evaluation using rare-event ranking metrics (PR-AUC/ROC-AUC), early-warning time-dependent AUC, and threshold-dependent metrics with out-of-fold threshold selection.

Figure 3-3 provides a visual summary of the complete methodology pipeline described in this section.

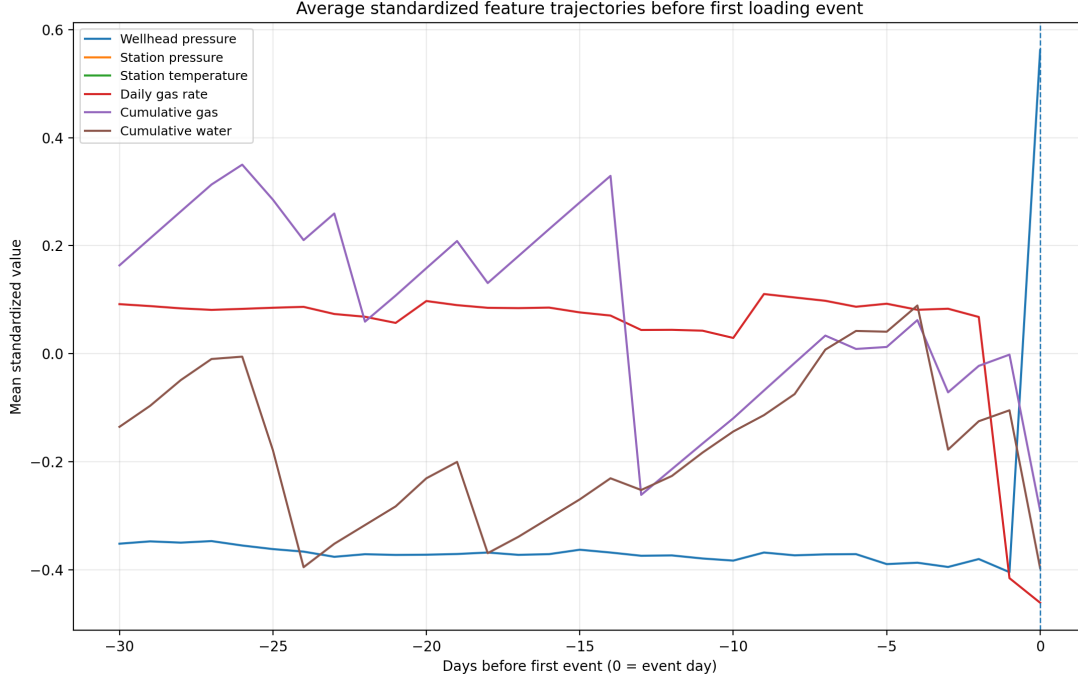


Figure 3-2: Average trend of standardized features for 30 days prior to first load; slow-moving weak trend changes indicate potential early warning but with noise.

3.3.1 Data Preprocessing

Field production datasets often arrive with messy formatting, sensor noise, missing data, and effects from shut-ins and interventions. The goal of preprocessing is to accomplish the following: (i) convert the raw log data into a standardized numerical format, (ii) handle missing data without exposing the model to information leaks, (iii) normalize the features properly depending on the family of models chosen, and (iv) handle outliers properly without losing important details in the data.

Cleaning and consistency checks. Let the raw dataset be a collection of well-level time series $\{(\mathbf{x}_{w,t}, y_{w,t})\}$ indexed by well w and time t . Cleaning steps include:

- **Date parsing and sorting.** Normalize the timestamps into a uniform date format and sort the data by (w, t) to preserve the correct ordering.
- **Numeric parsing.** Convert string representations of numbers to floating point numbers, handling locale-specific decimal marks and Excel serial date encodings as necessary.

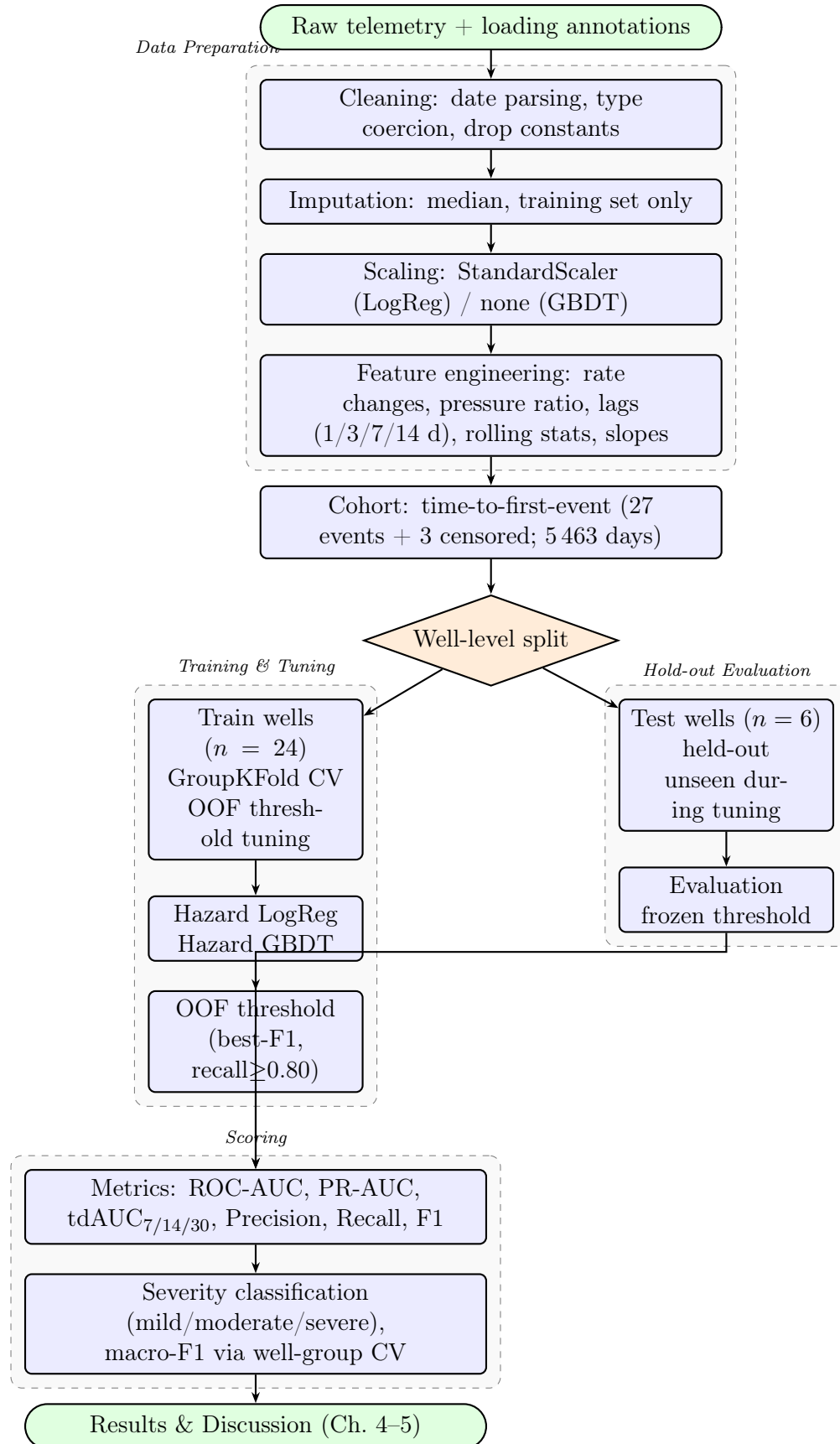


Figure 3-3: End-to-end methodology pipeline for liquid loading prediction.

- **Unit and range sanity checks.** Detect and fix any impossible numbers, for instance negative total production, with the use of knowledge about the problem domain.
- **Constant-column removal.** Remove features that have zero variance, as they cannot help with the prediction and can cause scaling problems.

In the combined set of data used for this thesis, there are no missing values for the main numerical features; however, constant columns are also removed, which means features with all zeros due to extraction issues.

Imputation of missing values. Missing data points are typically caused by sensor failures, communication losses, or data-integration issues. When missing values do occur, imputation must be based solely on the training data. The methods considered are:

- **Forward fill (time series).** If the feature is expected to change gradually, missing values at time t can be imputed by using the last observed value from time $t - \Delta t$.
- **Interpolation.** If the missing values are due to a gap in the trajectory, it is also possible to use linear interpolation over time, provided the gap is not too long so that the result does not become too smoothed.
- **Statistical imputation.** On the other hand, mean or median imputation is calculated from the training set and is often employed when temporal continuity cannot be trusted.

If $x_{i,j}$ is the value of the feature j for the sample i and the training set is denoted by $\mathcal{T}$, then the mean imputation is calculated by:

$$x_{i,j} \leftarrow \begin{cases} x_{i,j}, & \text{if observed,} \\ \mu_j, & \text{if missing,} \end{cases} \quad \mu_j = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} x_{i,j}. \quad (3.9)$$

In operational time-series settings, forward fill and interpolation are often preferable because they preserve within-well continuity more effectively than global-mean imputation.

Scaling and normalization. When distance-based or linear models are used, including logistic regression, SVMs, and neural networks, feature scaling is important. Tree-based models are much less sensitive to monotonic rescaling. Two common approaches are:

StandardScaler (z-score standardization).

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}, \quad (3.10)$$

where μ_j and σ_j are the mean and standard deviation computed from training data only.

MinMaxScaler (range normalization).

$$x'_{i,j} = \frac{x_{i,j} - \min(x_{\mathcal{T},j})}{\max(x_{\mathcal{T},j}) - \min(x_{\mathcal{T},j})}, \quad (3.11)$$

which maps values into $[0, 1]$ using training-set extrema.

Log transforms for skewed variables. Rate and cumulative variables are often right-skewed and have a large number of zeros. Monotone stabilizing transform is:

$$\tilde{x}_{i,j} = \log(1 + x_{i,j}), \quad (3.12)$$

This transform reduces the impact of extreme outliers and, in some cases, increases linear separability.

Outlier handling. Production data outliers may be caused by errors in sensors; however, they can sometimes represent real states of operations like shut-ins, slugs, unloading, or anything else. This is why they are treated with utmost care:

- **Detection:** Mark the extreme points according to the IQR-based criterion or robust z-score.

- **Treatment:** Do not remove them; rather employ robust techniques like transformation.

The outlier detection formula using the IQR criterion is as follows:

$$\text{IQR} = Q_{0.75} - Q_{0.25}, \quad [x_{\min}, x_{\max}] = [Q_{0.25} - 1.5\text{IQR}, Q_{0.75} + 1.5\text{IQR}], \quad (3.13)$$

where samples outside these thresholds will be marked. In the current thesis, those samples are kept by default after log transformation and robust tree ensembles, while winsorization can be used as an optional technique for other outlier-prone models like SVMs.

$$x_{i,j} \leftarrow \min(\max(x_{i,j}, x_{\min}), x_{\max}). \quad (3.14)$$

3.3.2 Pipeline Architecture

The architecture presented in Figure 3-4 encompasses all aspects implemented in the final pipeline. An important design decision was made to split the processing pipeline into two branches. One branch builds a discrete-time survival cohort and estimates a day-level hazard score for the *first* event in each well. Another branch runs when there are labeled events and performs classification of the severity of the event (either mild, moderate or severe). It should be noted that such architecture helps to avoid combining two different targets within one model.

In the hazard modeling, we started from per-well daily telemetry data and the engineer-annotated loading events. The initial records were aggregated into the first-event cohort where each well was represented by all the days before the first annotated event or all the available days if the well was censored. Once processed and leak-proofed features are extracted from each record, it becomes possible to train logistic regression and gradient-boosted trees on a day-level hazard target.

The evaluation logic is equally important. Thresholds are not tuned on the test wells. Instead, train wells are split with GroupKFold, out-of-fold hazard scores are collected, and an operating threshold is selected from those out-of-fold predictions. That threshold is then frozen and applied once to the six unseen wells. Figure 3-5

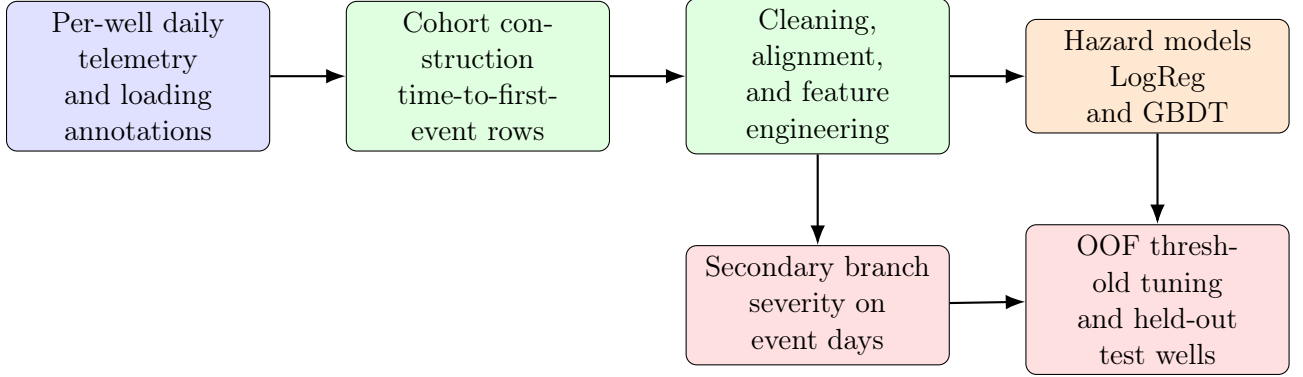


Figure 3-4: Overall thesis pipeline.

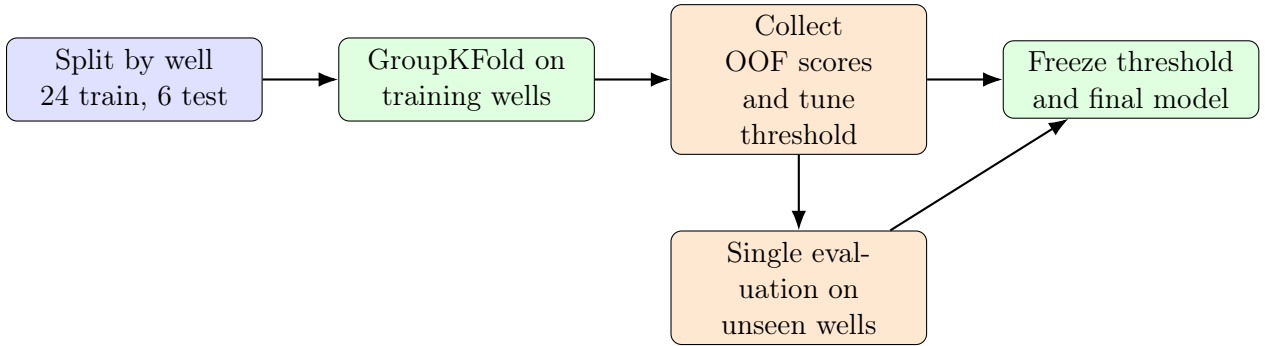


Figure 3-5: Leakage-resistant validation: thresholds are determined by training-well out-of-fold predictions alone and are thereafter fixed when assessing held-out wells.

visualizes this protocol, which prevents threshold leakage from the held-out wells into model selection.

3.3.3 Feature Engineering

The process of feature engineering involves applying one’s physical understanding of the concept of loading in liquids in terms of predictors that can be used in supervised learning algorithms. This thesis adopts a dual approach in constructing features to be considered in supervised learning based on the fact that loading tends to happen as a consequence of changes in the dynamics of pressure-production and gradual worsening of the state.

A logical place to start would be the history of the daily gas rate. Define q_t as the daily gas rate at time t . In order to quantify production decline, we can calculate the

first difference and the growth rate:

$$\Delta q_t = q_t - q_{t-1}, \quad r_t = \frac{q_t - q_{t-1}}{q_{t-1} + \epsilon}, \quad (3.15)$$

where ϵ represents a small constant, added to alleviate problems associated with numerical instability as the denominator tends to approach zero. Constant negative values for these variables could indicate well degradation and could therefore be considered reasonable early indicators of increasing load risk.

Aside from production characteristics, we also establish simple pressure-rate analogs, intended to capture the changing deliverability situation. If P_t is the calculated pressure proxy at the wellhead/surface level, then we have:

$$\phi_t = \frac{P_t}{q_t + \epsilon}. \quad (3.16)$$

While this is not an equation that mechanistically models flow, this is a good way of summarizing the pressure associated with a particular production rate. Higher numbers for ϕ_t might suggest lower deliverability or increased effective backpressure.

The history of a well's production can also provide clues about its maturity and loading characteristics. If we let G_t be the cumulative gas production and W_t be the cumulative water production, then their daily change can be defined as:

$$\Delta G_t = G_t - G_{t-1}, \quad \Delta W_t = W_t - W_{t-1}. \quad (3.17)$$

They help distinguish between long-term production build-up and short-term changes in production behavior. In practice, the cumulative variable acts as an index of maturity, while its change gives us a local insight into changing gas and water production dynamics.

When both pressures are known in the tubing and casing, new physically relevant pressure differences are formed.

$$\Delta P_t = P_{\text{casing},t} - P_{\text{tubing},t}, \quad (3.18)$$

These measured quantities can be indicative of changes in fluid flow regime, multi-phase flow dynamics, and liquid holdup within the wellbore itself. Where estimates of downhole pressures are available, the above-mentioned variables can be extended to estimate hydrostatic terms by employing a breakdown similar to Equation (ref: dpdz). In this study, however, the available dataset only includes a surrogate measure for wellhead pressure which is consistently available. Therefore, these differential and gradient-derived quantities are considered as extensible variables for future data sets and not fundamental variables for this analysis.

Since liquid loading is usually a progressive process and not an immediate one, temporal information is introduced using lagged values and moving statistics. In general, where x_t represents a variable, the lagged measures are calculated as:

$$x_t^{(k)} = x_{t-k}, \quad k \in \{1, 3, 7, 14\}, \quad (3.19)$$

To facilitate the use of short and medium-term history by the model, rolling statistics have been used to provide an indication of recent trends:

$$\bar{x}_t^{(w)} = \frac{1}{w} \sum_{i=0}^{w-1} x_{t-i}, \quad s_t^{(w)} = \sqrt{\frac{1}{w} \sum_{i=0}^{w-1} (x_{t-i} - \bar{x}_t^{(w)})^2}. \quad (3.20)$$

A window size of 7 days, 14 days, and 30 days has been chosen as the average period for monitoring operations. The rolling mean will capture any short term level shift, whereas the rolling standard deviation will capture any rising level of instability.

In order to make directional changes clearer, slopes for linear trends are calculated using moving averages:

$$\beta_t^{(w)} = \arg \min_{\beta} \sum_{i=0}^{w-1} (x_{t-i} - (\alpha + \beta i))^2, \quad (3.21)$$

where $\beta_t^{(w)}$ is the slope coefficient within a window of size w . If there are negative slopes on gas rate-related variables or positive slopes on pressure-related variables, it might imply progressive degradation.

Lastly, ratio variables are used to incorporate information that can be obtained from ratios rather than absolute numbers.

$$\psi_t = \frac{\Delta W_t}{\Delta G_t + \epsilon}, \quad \eta_t = \frac{W_t}{G_t + \epsilon}, \quad (3.22)$$

The interaction term, $P_t \cdot q_t$, is useful, especially in linear models where it is not inherent to learn nonlinearities. For tree-based models, interactions are often implicitly accounted for via the split structure and therefore are less necessary to include explicitly.

Overall, the feature engineering procedure utilizes physical intuition regarding proxy variables together with summaries of the time series data, which allows using regular machine learning models.

3.3.4 Train–Test Split and Validation

The generalization performance is measured using a more realistic deployment scenario where the data is partitioned on a per-well basis rather than on a per-day basis. In particular, the partition strategy employed here partitions the data using a well-based approach, reserving a subset of wells for validation (24 wells used for training and 6 wells reserved for testing). The order of events in time is maintained for each well when forming cohorts and constructing features, but the most prominent way of controlling leakage in the end is through the well-level holdout method.

Threshold selection is treated as part of the modeling pipeline. To prevent optimistic bias, decision thresholds are tuned using *out-of-fold* predictions on the training wells via GroupKFold cross-validation, and the selected threshold is then applied once to the held-out test wells.

In order to improve reproducibility, Table 3.2 describes the specific pre-processing pipeline, class handling mechanism, random seed numbers, and model architecture used during implementation, instead of just describing general model families.

This yields operating points that are more representative of real surveillance deployment under extreme class imbalance.

3.3.5 Evaluation Metrics

The metric framework is designed such that it would stay informative in the face of high class imbalance and would take into account the role of the model as a monitoring or warning system, rather than a classifier. In turn, along with a horizon-based performance metric for predicting lead-time utility, the thesis provides the results of classification and ranking performance metrics.

Denoting TP , FP , TN , and FN as the number of true positives, false positives, true negatives, and false negatives, respectively, under fixed threshold τ the threshold dependent measures of classification performance are defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Such evaluation metrics are important since they correspond to a specific decision rule: an alert is raised whenever a score is greater than the threshold. Precision refers to the proportion of alerts that are correctly identified, recall refers to the proportion of true events detected, and the F1-score combines both metrics. Since such metrics are dependent on the threshold value, the threshold value is calculated using out-of-fold predictions obtained on the training well group, and the resulting point on the ROC curve is determined in advance before evaluation of the model on the testing well group.

In addition to the aforementioned threshold-dependent metrics, threshold-independent ranking-based metrics are also considered. The ROC-AUC metric represents the trade-off between the true and false positive rates. However, the ROC-AUC metric may become somewhat inflated when used for an unbalanced problem since the classifier performance with regard to the predominant negative class affects this metric. Therefore, precision-recall area under the curve (PR-AUC) is also reported. PR-AUC may provide a more adequate representation of the algorithm’s predictive capabilities in detecting rare events due to its focus on the target minority class.

Since the goal involves both the identification and prediction of events prior to their occurrence, making the system operationally effective, a measure involv-

ing early warning within a timeframe is also included. For early warning periods $H \in \{7, 14, 30\}$, the labels were created from pre-event days as

$$y_{w,t}^{(H)} = \mathbb{I}[0 < (T_w - t) \leq H],$$

T_w represents the event time at well w , and $\mathbb{I}[\cdot]$ stands for the indicator function. The specification uses the observations that happen within the interval of H days before the event as positives. For each time horizon used, the time-dependent area under the receiver operating characteristic curve (AUC) is calculated in order to evaluate if the risk score assigned to observations that are closer to a future event is greater than those assigned to non-event observations. Greater time horizons indicate early warnings, which can be harder or easier to predict, based on the nature of the data. In general, it is useful to evaluate how well the model predicts at various time horizons to know if it reacts only when there is an impending failure.

These measures provide a comprehensive set of measures for discriminating ability, operating point characteristics, and early warnings. This is especially important for this thesis work because high discrimination is not sufficient to produce an effective alarm scheme, especially in situations of extreme class imbalance and multiple wells.

3.3.6 Summary

This process involves conservative preprocessing, feature extraction based on physical insight, well-level testing, and rare-event scoring criteria. Rather than optimizing one metric in isolation, the goal is to compare real-world baseline models in conditions that replicate the intended surveillance scenario.

3.4 Model Development and Validation

In this section, the employed predictive algorithms, tuning schemes, and validation protocols are outlined in detail. The chosen classes of models conform to established guidelines in machine learning literature for classification and ensemble methods [19,

20], whereas their application to liquid loading problems stems from more recent petroleum engineering research [12, 14, 17].

3.4.1 Models Implemented

Let $\mathbf{x}_i \in \mathbb{R}^d$ be the feature vector for the i -th sample and $y_i \in \{0, 1\}$ be the binary label indicating liquid loading. Each model produces a score $s(\mathbf{x}_i)$ that is transformed into an estimated probability $\hat{p}_i = P(y_i = 1 \mid \mathbf{x}_i)$. A classification threshold τ yields a predicted label $\hat{y}_i = \mathbb{I}[\hat{p}_i \geq \tau]$. Because positives are extremely rare, the modeling and validation focus on probability ranking quality and threshold-stable metrics rather than accuracy. .

Logistic Regression (baseline). Logistic regression provides a clear and interpretable baseline and a well-calibrated scoring function.

$$\hat{p}(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + \exp(-(\mathbf{w}^\top \mathbf{x} + b))}. \quad (3.23)$$

Parameters $\mathbf{w}, b$ are learned by minimizing a regularized cross-entropy objective:

$$\min_{\mathbf{w}, b} \sum_{i=1}^N \ell_{\text{BCE}}(y_i, \hat{p}_i) + \lambda \|\mathbf{w}\|_2^2, \quad (3.24)$$

The model uses the standard setup with λ controlling the strength of the L2 regularization term [20]. In the case of imbalanced classes, class-weighted cross-entropy is used. The role of logistic regression here is to provide a stable baseline against which nonlinear models can be judged.

Strengths. Fast training, strong baseline performance, high interpretability through coefficient sign and magnitude, and good calibration.

Limitations. However, it has a tendency to underfit when the decision boundary is highly nonlinear, involves interactions, and has regime changes, as is common for multiphase well behavior [17].

Random Forest. Random Forest (RF) is an ensemble of bagged decision trees that can handle nonlinearities and interactions without requiring complex preprocessing.

A forest of M decision trees gives:

$$\hat{p}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \hat{p}_m(\mathbf{x}), \quad (3.25)$$

Each tree is based on a bootstrap sample, with split variables chosen from a random subset of features to minimize inter-tree correlation.

Reasoning for using RF for liquid loading. Random Forests can handle unbalanced features and mild outlier issues, can handle features at different scales without heavy normalization, and provide features with importance. These features make RF particularly attractive in field settings where interpretability and robustness matter. RF is often used in many reviews as a good classical baseline for liquid loading and well event detection[17].

Imbalance handling. Random Forest also allows the use of class balancing using parameters such as `class_weight = 'balanced'`, or by balancing samples in the training process. This feature is beneficial when we are dealing with rare event cases since it helps avoid the model’s bias towards the larger class.

Gradient Boosting (XGBoost/LightGBM). Gradient Boosting Decision Trees (GBDT) are an ensemble of trees that grow iteratively to fit the residual errors, which results in a strong learner for tabular data. The boosted model can be expressed as:

$$f_K(\mathbf{x}) = \sum_{k=1}^K \eta h_k(\mathbf{x}), \quad \hat{p}(\mathbf{x}) = \sigma(f_K(\mathbf{x})), \quad (3.26)$$

where h_k is the k -th tree, η is the learning rate, and K is the number of boosting rounds. Practical implementations include XGBoost and LightGBM [23, 24].

Why GBDT is expected to perform strongly. According to recent studies on the topic of liquid loading and its relation to regression and classification, the performance of boosted algorithms is adequate if they are properly calibrated and regularized [12]. In particular, boosted decision trees can represent non-linear interactions between variables, as well as monotonic constraints if necessary.

Imbalance handling. GBDT supports scale-pos-weight (like XGBoost), class weights,

and custom loss functions. In handling rare events, it is important to carefully tune tree depth, minimum leaf size, and subsampling to avoid fitting noise.

Neural network (optional). Lastly, we also consider a simple feed-forward neural network as a nonlinear benchmark, MLP:

$$\mathbf{h}_1 = \phi(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1), \quad \mathbf{h}_2 = \phi(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2), \quad \hat{p} = \sigma(\mathbf{w}^\top \mathbf{h}_2 + b), \quad (3.27)$$

with nonlinearity $\phi(\cdot)$ (e.g., ReLU). The regularization through dropout and early stopping is critical in dealing with imbalance and nonstationarity.

When neural models matter. Modeling the temporal sequences directly, such as through LSTMs or attention, tends to perform better, as recent work indicates that such techniques can learn precursor patterns at the onset of loading[14]. On purely tabular data with a small feature set, tree ensemble models often perform as well as or better than MLPs.

3.4.2 Hyperparameter Tuning

Hyperparameter tuning is the process of selecting the model configurations that achieve good performance on the validation data with the awareness of time. Since the data is imbalanced, the hyperparameter tuning is done using appropriate metrics for rare event detection problems, such as precision-recall, average precision, and recall at certain false alarm rates, rather than using accuracy as a metric.

Search strategy: Grid Search (and practical variants). **Grid Search.** During hyperparameter tuning, we begin with a set of candidate hyperparameters given as $\Theta = \Theta_1 \times \cdots \times \Theta_m$. Grid search involves training as well as evaluating the model over all possible values of the hyperparameters given as $\theta \in \Theta$. It's a thorough method, though at times expensive. For boosted trees, we have employed a tuned grid over the hyperparameters with the most significant impact.

Practical variant. In certain situations, random search can be highly efficient in a very large space. However, in thesis, we have preferred to stick with grid search to

enhance reproducibility as well as transparency. We have employed limited parameter ranges as well as early stopping in certain situations, such as in GBDT.

Cross-validation protocol for tuning. Standard random k -fold CV is leakage-prone for time series. Therefore, it is not applicable in this case. The hyperparameters are selected using either:

- **Forward-chaining CV within wells:** train on earlier and validate on later segments.
- **Blocked time splits:** validate on contiguous time blocks.

In cases where it is possible, we follow a careful evaluation as an additional generalization check, which corresponds to the cross-well robustness issues raised in the literature [17].

Representative tuned hyperparameters. Table 3.3 is a summary of the representative hyperparameters that were tuned for each family of models. The ranges were chosen to limit overfitting and be computationally tractable.

3.4.3 Model Complexity Analysis and Diagnostics

Other than the large numbers, validation also involves ensuring that the model is interpretable and that there is no overfitting. This is particularly important for predicting liquid loading, where false positives can result in costly actions and where relationships may be artifacts of system operation rather than physical effects.

Feature importance (tree ensembles). Tree-based models provide impurity-based feature importance, which includes RF and GBDT. This is a measure of the decrease in impurity over the ensemble for each feature. They are efficient and convenient, but they can be biased towards continuous features with many possible split points. We therefore use impurity-based feature importance qualitatively but rely on SHAP values.

A representative feature-importance plot is shown in Fig. 3-6.

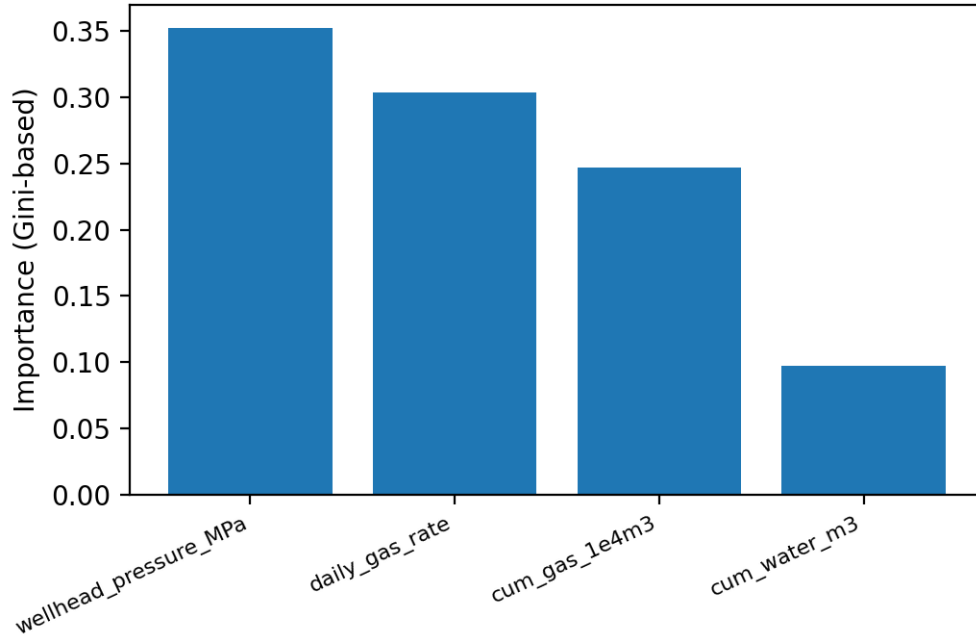


Figure 3-6: Importance of features in tree-based methods: higher values mean greater contribution to reducing impurity.

SHAP explanations (optional). To increase interpretability and reduce bias from impurity-based importance, SHAP (Shapley Additive Explanations) can be used to decompose a prediction into feature contributions [25]:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^d \phi_j, \quad (3.28)$$

Here ϕ_j is the contribution of feature j to the final prediction. SHAP is useful in operational review because it allows engineers to inspect whether a high-risk day is being driven by physically plausible signals such as declining gas rate, rising pressure proxies, or unusual variability.

Overfitting diagnostics. Overfitting is detected using three different checks:

(1) **Train–validation divergence.** The first check is the large discrepancy between the performance on the training data and the forward-chaining validation performance, which indicates overfitting or memorization. The first check uses early stopping in the case of boosted trees, observing the validation loss and stopping the

boosting process before overfitting occurs.

(2) Stability across folds and wells. The second check is the performance variation over time blocks or across wells, indicating the model may be overfitting to the wells’ particular characteristics. Because of the generalization difficulties across wells and operating regimes [17], stability is an important factor when choosing the model.

(3) Probability calibration and threshold sensitivity. The third check is the performance on the validation metrics, even when the metrics are good, indicating the model’s confidence may not be meaningful. The calibration curve or reliability diagram and sensitivity analysis using precision/recall with varying thresholds on tau are useful checks. Post-calibration methods such as Platt scaling and isotonic regression may be employed if necessary on the validation set.

3.4.4 Validation Summary and Model Selection Criteria

Model selection is based on:

- **Rare-event performance:** It is all about accurate performance in terms of the ability of the model to recall the correct signals and the false alarm rate being acceptable.
- **Generalization:** It should be consistent in terms of time forward validation and, wherever possible, across multiple wells.
- **Interpretability:** It should be possible to justify the alerts based on the importance of the features or SHAP values.
- **Operational simplicity:** It should be robust to changes in the data distribution and should not need extensive tuning or be difficult to deploy.

This is all in line with the more recent research that suggests boosted trees are good quantitative tabular models [12], while deep temporal models show promise for identifying precursor patterns [17].

Table 3.2: Final model specification used in the experiments (reproducibility checklist).

Specification item	Value used
Primary task	Time-to-event hazard model for time to first loading event. Every well has observations starting at the first observation time point until the day that the first loading event is observed; wells without any event occurrences are considered right-censored.
Secondary task	Classification of event severity on event days only based on severity (mild / moderate / severe).
Data sources	Per-well Excel files with two sheets: (i) daily measurements (ii) loading annotations
Cohort size (final)	30 wells; 5,463 daily cohort rows; 27 first-event wells; 3 censored wells (no event).
Features (final)	6 numeric predictors: wellhead pressure , daily gas rate , cumulative gas, cumulative water, daily methanol injection , auxiliary pressure-like channel.
Window sizes / horizons	Lagged feature windowing is not included in the hazard baseline, which relies on day-level features only. For early warning evaluation, horizon label H7, 14, 30 days is used, as indicated by time-dependent AUC. Risk ramp-up chart is generated by a 30-day window prior to an event.
Preprocessing	In terms of data preprocessing, numeric coercion is used; feature columns that have missing value or constant value are dropped (based on training data only). Imputation for missing values is done via median(SimpleImputer(strategy=median)).
Scaling / transforms	Logistic Regression (Hazard LogReg): Median imputation together with StandardScaler. GBDT tree-based models: Median imputation alone without any scaling. Log-transformations were not considered at all.
Train-test split	Well-held-out split (GroupShuffleSplit by <code>well_id</code>): 24 training wells, 6 unseen test wells.
Threshold selection policy	The optimal decision thresholds were chosen using out-of-fold predictions based on the GroupKFold cross-validation of training wells to avoid optimistic estimates and then used for test wells. Two decision rules were obtained: (i) OOF best-F1 and (ii) constrained by recall (recall 0.80).
Evaluation metrics	Hazard prioritization: ROC-AUC and PR-AUC calculated on out-of-sample wells. Early warning system accuracy: time-varying AUC for $H = \{7, 14, 30\}$. Threshold-based measures: precision, recall, F1-score, and confusion matrix at threshold levels obtained from OOF tuning. Severity analysis: macro F1-score using cross-validation between well groups; confusion matrix for out-of-sample wells, where possible
Model 1: Hazard LogReg	Pipeline: <code>SimpleImputer(median)</code> $\rightarrow$ <code>StandardScaler</code> $\rightarrow$ <code>LogisticRegression</code> . Parameters: <code>max_iter=2000</code> , <code>solver=liblinear</code> , <code>class_weight=balanced</code> .
Model 2: Hazard GBDT (final best)	<code>HistGradientBoostingClassifier</code> with <code>max_depth=4</code> , <code>learning_rate=0.05</code> , <code>max_iter=250</code> , <code>min_samples_leaf=50</code> , <code>random_state=42</code> .
Secondary model: Severity GBDT	Pipeline: <code>SimpleImputer(median)</code> $\rightarrow$ <code>HistGradientBoostingClassifier</code> with <code>max_depth=3</code> , <code>learning_rate=0.08</code> , <code>max_iter=200</code> , <code>random_state=42</code> .
Seeds / determinism	<code>random_state=42</code> used for the well split and tree models.
Training stopping criteria	LogReg stops at <code>max_iter=2000</code> (convergence-based). GBDT training stops at <code>max_iter</code> (250 for hazard; 200 for severity).

Table 3.3: Models implemented and hyperparameter values used in the final experiments.

Model	Hyperparameter values used
Logistic Regression	Median imputation; feature scaling with StandardScaler; inverse regularization strength $C = 1.0$; penalty = L2; solver = liblinear; class_weight = balanced; max_iter = 2000.
Gradient Boosting (hazard model)	HistGradientBoostingClassifier with learning rate $\eta = 0.05$; maximum depth = 4; number of boosting iterations = 250; min_samples_leaf = 50; random_state = 42.
Gradient Boosting (severity model)	HistGradientBoostingClassifier with learning rate $\eta = 0.08$; maximum depth = 3; number of boosting iterations = 200; random_state = 42; median imputation applied before training.

Chapter 4

Results

4.1 Results

This section reports results for liquid loading surveillance under a well-held-out evaluation protocol. The primary task is discrete-time hazard modeling for time-to-first-event prediction. Models are trained on 24 wells and evaluated on 6 unseen wells. Operating thresholds are selected using out-of-fold (OOF) predictions on training wells to avoid optimistic bias.

4.1.1 Hazard prediction on unseen wells

Table 4.1 reports ranking and early-warning metrics on the held-out wells. Gradient-boosted decision trees (GBDT) provide the strongest ranking signal (ROC-AUC = 0.929, PR-AUC = 0.145). Time-dependent AUC increases with horizon for GBDT (0.458 at 7 days, 0.530 at 14 days, 0.566 at 30 days), which is consistent with the operational expectation that longer lead windows are easier to discriminate than immediate next-day alarms.

4.1.2 Operating point on unseen wells (OOF-tuned threshold)

Figure 4-1 depicts the ROC curves for both the hazard models using the six validation wells. The GBDT outperforms the logistic regression method in nearly all operating

Table 4.1: Hazard prediction performance on unseen wells (6 wells). Ranking metrics are computed from test probabilities; time-dependent AUC uses horizon labels for $H \in \{7, 14, 30\}$.

Model	ROC-AUC	PR-AUC	tdAUC ₇	tdAUC ₁₄	tdAUC ₃₀
Hazard LogReg	0.857	0.118	0.387	0.336	0.312
Hazard GBDT	0.929	0.145	0.458	0.530	0.566

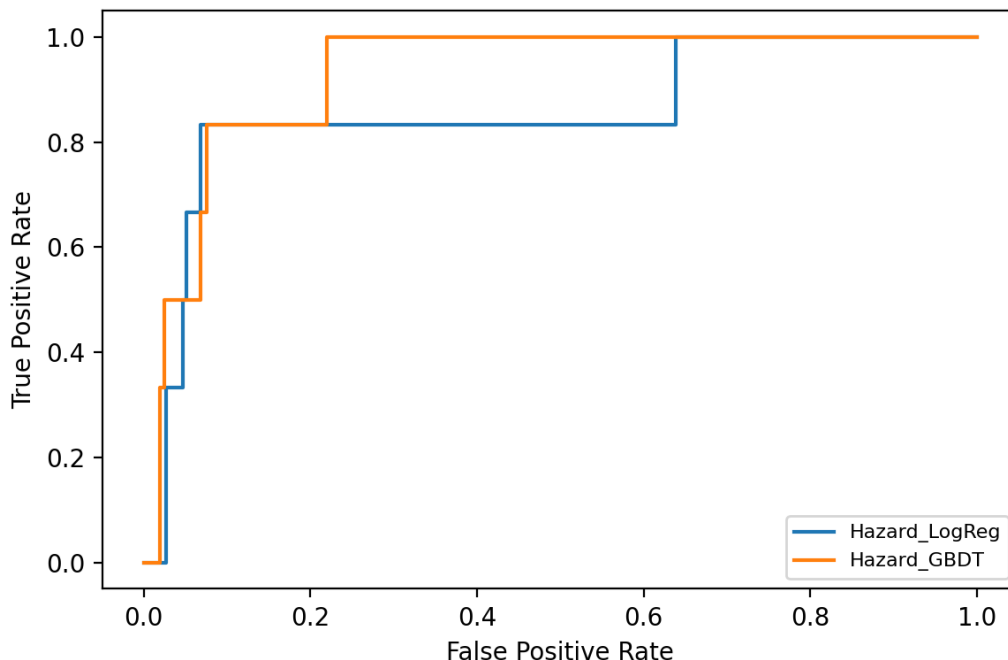


Figure 4-1: ROC curves on the held-out wells for the two hazard models. GBDT dominates logistic regression across most operating regions, consistent with the higher test ROC-AUC.

points, in line with its higher test ROC-AUC value of 0.929 compared to 0.857 of the logistic regression. Importantly, the GBDT curve shows an especially steep slope for low values of the FPR, meaning that many true positives could be captured before reaching an unacceptable number of false positives, making it valuable for monitoring operations.

Figure 4-2 depicts the precision-recall curves for both evaluation runs. Since the positive category consists of less than 0.2% of all days, it makes the PR curve a much more informative measure of model performance than ROC: while a model always scoring days as negative will have an ROC-AUC close to perfect, it will plot a horizontal line near zero on the PR plot. For both models, the PR plots demonstrate

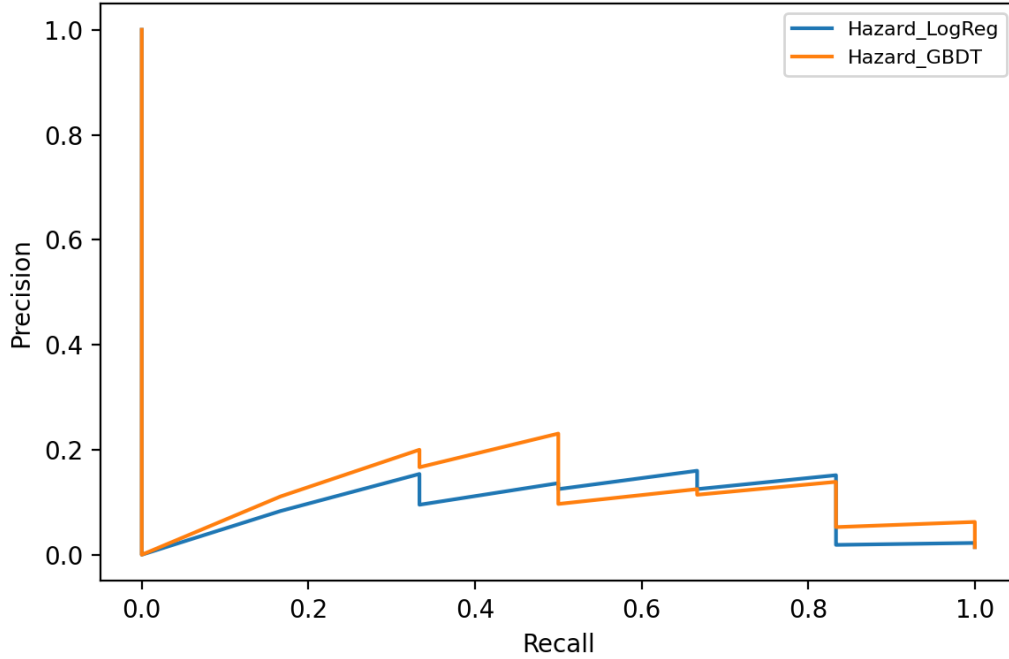


Figure 4-2: PR Curves - held out: rare events favor PR-AUC over accuracy measure because GBDT is a good predictor but precision can be improved.

a typical pattern of rare event detectors, where precision stays low at any level of recall, although the GBDT model outperforms in PR-AUC (0.145 vs. 0.118).

Using the OOF tuned best F1 threshold obtained during training on the training wells, work can obtain the following results for the unseen wells: precision = 0.116 and recall = 0.833 for the GBDT hazard model (F1 = 0.204). The corresponding confusion matrix is:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} = \begin{bmatrix} 371 & 38 \\ 1 & 5 \end{bmatrix}.$$

That is, 5 out of 6 first loading events are detected, at the cost of 38 false alarms over 415 test days — approximately one spurious alert every 11 days. In terms of real-life applicability, the risk scores can be used for daily well prioritisation, but the false-alarm rate remains too high for unsupervised automated alerting. The corresponding confusion matrix is:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} = \begin{bmatrix} 371 & 38 \\ 1 & 5 \end{bmatrix}.$$

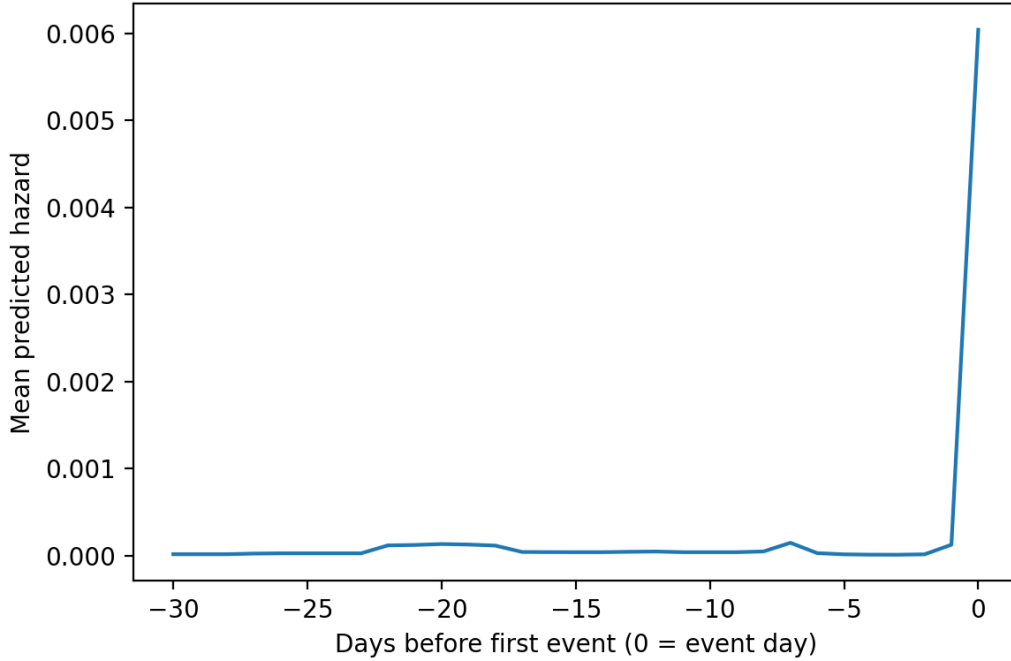


Figure 4-3: Pre-event mean hazard prediction: gradual ramping up before the first event is responsible for better discrimination over a longer window compared to instantaneous hazard timing.

That is, out of 6 initial loading events, 5 will be predicted, resulting in 38 false alarms in 415 days. In terms of real-life applicability, the model can be used for prioritizing the wells on the basis of their risk levels, but it cannot yet be used to alert automatically.

4.1.3 Secondary task: severity classification

As a secondary experiment, we predict loading severity (mild/moderate/severe) on event days only. Due to limited event counts, stability is assessed using well-group cross-validation. The severity model achieves $\text{macro-F1} = 0.48 \pm 0.27$ across 5 folds, while a small held-out subset of test wells yields $\text{macro-F1} = 0.40$ ($n = 3$ event days). The confusion pattern is shown in Fig. 4-5.

This implies that severity prediction can only serve as an exploratory secondary objective within the framework of this thesis work. This is due to the small size of the samples used in the analysis, whereby severity prediction can only indicate possibility

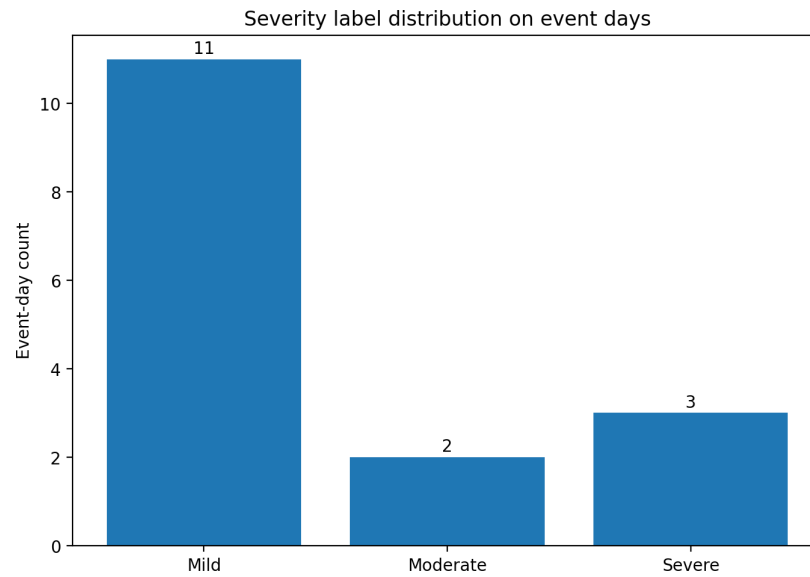


Figure 4-4: Severity distribution on event day: unevenly distributed classes – particularly moderate severity – affect macro-F1 stability across folds.

and not readiness.

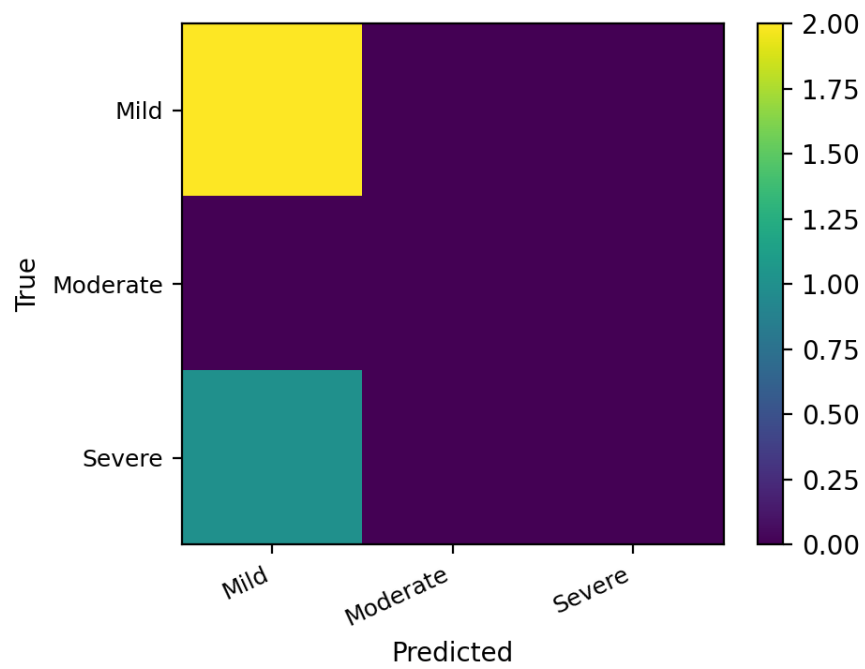


Figure 4-5: Confusion Matrix for Severity: The classification of mild severity is more accurate compared to moderate and severe severity because there are fewer labels for event days.

Chapter 5

Discussion

5.1 Interpretation of the Results

As is evident from Table 4.1, the GBDT model is able to learn a signal in the operational telemetry data indicating the risk of liquid loading, yet its accuracy is not sufficiently high to use it for automatic detection.

Hazard ranking. According to the results presented in Table 4.1, the GBDT model is superior to logistic regression with ROC-AUC and PR-AUC equal to 0.929 and 0.145 versus 0.857 and 0.118, respectively. In addition, the difference in accuracy metrics of the two models can be observed from their ROC and precision–recall curves plotted in Figures 4-1 and 4-2, respectively. Due to the small fraction of days with liquid loading (less than 0.2% per day), one might expect the difference in ROC-AUC and PR-AUC of the two models, which makes the latter metric more informative.

Operating point. Using the optimal OOF threshold, the GBDT hazard model identifies 5 out of 6 first loading episodes on unseen wells for which recall is estimated as $= 0.833$. The penalty of such an approach is the generation of 38 false alarms across 415 testing days, leading to precision $= 0.116$ and F1-score $= 0.204$. The confusion matrix includes 371 true negatives, 38 false positives, 1 false negative, and 5 true positives. From a practical perspective, it means that the model triggers approximately one wrong alarm every 11 days. At the same time, this level of false positives prevents automatic system operation, yet this scoring method can be utilized

within the human-in-the-loop framework for daily well prioritization.

Early-warning horizon. The findings based on the horizon-based approach presented in Table 4.1 reveal that the time-dependent AUC metric rises with increasing length of the lead window: $\text{tdAUC}_7 = 0.458$, $\text{tdAUC}_{14} = 0.530$, and $\text{tdAUC}_{30} = 0.566$ for GBDT. Such behavior is consistent with the mean risk trajectories depicted in Figure 4-3. It is evident

Severity classification. The secondary severity model obtains macro-F1 score $= 0.48 \pm 0.27$ across five folds, with macro-F1 score $= 0.40$ on the small test subset ($n = 3$ event days). As shown in Figure 4-4, mild events comprise the bulk of the severity labels, whereas moderate and severe events are rare and insufficiently frequent for stable training. The confusion matrix in Figure 4-5 further demonstrates that the model exhibits a bias toward the majority class. As such, severity classification must be considered exploratory in nature, and not ready for production use.

Overall. Overall, the experiments reveal that operational telemetry contains valuable signals for risk assessment: it is sufficient for risk stratification but insufficient for automated alert generation. The primary limitation is the lack of data rather than a specific model type, given that the final cohort size is limited to 27 first events out of 30 wells, 6 of which can be used for testing. Increasing the accuracy of alarm prediction without losing recall is the critical path forward.

Potential Directions for Performance Improvement

Adopting additional features from the source dataset. Current system utilizes only a few process signals: wellhead pressure, daily gas flow rate, cumulative gas production, cumulative water production, daily methanol injection, and an additional pressure-like signal. The source data from Wang et al. [18] provides more signals such as tubing pressure, casing pressure, wellhead temperature, and injection rates that are missing in the integrated modeling file employed for this thesis. Using these extra signals would overcome one of the main shortcomings of the pipeline, as discussed in Section 5.2, which concerns the lack of pressure data and completion details. Among all these, tubing-casing pressure difference, $\Delta P_t = P_{\text{casing},t} - P_{\text{tubing},t}$, is of physical

relevance because it can represent flow regime shift, back-pressure, and liquid hold-up.

Systematic analysis of raw versus engineered features. A number of engineered features are suggested by the methodology, including rate changes, pressure-rate ratios, production ratio, and rolling features. Nonetheless, the model hazard baseline presented in Table 3.2 is primarily based on daily level information. A future ablation test is required comparing the raw features with engineered features to understand what inputs are driving better predictions. SHAP value analysis [25] or permutation importance tests can help assess whether features like gas-rate change, water/gas ratio, and pressure-rate interaction are actually adding early warning signals or not. This will result in a simpler and deployable set of features.

Synthetic data generation for rare-event augmentation. The major challenge faced by our model is the limited number of labeled first loading events: there are only 27 loading events in 30 wells. This poses a huge limitation in developing an efficient rare-event classifier. One possible approach is the use of synthesized loading trajectories, which can be derived from physics-based simulation models. The simulation may be done using the mass balance principle in Equation 3.5 such that the liquid holdup increases if the liquid input is greater than the liquid removal rate. The synthetic daily telemetry trajectories can be used to train the model or conduct sensitivity analysis before evaluating its performance on held-out wells.

Another method is augmentation using interpolation between actual trajectories of wells, although the augmented data must still make physical sense. Any form of synthesis of data will need to be carefully validated to avoid introducing bias to the model. Nonetheless, since the present PR-AUC and F1 measures are highly constrained by the limited events, augmentation of rare events is a promising avenue for improvement of the alarm system.

5.2 Limitations

The first limitation lies in the data itself, which can be influenced by more than just the reservoir response. It can be influenced by things like interventions, sensor

inaccuracies, transmission errors, shut-ins, and operational strategy changes. With daily sampling, one may lose information about flow instabilities that come and go quickly, and the observations will not distinguish well between the precursors to an action taken versus the effects of the action itself. The second limitation is related to the frequency of the events. After the creation of our cohort of wells, we have only 27 total first events, including only 6 on the holdout wells, which means that each true positive or false negative makes a material difference. The third data-related limitation is incompleteness of observations. Important process information like tubing configuration, composition, bottom hole conditions, and intervention history are unknown and unavailable to us.

Modeling constraints exist as well. Inter-well generalization becomes the central challenge since the wells can vary in their completion design, maturity, proclivity to produce water, and operational policies. A model that performs well on one batch of wells will have a degree of dependence on well-level signals instead of universal signals for precursors. This issue is further complicated by concept drift: as the wells get older and technologies change, the correlation between telemetry data and the labels could be different. The second significant restriction is the threshold dependency. In rare event monitoring, ranking and operational efficiency cannot be considered equivalent: while the GBDT model ranks well, the threshold at which it can identify 5 out of 6 test events creates 38 false positives over 415 days of testing.

Lastly, there are limitations to implementation. In practice, a working surveillance system would have to ingest information reliably, create features automatically, detect drift, implement versioning for the models, and define the process of human intervention. The output from the model itself would have to be used as input for existing engineering practices, not as a directive. Hence, it is reasonable to treat the present findings as a starting point for analysis, not as an independent warning system.

5.3 Ethical and Safety Considerations

Even if the current thesis does not deal with human subjects' predictions but rather with industrial analytics, some ethical questions remain concerning data management, automation bias, and the consequences of reacting to false alerts. Data used in the process are commercial, and therefore their security, accessibility, and proper handling need to be considered even in the context of scholarly research. Due to the predictive capabilities of the liquid loading model's output, it needs to be approached from the perspective of a decision support system and not as a replacement for an engineer's experience and judgment. It is crucial to keep in mind that the model needs to exhibit consistent performance in relation to various wells or wells categories.

Security issues persist since erroneous predictions could cause disruptions in production, unnecessary measures, or even delays in necessary actions. Erroneous negatives, for instance, could mean delaying an intervention when it is necessary to prevent the accumulation of liquids, unstable flow, and further reduction in production. An error of the positive sort means that unnecessary actions are performed; such actions could decrease production and disturb the operational state of affairs. For that reason, both kinds of mistakes preclude the direct automation of the process; the better solution would be to have a multistep process whereby increased risks of model errors call for reconfirmation rather than automatic measures. In case of partial data or data clearly outside the training set, the algorithm must act by downgrading itself to advice-giving status. In essence, security and production integrity rest not on model results but on their integration into procedures.

Chapter 6

Conclusion

In this thesis, liquid loading monitoring was recast as a time-to-event classification problem. A discrete hazard pipeline was created from daily data that corresponds to the engineer’s labeling. With leakage-resistant validation protocol including hold-out testing and out-of-fold threshold selection, gradient-boosted decision trees achieved the best baseline performance ($\text{ROC-AUC} = 0.929$, $\text{PR-AUC} = 0.145$) on previously unobserved wells. The horizon approach was able to discriminate slightly better for longer lead times than shorter lead times.

This, then, is the primary contribution of the thesis – a replicable standard for testing, rather than an implemented warning system in itself. While even the optimal hazard predictor identified most out-of-sample first cases, there were plenty of false positives, and the severity study was constrained by an extremely limited amount of data. Overall, these findings point toward the existence of predictive information in the operational telemetry data, but not quite sufficient to claim much more than its presence until further improvements in sensing, calibration, and deployment.

The following avenues are worth exploring in future research efforts. First and foremost, it would be useful to incorporate additional process features, like tubing and casing pressure, bottom hole parameters, intervention data, and completion data, since the current data set lacks some descriptors that may have been crucial for the classification task. Second, external validation against a different data source would help determine how portable the detected patterns are across other settings beyond

the existing data set and well-specific operational regimes. Finally, the discrete time hazard baseline could be compared to sequential methods that are better at leveraging the temporal dimension, such as LSTM and attention models.

Bibliography

- [1] M. Chen, Z. Sun, and W. Chen, “A summary of wellbore fluid accumulation and gas well drainage technology,” *IOP Conference Series: Earth and Environmental Science*, vol. 621, no. 1, Art. no. 012113, 2021, doi: 10.1088/1755-1315/621/1/012113.
- [2] R. G. Turner, M. G. Hubbard, and A. E. Dukler, “Analysis and prediction of minimum flow rate for the continuous removal of liquids from gas wells,” *Journal of Petroleum Technology*, vol. 21, no. 11, pp. 1475–1482, 1969, doi: 10.2118/2198-PA.
- [3] B. Guo, A. Ghalambor, and C. Xu, “A systematic approach to predicting liquid loading in gas wells,” *SPE Production & Operations*, vol. 21, no. 1, pp. 81–88, 2006, doi: 10.2118/94081-PA.
- [4] R. P. Sutton, S. A. Cox, J. F. Lea, and O. L. Rowlan, “Guidelines for the proper application of critical velocity calculations,” *SPE Production & Operations*, vol. 25, no. 2, pp. 182–194, 2010, doi: 10.2118/120625-PA.
- [5] C. A. M. Veeken, J. Hu, and H. M. Schiferli, “Gas-well liquid-loading field-data analysis and multiphase-flow modeling,” *SPE Production & Operations*, vol. 25, no. 3, pp. 275–283, 2010, doi: 10.2118/123657-PA.
- [6] X. Liu, S. Falcone, J. R. Barbosa, and coauthors, “Liquid loading in gas wells: From core-scale transient flow to field-scale numerical simulation,” *Journal of Petroleum Science and Engineering*, vol. 157, pp. 1056–1069, 2017, doi: 10.1016/j.petrol.2017.08.027.

- [7] A. Joseph and P. D. Hicks, “Modelling mist flow for investigating liquid loading in gas wells,” *Journal of Petroleum Science and Engineering*, vol. 170, pp. 476–484, 2018, doi: 10.1016/j.petrol.2018.06.023.
- [8] X. H. Tan, Z. Yang, and Y. Liu, “Prediction of liquid accumulation in horizontal gas wells using liquid film theory,” *Journal of Industrial and Engineering Chemistry*, 2025, doi: 10.1080/12269328.2025.2556839.
- [9] C. Fu, X. Li, M. Cai, W. Zhang, and B. Huang, “A prediction model of critical liquid-carrying velocity of gas well considering the mechanism of liquid film shear and droplet entrainment,” *Case Studies in Thermal Engineering*, vol. 73, Art. no. 106555, 2025, doi: 10.1016/j.csite.2025.106555.
- [10] H. Jia, J. Zhu, G. Cao, Y. Lu, B. Lu, and H. Zhu, “A model ranking approach for liquid loading onset predictions,” *SPE Production & Operations*, vol. 37, no. 3, pp. 370–384, 2022, doi: 10.2118/209578-PA.
- [11] Y. Chen, Y. Huang, B. Miao, X. Shi, and P. Li, “Adaptive anomaly detection-based liquid loading prediction in shale gas wells,” *Journal of Petroleum Science and Engineering*, vol. 214, Art. no. 110522, 2022, doi: 10.1016/j.petrol.2022.110522.
- [12] W. Xia, B. Liu, and H. Xiang, “Prediction of liquid accumulation height in gas well tubing using integration of crayfish optimization algorithm and XGBoost,” *Processes*, vol. 12, no. 9, Art. no. 1788, 2024, doi: 10.3390/pr12091788.
- [13] P. Chen, Y. Chen, C. Yang, Y. Xu, and G. Feng, “Gas well production optimization: Classifying liquid loading severity in shale gas wells using semi-supervised learning,” *Gas Science and Engineering*, vol. 128, Art. no. 205394, 2024, doi: 10.1016/j.jgsce.2024.205394.
- [14] Y. Chen, B. Miao, Y. Wang, Y. Huang, Y. Jiang, and X. Shi, “A deep regression method for gas well liquid loading prediction,” *SPE Journal*, vol. 29, no. 4, pp. 1847–1862, 2024, doi: 10.2118/218104-PA.

- [15] M. Eita, P. Aditama, P. Jaeger, and N. Perozo, “Prediction of liquid loading using physics-based and data driven machine learning model for a German gas field,” in *Proc. SPE Europe Energy Conf. and Exhib.*, 2025, Paper SPE-225515-MS, doi: 10.2118/225515-MS.
- [16] F. Zeng, P. Chen, J. Guo, C. Yang, and G. Feng, “Enhancing the performance of data-driven liquid loading severity grading models for shale gas wells using contrastive learning,” *Expert Systems with Applications*, vol. 313, Art. no. 131580, 2026, doi: 10.1016/j.eswa.2026.131580.
- [17] A. M. Salisu, I. Ayuba, A. Abdulrasheed, and A. Usman, “Predicting liquid loading in gas condensate wells using machine learning to enhance production efficiency,” *Petroleum Science and Engineering*, vol. 9, no. 2, pp. 55–66, 2025, doi: 10.11648/j.pse.20250902.12.
- [18] Y. Wang, J. Lian, and C. Li, “A dataset of natural gas and liquid level for oil field production prediction in China,” *Scientific Data*, vol. 12, Art. no. 1071, 2025, doi: 10.1038/s41597-025-05309-w.
- [19] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
- [21] Equinor, “Volve field data village,” 2018. [Online]. Available: <https://www.equinor.com/energy/volve-data-sharing>
- [22] R. V. Vargas, A. C. C. P. de Souza, and R. J. Poppi, “3W dataset: A realistic and public dataset with rare undesirable real events in oil wells,” GitHub repository, 2019. [Online]. Available: https://github.com/ricardovvargas/3w_dataset

- [23] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, 2017.
- [25] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017.