

Orderbook Feature Learning and Asymmetric Generalization in Intraday Electricity Markets

Runyao Yu^{*†}, Ruochen Wu^{*}, Yongsheng Han^{*}, Jochen L. Cremer^{*†}

^{*} Delft University of Technology, Delft, The Netherlands

[†] Austrian Institute of Technology, Vienna, Austria

Abstract—Accurate probabilistic forecasting of intraday electricity prices is critical for market participants to inform trading decisions. Existing studies rely on specific domain features, such as Volume-Weighted Average Price (VWAP) and the last price. However, the rich information in the orderbook remains underexplored. Furthermore, these approaches are often developed within a single country and product type, making it unclear whether the approaches are generalizable. In this paper, we extract 384 features from the orderbook and identify a set of powerful features via feature selection. Based on selected features, we present a comprehensive benchmark using classical statistical models, tree-based ensembles, and deep learning models across two countries (Germany and Austria) and two product types (60-min and 15-min). We further perform a systematic generalization study across countries and product types, from which we reveal an asymmetric generalization phenomenon: models trained on more liquid markets or products transfer well to less liquid ones, whereas the reverse transfer leads to substantial performance degradation. The project page is at <https://runyao-yu.github.io/AsymGen/>.

Index Terms—Intraday Electricity Market, Feature Selection, Machine Learning, Generalization, Probabilistic Forecasting

I. INTRODUCTION

Accurate probabilistic forecasting of intraday electricity price plays a vital role in enhancing decision-making for market participants under uncertainties [1]. In continuous intraday (CID) markets, several studies have identified the volume-weighted average price (VWAP) from the most recent 15 minutes as a strong predictor of the future price index (ID₃), an exchange-defined reference price computed as the VWAP of all matched trades in the final three hours before delivery [2], [3], [4]. Previous works even argue that the last price already reflects past information, assuming weak-form efficiency [5], and report that incorporating fundamental features, such as day-ahead forecasts of renewable generation and load, offers no or very limited improvement [6], [7], [8], [9], [10], [11], thereby motivating that using only the last price as input may yield similar performance to using it together with additional exogenous features. However, this assumption does not consider the rich information available in the orderbook. A wide range of orderbook features, such as price percentiles, price momentum, and traded volumes, are not explored and could potentially enhance forecasting performance.

In the context of intraday electricity price forecasting, prior works primarily rely on classical statistical methods, such as linear regression and its variants [3], [4], [7], [9], while more

recent studies explore deep learning approaches, such as Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Transformer variants [8], [12], [13], [14], [15], [16], to better capture non-linear patterns in electricity prices. Most of these works have primarily focused on the German market, motivated by its high liquidity and large market size [4], [11]. Concurrently, there has been a notable shift in research focus from hourly (60-minute) products to quarter-hourly (15-minute) products [7], [10], [11], [12], which provide finer temporal resolution. However, existing studies typically focus on a single type of model, country, and product type, resulting in a fragmented view of model performance. This highlights the need for a unified benchmarking study that systematically compares various machine learning models across countries and product types.

As most prior studies focus on a single country and product type, it remains unclear whether the selected features and trained models generalize well across different settings. For example, a feature set optimized for the Austrian market may not perform equally well in Germany, and a model trained on 15-min products may fail to capture the dynamics of 60-min prices. This raises important questions about cross-country and cross-product-type generalization. Thus, a systematic investigation into such generalizability is necessary to understand the robustness of derived features, support the transferable development of trained models, and offer actionable insights to stakeholders operating across multiple European markets.

In this paper, we extract 384 features from the orderbook and select the optimal features (Section III). Then, we provide a comprehensive benchmarking study using classical statistical models, tree-based ensembles, and deep learning models (Section IV). Lastly, we assess the cross-country and cross-product-type generalization of the derived optimal features and trained models (Section V). We reveal an asymmetric phenomenon: while the optimal feature set and trained model derived from a more liquid market transfer well to a less liquid one, the reverse does not hold. An overview of the workflow is illustrated in Fig. 1 (a). Our main contributions are summarized as follows:

- We extract an exhaustive set of 384 statistical features from the orderbook, including price percentiles, extreme prices, and VWAPs, and reveal a set of powerful features.
- We present a comprehensive benchmark of probabilistic forecasting performance using multiple machine learning

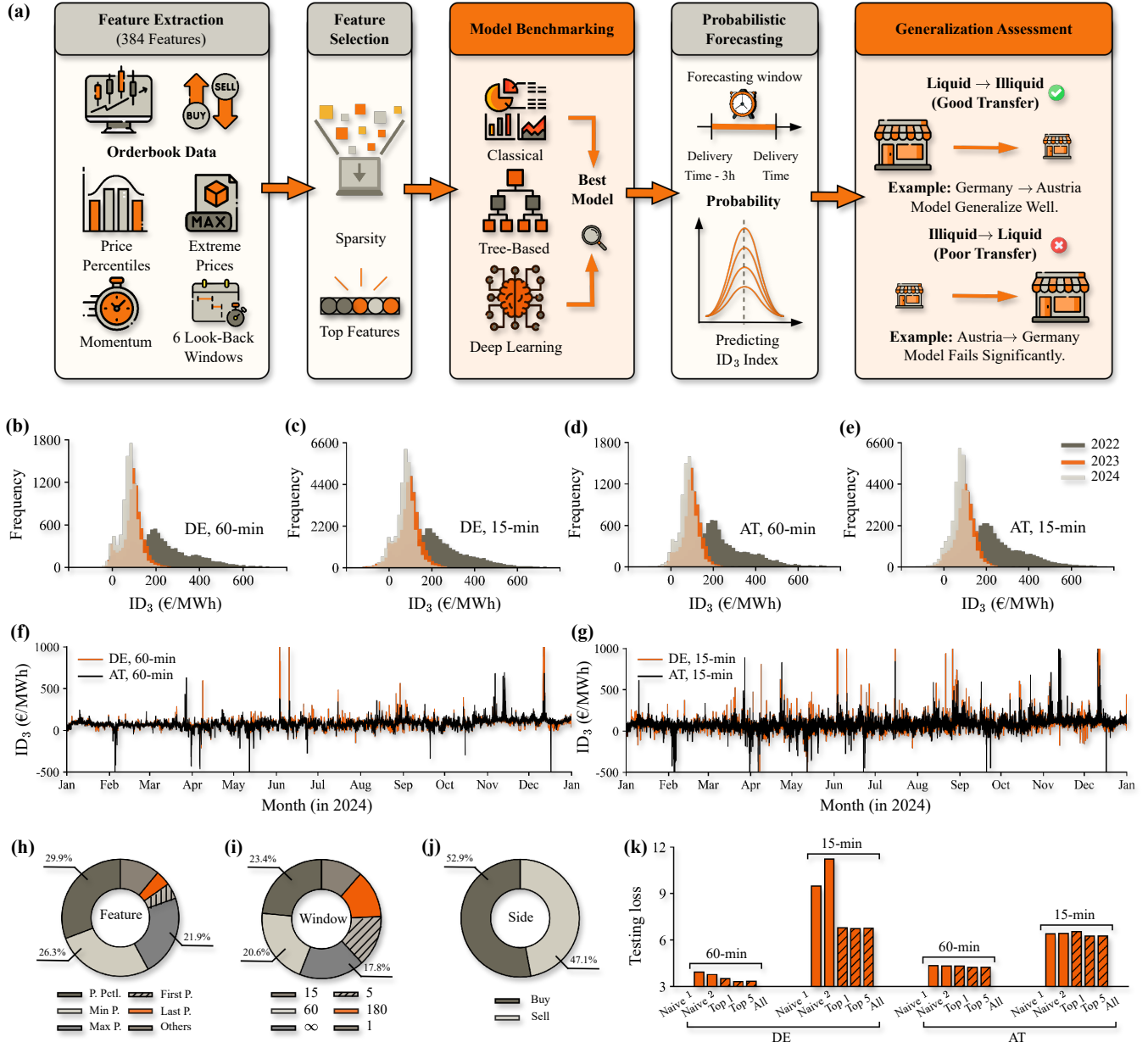


Fig. 1. (a) Overview of the workflow: extracting 384 orderbook features, performing sparse feature selection, benchmarking classical, tree-based, and deep models, producing probabilistic forecasts of the ID₃ index, and assessing cross-market and cross-product generalization. (b)-(e) Histograms of 60-min and 15-min ID₃ from Germany and Austria. The price indices exhibit high skewness and dispersion during the energy crisis in 2022, gradually reverting to a more stable distribution in 2023 and 2024. (f)-(g) ID₃ trajectories in 2024 (range limited to [-500, 1000] €/MWh for better visual comparison). Volatility increases in the order: AT, 60-min < DE, 60-min < AT, 15-min < DE, 15-min. (h)-(j) Distribution of absolute feature importance by feature type, look-back window size, and market side, respectively. (k) Testing loss (AQL) for different feature sets across countries and resolutions.

models across two countries (Germany and Austria) and two product types (60-min and 15-min).

- We systematically assess the generalizability across countries and product types. Our analysis reveals an asymmetric generalization phenomenon.

II. PRELIMINARY

The forecasting target is the widely used ID₃, as visualized in Fig. 1 (b)-(g). The ID₃ is defined as the VWAP of trades executed within a specific time window before delivery:

$$\text{ID}_3 = \frac{\sum_{s \in S} \sum_{t \in T_f} P_t^s V_t^s}{\sum_{s \in S} \sum_{t \in T_f} V_t^s}, \quad (1)$$

where the market side $s \in S = \{+, -\}$ corresponds to buy and sell orders, respectively. The forecasting time is defined as $t_f = t_d - \Delta$, with t_d denoting the delivery time and $\Delta = 180$ min representing the lead time specific to ID₃. The transaction time is defined as $t \in \mathcal{T}_f = [t_f, t_d - \delta_m]$, where $\mathcal{T}_f$ is the forecasting (trading) window. δ_m denotes the market-area-specific cutoff (minutes) set by EPEX that defines the end of the ID₃ trading window; equivalently, trades after $t_d - \delta_m$ are excluded from the index (e.g., $\delta_m = 0$ for AT, 30 for DE, and 60 for DK)¹. Here, P_t^s and V_t^s denote the price and traded volume, respectively.

III. FEATURE EXTRACTION AND SELECTION

A. Feature Extraction

We extract an exhaustive set of features from both the buy (+) and sell (−) sides across multiple look-back windows $\mathcal{T}_w = [t_f - \delta_w, t_f]$, where $\delta_w \in \{1, 5, 15, 60, 180, \infty\}$ (in minutes), and ∞ denotes the full available trading history. The full list of extracted features is summarized in Table I. If no trades are recorded within a given window (e.g., $\delta_w = 1$), we fall back to the next longer window (e.g., $\delta_w = 5$) to extract features. If no trades are observed within the full history window ($\delta_w = \infty$), the corresponding sample is discarded. Feature types include price and volume statistics (e.g., min, max, mean, percentiles), with percentile levels $p \in \mathcal{P} = \{10\%, 25\%, 45\%, 50\%, 55\%, 75\%, 90\%\}$. For example, *Price Percentile* $|_{[t_f - 15 \text{ min}, t_f], 50\%}^+$ denotes the median (50% quantile) of prices on the buy side (+) within the 15-minute look-back window.

B. Feature Selection

The extracted feature set may contain redundant or noisy features that harm generalization. Following prior works in utilizing ℓ_1 -penalized linear regression, also known as Least Absolute Shrinkage and Selection Operator (LASSO), to encourage sparse feature sets for pointwise prediction [9], we extend this idea to the probabilistic forecasting setting by applying ℓ_1 -penalized Linear Quantile Regression (LQR) to provide a transparent and interpretable baseline for identifying informative orderbook-derived features.

Given an input feature matrix $X_i \in \mathbb{R}^{N \times D}$ and target quantile vector $y_{i,\tau} \in \mathbb{R}^N$, we estimate the coefficient vector $\beta \in \mathbb{R}^D$ by solving the following optimization problem:

$$\hat{\beta} = \arg \min_{\beta} L_{\tau}(y_{i,\tau}, X_i \beta) + \alpha \|\beta\|_1 \quad (2)$$

where $L_{\tau}(\cdot)$ denotes the quantile loss:

$$L_{\tau}(y_{i,\tau}, X_i \beta) = \sum_{n=1}^N \left(y_{i,\tau}^{(n)} - X_i^{(n)} \beta \right) \cdot \left(\tau - \mathbb{I}_{\{y_{i,\tau}^{(n)} < X_i^{(n)} \beta\}} \right) \quad (3)$$

The hyperparameter $\alpha > 0$ controls the degree of sparsity by penalizing the absolute magnitudes of the coefficients and is optimized based on validation (quantile) loss. After

¹ δ_m can be retrieved from EPEX Spot download center.

TABLE I
EXTRACTED FEATURES AND DEFINITIONS.

Feature	Mathematical Definition
Price Percentile $ _{\mathcal{T}_w, p}^s$	percentile P_t^s $t \in \mathcal{T}_w, p$
Min Price $ _{\mathcal{T}_w}^s$	$\min_{t \in \mathcal{T}_w} P_t^s$
Max Price $ _{\mathcal{T}_w}^s$	$\max_{t \in \mathcal{T}_w} P_t^s$
First Price $ _{\mathcal{T}_w}^s$	first P_t^s $t \in \mathcal{T}_w$
Last Price $ _{\mathcal{T}_w}^s$	last P_t^s $t \in \mathcal{T}_w$
Mean Price $ _{\mathcal{T}_w}^s$	$\bar{P}_{\mathcal{T}_w}^s$
Price Volatility $ _{\mathcal{T}_w}^s$	$\sqrt{\frac{1}{n_{\mathcal{T}_w}^s} \sum_{t \in \mathcal{T}_w} (P_t^s - \bar{P}_{\mathcal{T}_w}^s)^2}$
Delta Price $ _{\mathcal{T}_w}^s$	last $P_t^s - \text{first } P_t^s$ $t \in \mathcal{T}_w$
Volume Percentile $ _{\mathcal{T}_w, p}^s$	percentile V_t^s $t \in \mathcal{T}_w, p$
Min Volume $ _{\mathcal{T}_w}^s$	$\min_{t \in \mathcal{T}_w} V_t^s$
Max Volume $ _{\mathcal{T}_w}^s$	$\max_{t \in \mathcal{T}_w} V_t^s$
First Volume $ _{\mathcal{T}_w}^s$	first V_t^s $t \in \mathcal{T}_w$
Last Volume $ _{\mathcal{T}_w}^s$	last V_t^s $t \in \mathcal{T}_w$
Mean Volume $ _{\mathcal{T}_w}^s$	$\bar{V}_{\mathcal{T}_w}^s$
Volume Volatility $ _{\mathcal{T}_w}^s$	$\sqrt{\frac{1}{n_{\mathcal{T}_w}^s} \sum_{t \in \mathcal{T}_w} (V_t^s - \bar{V}_{\mathcal{T}_w}^s)^2}$
Delta Volume $ _{\mathcal{T}_w}^s$	last $V_t^s - \text{first } V_t^s$ $t \in \mathcal{T}_w$
Sum Volume $ _{\mathcal{T}_w}^s$	$\sum_{t \in \mathcal{T}_w} V_t^s$
Trade Count $ _{\mathcal{T}_w}^s$	$n_{\mathcal{T}_w}^s$
VWAP $ _{\mathcal{T}_w}^s$	$\frac{\sum_{t \in \mathcal{T}_w} P_t^s V_t^s}{\sum_{t \in \mathcal{T}_w} V_t^s}$
Momentum $ _{\mathcal{T}_w}^s$	$\frac{\text{last } P_t^s - \text{VWAP}^s}{\text{VWAP}^s}$

optimization, only features with non-zero coefficient magnitudes are retained, yielding a reduced sparse feature matrix $X_i^{(\ell_1)} \in \mathbb{R}^{N \times D(\ell_1)}$, where $D(\ell_1) \ll D$, that serves as the input to downstream quantile forecasting models.

IV. MODEL COMPARISON

Based on the optimal feature set identified in the previous step, we compare several machine learning models, spanning classical statistical models, tree-based ensembles, and deep learning models.

A. Classical Statistical Models

- **Linear Quantile Regression (LQR).** LQR models conditional quantiles as linear functions of the input variables. It is highly interpretable and computationally efficient, making it well-suited for high-frequency forecasting tasks [17]. LQR has been widely adopted in the context of intraday electricity price forecasting due to its simplicity and fast training time.
- **Quantile K-Nearest Neighbors (QKNN).** QKNN performs non-parametric quantile regression by computing empirical quantiles from the nearest neighbors in feature

space. It makes no assumptions about the underlying data distribution and is capable of capturing strong local nonlinearities, which can be useful in modeling complex market dynamics [18].

B. Tree-Based Ensemble Learning Models

- **Quantile LightGBM (QLGBM).** QLGBM extends LightGBM to support quantile regression via gradient boosting and histogram-based tree construction. It efficiently handles large-scale data and captures complex feature interactions [19]. QLGBM is a widely used model for day-ahead electricity price forecasting due to its speed and robustness across diverse feature sets.
- **Quantile XGBoost (QXGB).** QXGB adapts XGBoost for quantile objectives using regularized decision tree ensembles. It is highly effective at modeling non-linear relationships and capturing long-range dependencies [20]. XGBoost is also a commonly used model for day-ahead electricity price forecasting.

C. Deep Learning Models

- **Quantile Multi-layer Perceptron (QMLP).** QMLP applies feedforward neural networks to directly estimate conditional quantiles. It learns complex non-linear mappings and scales well with high-dimensional inputs [21]. QMLP has become a common deep learning baseline for intraday electricity price forecasting.
- **Quantile Kolmogorov–Arnold Networks (QKAN).** QKAN is a recent neural architecture based on the Kolmogorov–Arnold representation theorem, which approximates multivariate functions using compositions of univariate functions. It is designed to learn highly expressive, structured representations [22]. To the best of our knowledge, this work represents the first application of QKAN in intraday electricity price forecasting.

D. Evaluation Metrics

Model performance is evaluated using probabilistic and pointwise metrics. For probabilistic forecasting, we employ the Average Quantile Loss (AQL) and the Average Quantile Crossing Rate (AQCR). For pointwise forecasting, we report the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).

- **Average Quantile Loss (AQL).** AQL is employed to jointly evaluate the accuracy of multiple quantiles [15]. It aggregates the quantile loss over all quantile levels:

$$\text{AQL} = \frac{1}{N|Q|} \sum_{i=1}^N \sum_{\tau \in Q} L_{\tau}(y_i, \hat{y}_{i,\tau}), \quad (4)$$

where y_i is the true price, $\hat{y}_{i,\tau}$ denotes the predicted quantile, and the pinball loss L_{τ} is defined as:

$$L_{\tau}(y_i, \hat{y}_{i,\tau}) = \begin{cases} \tau \cdot (y_i - \hat{y}_{i,\tau}), & \text{if } y_i \geq \hat{y}_{i,\tau}, \\ (1 - \tau) \cdot (\hat{y}_{i,\tau} - y_i), & \text{otherwise.} \end{cases} \quad (5)$$

TABLE II
HYPERPARAMETER SEARCH RANGE.

Model	Search Range
LQR	ℓ_1 regularization: [1e-8, 1]
QKNN	n_neighbors: [5, 100] distance_metric: {euclidean, manhattan} weights: {uniform, distance}
QLGBM	n_estimators: [50, 500] max_depth: [3, 12] learning_rate: [1e-3, 1e-1] subsample: [0.5, 1.0] colsample_by_tree: [0.5, 1.0] reg_lambda: [0.0, 10.0]
QXGB	n_estimators: [50, 500] max_depth: [3, 12] learning_rate: [1e-3, 1e-1] reg_alpha: [0.0, 5.0] reg_lambda: [0.0, 10.0]
QMLP	hidden_size: [32, 1024] n_layers: [2, 6] dropout_rate: [0.0, 0.5] learning_rate: [1e-5, 1e-1] batch_size: [64, 1024]
QKAN	kan_units: [32, 1024] n_layers: [2, 6] grid_intervals: [5, 16] spline_order: [2, 4] learning_rate: [1e-5, 1e-1] batch_size: [64, 1024]

The quantile loss penalizes underestimation more heavily at higher quantiles and overestimation more heavily at lower quantiles.

- **Average Quantile Crossing Rate (AQCR).** AQCR quantifies the frequency of quantile crossing violations [23], i.e., instances where a lower quantile prediction exceeds a higher quantile prediction. For each sample i , and any quantile pair (τ_l, τ_u) with $\tau_l < \tau_u$, the crossing indicator is defined as:

$$C_{\tau_l, \tau_u}(\hat{y}_{l,i}, \hat{y}_{u,i}) = \mathbb{I}(\hat{y}_{l,i} > \hat{y}_{u,i}), \quad (6)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The overall AQCR is then computed as:

$$\text{AQCR} = \frac{1}{N} \sum_{i=1}^N C_{\tau_l, \tau_u}(\hat{y}_{l,i}, \hat{y}_{u,i}). \quad (7)$$

A lower AQCR indicates better consistency of quantile predictions, with fewer violations across quantile levels.

- **Root Mean Squared Error (RMSE).** RMSE evaluates the overall predictive quality and is sensitive to outliers:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}. \quad (8)$$

TABLE III
TOP 5 FEATURES PER MARKET, PRODUCT TYPE, AND QUANTILE.

Market	Product Type	Quantile	Top 5 features					
DE	60-min	0.1	Min P. $\left \begin{smallmatrix} + \\ \tau_{15} \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{60} \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{\infty} \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_{\infty} \end{smallmatrix} \right $	First P. $\left \begin{smallmatrix} - \\ \tau_1 \end{smallmatrix} \right $	
		0.5	P. Pctl. $\left \begin{smallmatrix} - \\ \tau_5, 90\% \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_{15}, 10\% \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} - \\ \tau_{15}, 75\% \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{15} \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_5, 10\% \end{smallmatrix} \right $	
		0.9	Max P. $\left \begin{smallmatrix} - \\ \tau_{15} \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_{\infty}, 10\% \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_5, 10\% \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_{60} \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_{180} \end{smallmatrix} \right $	
	15-min	0.1	Min P. $\left \begin{smallmatrix} + \\ \tau_{60} \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{180} \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_1 \end{smallmatrix} \right $	Mean P. $\left \begin{smallmatrix} - \\ \tau_5 \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{15} \end{smallmatrix} \right $	
		0.5	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_{60}, 10\% \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} - \\ \tau_{15}, 90\% \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{15} \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} - \\ \tau_{60}, 45\% \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_{60}, 25\% \end{smallmatrix} \right $	
		0.9	Max P. $\left \begin{smallmatrix} - \\ \tau_{60} \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} - \\ \tau_{60}, 90\% \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_1 \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_{15} \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_{180} \end{smallmatrix} \right $	
	AT	60-min	0.1	Min P. $\left \begin{smallmatrix} + \\ \tau_{15} \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{\infty} \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} - \\ \tau_5, 45\% \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} - \\ \tau_5 \end{smallmatrix} \right $	First P. $\left \begin{smallmatrix} - \\ \tau_5 \end{smallmatrix} \right $
			0.5	Last P. $\left \begin{smallmatrix} - \\ \tau_{\infty} \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_5 \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} - \\ \tau_5, 45\% \end{smallmatrix} \right $	Last P. $\left \begin{smallmatrix} + \\ \tau_{\infty} \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_1 \end{smallmatrix} \right $
			0.9	Max P. $\left \begin{smallmatrix} - \\ \tau_{180} \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_{60}, 75\% \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_1 \end{smallmatrix} \right $	First P. $\left \begin{smallmatrix} + \\ \tau_{15} \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} - \\ \tau_1, 10\% \end{smallmatrix} \right $
15-min		0.1	Min P. $\left \begin{smallmatrix} - \\ \tau_1 \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_{180}, 10\% \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{15} \end{smallmatrix} \right $	P. Pctl. $\left \begin{smallmatrix} + \\ \tau_{\infty}, 10\% \end{smallmatrix} \right $	Min P. $\left \begin{smallmatrix} + \\ \tau_{180} \end{smallmatrix} \right $	
		0.5	Mean P. $\left \begin{smallmatrix} + \\ \tau_{\infty} \end{smallmatrix} \right $	Last P. $\left \begin{smallmatrix} - \\ \tau_{\infty} \end{smallmatrix} \right $	VWAP $\left \begin{smallmatrix} + \\ \tau_1 \end{smallmatrix} \right $	VWAP $\left \begin{smallmatrix} - \\ \tau_{180} \end{smallmatrix} \right $	Mean P. $\left \begin{smallmatrix} + \\ \tau_{15} \end{smallmatrix} \right $	
		0.9	Max P. $\left \begin{smallmatrix} + \\ \tau_1 \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} + \\ \tau_{180} \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} - \\ \tau_{180} \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} + \\ \tau_1 \end{smallmatrix} \right $	Max P. $\left \begin{smallmatrix} + \\ \tau_{\infty} \end{smallmatrix} \right $	

- **Mean Absolute Error (MAE).** MAE measures the average magnitude of the prediction errors:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (9)$$

- **Coefficient of Determination (R^2).** The R^2 score quantifies the proportion of variance in the target variable explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (10)$$

where $\bar{y}$ is the mean of the true values.

V. GENERALIZATION ASSESSMENT

To assess the model's generalizability across countries and product types, we conduct two sets of experiments: (1) *cross-country generalization*, where the model is transferred between DE and AT; and (2) *cross-product-type generalization*, where the model is transferred between 60-min and 15-min products. For both experiments, the following three transfer strategies are applied:

- **A $\rightarrow$ A:** Use the optimal feature set derived from domain A; train and optimize the model on data from domain A; test it on domain A.
- **B $\rightarrow$ A:** Use the optimal feature set derived from domain B; train and optimize the model on data from domain B; test on domain A.
- **A + B $\rightarrow$ A:** Use the union of optimal feature sets from both domains A and B; train and optimize the model on combined data from A and B; test on domain A.

Furthermore, we introduce two measures to quantify the phenomenon of *asymmetric generalization*: the *loss ratio* $\mathcal{L}$ and the *trade-count ratio* $\mathcal{C}$, defined as follows:

$$\mathcal{L} = \frac{\text{AQL}(\text{B} \rightarrow \text{A})}{\text{AQL}(\text{A} \rightarrow \text{A})}, \quad (11)$$

where $\text{AQL}(\text{B} \rightarrow \text{A})$ and $\text{AQL}(\text{A} \rightarrow \text{A})$ are testing loss and can be retrieved from Table V. A higher value of $\mathcal{L}$ indicates that a model transferred from domain B performs worse on domain A than a model trained directly on domain A;

$$\mathcal{C} = \frac{N_B}{N_A}, \quad (12)$$

where N_A and N_B represent the average count of matched trades across testing samples for domains A and B, respectively. These values reflect the market liquidity and are illustrated in Fig. 2 (a), where DE 60-min exhibits the highest trade count (most liquid), and AT 15-min the lowest (least liquid). A higher value of $\mathcal{C}$ indicates that transfer learning is performed from a more liquid domain to a less liquid one.

A. Cross-Country Generalization

In this setting, domains A and B refer to countries. Specifically, A can be either DE or AT, and B is the other. We evaluate model generalization across countries for the 60-min and 15-min product types, respectively.

B. Cross-Product-Type Generalization

In this setting, domains A and B refer to product types. Specifically, A can be either the 60-min or 15-min product, and B is the other. We evaluate model generalization across product types for the DE and AT markets, respectively.

VI. EXPERIMENT

The orderbook is split into training (2022-01-01 to 2024-01-01), validation (2024-01-01 to 2024-07-01), and testing (2024-07-01 to 2025-01-01) periods. The testing window is chosen to examine the model's performance on more recent, up-to-date data. For the 60-min and 15-min products, a prediction is generated every 60 minutes and 15 minutes, respectively. Each model is optimized with 100 trials using Optuna, which applies Bayesian optimization for efficient hyperparameter tuning [24]. The corresponding hyperparameter search ranges are summarized in Table II.

TABLE IV
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS.

Market	Product Type	Model	AQL ↓	AQCR ↓	RMSE ↓	MAE ↓	R2 ↑
DE	60-min	LQR [†]	3.42±0.00	0.05±0.00	27.94±0.00	10.05±0.00	0.87±0.00
		QKNN [†]	3.51±0.00	0.06±0.00	28.41±0.00	10.33±0.00	0.86±0.00
		QLGBM	3.62±0.00	0.14±0.03	29.88±0.14	11.45±0.29	0.85±0.01
		QXGB	3.59±0.00	0.08±0.03	28.75±0.36	10.95±0.10	0.86±0.01
		QMLP	3.30±0.01	0.01±0.00	27.67±0.24	10.02±0.02	0.87±0.00
		QKAN	<u>3.32±0.02</u>	<u>0.01±0.00</u>	<u>27.69±0.17</u>	<u>10.06±0.05</u>	0.87±0.00
	15-min	LQR [†]	6.79±0.00	0.09±0.00	52.68±0.00	18.41±0.00	0.69±0.00
		QKNN [†]	6.80±0.00	0.10±0.00	52.78±0.00	18.56±0.00	0.69±0.00
		QLGBM	6.82±0.05	0.19±0.09	52.93±0.37	18.37±0.20	0.69±0.01
		QXGB	6.75±0.04	0.14±0.05	52.40±0.41	17.97±0.33	0.70±0.01
		QMLP	<u>6.56±0.03</u>	0.01±0.00	50.40±0.32	17.84±0.28	0.72±0.00
		QKAN	6.56±0.02	<u>0.01±0.00</u>	<u>50.42±0.27</u>	<u>17.88±0.09</u>	<u>0.72±0.00</u>
	60-min	LQR [†]	4.33±0.00	0.05±0.00	28.61±0.00	11.91±0.00	0.80±0.00
		QKNN [†]	4.46±0.00	0.04±0.00	29.29±0.00	12.23±0.00	0.79±0.00
		QLGBM	4.49±0.05	0.10±0.04	29.41±0.42	12.14±0.17	0.78±0.01
		QXGB	4.45±0.03	0.04±0.01	28.90±0.38	11.97±0.10	0.79±0.01
		QMLP	<u>4.20±0.01</u>	0.01±0.00	<u>28.33±0.23</u>	11.69±0.06	0.80±0.00
		QKAN	4.19±0.01	<u>0.01±0.00</u>	28.27±0.14	<u>11.75±0.10</u>	<u>0.80±0.00</u>
	15-min	LQR [†]	6.44±0.00	0.04±0.00	52.99±0.00	17.82±0.00	0.57±0.00
		QKNN [†]	6.44±0.00	0.09±0.00	52.02±0.00	17.84±0.00	0.57±0.00
		QLGBM	6.38±0.11	0.05±0.01	51.97±0.24	17.42±0.18	0.57±0.01
		QXGB	6.47±0.08	0.07±0.02	51.67±0.18	17.32±0.15	0.58±0.01
		QMLP	6.22±0.06	0.01±0.00	<u>51.24±0.22</u>	<u>17.18±0.21</u>	0.58±0.00
		QKAN	<u>6.22±0.08</u>	<u>0.01±0.00</u>	51.15±0.11	17.09±0.14	<u>0.58±0.00</u>

A. Feature Extraction and Selection

We retain features with non-zero coefficients and analyze their importance by summing the absolute coefficient magnitudes across feature type, look-back window size, and market side, respectively. As shown in Fig. 1 (h), the most important feature types are *price percentiles* (29.9%), *minimum prices* (26.3%), and *maximum prices* (21.9%), while the volume-based features do not contribute much in terms of coefficient magnitude. In Fig. 1 (i), features extracted from the last 15 and 60 minutes contribute the most (23.4% and 20.6%). Surprisingly, the last 1-minute window contributes only 11.3% of total importance. This may be due to the volatility and noise in short-term trading activity. Fig. 1 (j) shows that buy-side features slightly dominate sell-side features, although the difference is marginal. We note that this feature-importance analysis reflects only linear associations between features and price labels, non-linear models may exploit complex interactions and non-linear transformations, which can shift the relative importance across feature groups.

To evaluate the impact of feature selection on performance, we rank all features by their absolute coefficient values per market, product type, and quantile. The LQR models are retrained using: only the top 1 feature; the top 5 features; and all the selected features. Additionally, we include two previously reported strong predictors as benchmarks: VWAP over the last 15 minutes (Naive¹) and the last price (Naive²). As shown in Fig. 1 (k), the selected full feature set significantly outperforms both naive baselines, achieving on average 10.53% and 11.87% lower testing loss compared to Naive¹

and Naive², respectively. Moreover, the top 5 features are often sufficient to match the performance of the full set, indicating redundancy among weaker features. While the top 1 feature yields comparable performance to the top 5 in the 15-minute product in DE, it performs worse in other settings. Therefore, we proceed by using the top 5 features, revealed in Table III, for the downstream forecasting task.

B. Model Comparison

The results of the model comparison are illustrated in Table IV, where all metrics are reported as mean±standard deviation over 5 independent runs. The best results are in **bold**. Models marked with [†] lack random-seed control; thus, the standard deviation is zero. The units of AQL, RMSE, and MAE are expressed in €/MWh, and AQCR in %. We observe substantial variation in AQL across probabilistic forecasting scenarios. Specifically, the difficulty in probabilistic forecasting is in the order: DE, 60-min < AT, 60-min < AT, 15-min < DE, 15-min, as indicated by the average AQL across models. This order contradicts the volatility order: AT, 60-min < DE, 60-min < AT, 15-min < DE, 15-min, as observed from Fig. 1 (f)-(g). One possible explanation is that the higher liquidity in the DE market provides richer and more stable predictive features for 60-min products, partially offsetting the impact of volatility. In contrast, the AT 60-min product combines lower volatility with lower liquidity compared to the DE 60-min product, and the lower liquidity increases the difficulty of probabilistic forecasting.

Among the six models compared, the deep learning approaches consistently outperform classical statistical models

TABLE V
CROSS-DOMAIN GENERALIZATION PERFORMANCE.

Cross-Country Generalization						
Product Type	Trans. Strategy	AQL ↓	AQCR ↓	RMSE ↓	MAE ↓	R2 ↑
60-min	DE → DE	3.30±0.01	0.01±0.00	27.67±0.24	10.02±0.01	0.87±0.00
	AT → DE	23.84±13.34	0.41±0.10	35.60±9.64	11.60±1.61	0.78±0.12
	DE + AT → DE	3.47±0.02	0.01±0.00	27.91±0.09	10.12±0.04	0.86±0.01
	AT → AT	4.20±0.01	0.01±0.00	28.33±0.23	11.69±0.06	0.80±0.00
	DE → AT	4.20±0.01	0.02±0.00	28.85±0.06	11.98±0.04	0.80±0.00
	DE + AT → AT	4.15±0.01	0.01±0.00	28.23±0.05	11.55±0.07	0.80±0.00
15-min	DE → DE	6.56±0.03	0.01±0.00	50.40±0.32	17.84±0.28	0.72±0.00
	AT → DE	80.15±11.23	0.08±0.01	88.99±8.16	39.48±2.84	0.12±0.16
	DE + AT → DE	8.89±0.41	0.06±0.04	55.29±2.08	21.94±1.33	0.66±0.03
	AT → AT	6.22±0.06	0.01±0.00	51.24±0.22	17.18±0.21	0.58±0.00
	DE → AT	6.27±0.12	0.02±0.04	52.89±0.13	17.79±0.19	0.58±0.00
	DE + AT → AT	6.20±0.03	0.00±0.00	51.00±0.04	17.04±0.09	0.58±0.00
Cross-Product-Type Generalization						
Market	Trans. Strategy	AQL ↓	AQCR ↓	RMSE ↓	MAE ↓	R2 ↑
DE	60-min → 60-min	3.30±0.01	0.01±0.00	27.67±0.24	10.02±0.01	0.87±0.00
	15-min → 60-min	4.45±0.42	0.04±0.03	28.90±0.07	10.34±0.11	0.86±0.00
	60-min + 15-min → 60-min	3.57±0.01	0.02±0.00	27.98±0.12	10.13±0.06	0.87±0.00
	15-min → 15-min	6.56±0.03	0.01±0.00	50.40±0.32	17.84±0.28	0.72±0.00
	60-min → 15-min	6.59±0.28	0.02±0.00	50.83±2.39	17.89±0.35	0.72±0.00
	60-min + 15-min → 15-min	6.33±0.02	0.01±0.00	50.10±0.36	17.63±0.07	0.72±0.00
AT	60-min → 60-min	4.20±0.01	0.01±0.00	28.33±0.23	11.69±0.06	0.80±0.00
	15-min → 60-min	8.17±3.90	0.13±0.23	29.99±0.24	12.41±0.10	0.78±0.00
	60-min + 15-min → 60-min	4.46±0.04	0.02±0.00	29.47±0.03	12.16±0.04	0.79±0.00
	15-min → 15-min	6.22±0.06	0.01±0.00	51.24±0.22	17.18±0.21	0.58±0.00
	60-min → 15-min	6.22±0.09	0.01±0.00	52.18±0.25	17.59±0.11	0.58±0.00
	60-min + 15-min → 15-min	6.21±0.05	0.00±0.00	51.15±0.05	17.05±0.05	0.58±0.00

and tree-based ensembles. In particular, QMLP achieves on average 3.45% and 4.59% lower AQL than LQR and QKNN, respectively, when averaged across markets and product types. Furthermore, QLGBM and QXGB result in 5.08% and 4.83% higher AQL on average compared to QMLP. In addition, the classical statistical methods and tree-based ensembles exhibit higher AQCR values, ranging from 0.04% to 0.19%, indicating more unreliable probabilistic forecasting. This issue is expected to be further magnified when predicting additional quantiles. We also note that QMLP and QKAN perform nearly identically across all metrics. However, QKAN requires approximately 9.7 times longer training time per epoch due to its neural decomposition and multivariate integration structure. Therefore, QMLP offers the best trade-off between computational efficiency and predictive performance and is selected for the downstream generalization assessment.

C. Generalization Assessment

In the *cross-country experiments*, models trained on the DE market generalize well to the AT market, while the reverse direction results in substantial degradation of performance, as observed from Table V. In both 60-min and 15-min settings, separate training achieves the best performance across all metrics when predicting DE prices, while joint training leads to the best or equivalent performance when predicting AT

prices. Notably, when directly transferring a model trained on AT orderbook data, the AQL increases drastically from 3.30 to 23.84 for the 60-min product and from 6.56 to 80.15 for the 15-min product (highlighted in orange in the table). In contrast, DE-trained models maintain similar performance when applied directly to the AT market (highlighted in gray in the table). These results highlight a clear asymmetric phenomenon, consistent with the higher liquidity of the DE market being associated with better transfer toward the less liquid AT market.

In the *cross-product-type experiments*, models trained on the 60-min product generalize well to the 15-min product, while the reverse direction again leads to inferior performance, as observed from Table V. In both DE and AT markets, separate training yields the best results across all metrics when predicting 60-min prices, whereas joint training improves or maintains performance when predicting 15-min prices. Notably, directly transferring a model trained on 15-min data results in AQL increases from 3.30 to 4.45 in DE and from 4.20 to 8.17 in AT (highlighted in orange in the table). Meanwhile, transferring from 60-min to 15-min retains similar performance compared to separate training (highlighted in gray in the table). These results again suggest that the asymmetric phenomenon is associated with liquidity differences, as 60-min products contain more trades. In contrast, 15-min products are sparser

and more volatile, limiting their generalizability to coarser timescales.

Fig. 2 (b) shows the scatter of $(\mathcal{C}, \mathcal{L})$ and its empirical fitting curve. For $\mathcal{C} \geq 1$, where transfer learning is performed from a more liquid domain to a less liquid one, the loss ratio remains at $\mathcal{L} = 1$, indicating performance equivalent to training directly on the target domain. In contrast, for $\mathcal{C} < 1$, where transfer occurs from a less liquid domain to a more liquid one, a clear exponential trend is observed: as $\mathcal{C}$ decreases, the loss ratio $\mathcal{L}$ increases sharply, indicating worse performance compared to target-only training. These observations confirm the role of liquidity in transfer performance and support the emergence of the asymmetric generalization phenomenon.

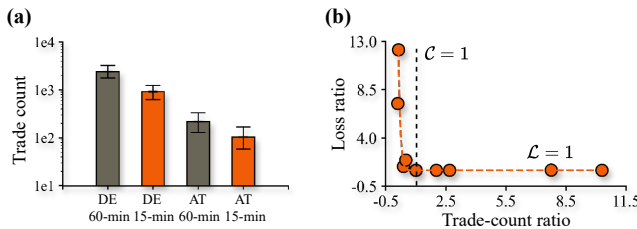


Fig. 2. Analysis of model performance against market liquidity. (a) Comparison of market liquidity. (b) Loss ratio versus trade-count ratio.

VII. CONCLUSION

In this paper, we developed a comprehensive feature set consisting of 384 orderbook features and revealed that price percentiles and extreme prices outperform the previously reported powerful features such as VWAP and last price. Moreover, through model comparison, we find that deep learning models consistently outperform classical statistical models and tree-based ensembles. In particular, QMLPs emerge as a strong baseline for probabilistic forecasting when using domain features. Finally, our generalization assessment uncovers a pronounced asymmetry in transferability: models trained on more liquid markets or products generalize well to less liquid domains, while the reverse transfer leads to substantial performance degradation. These findings underscore the importance of market liquidity in designing better models for probabilistic intraday electricity price forecasting.

VIII. LIMITATION AND FUTURE WORK

First, the extracted features in this study are empirical and may benefit from exploring a broader feature set in future work. Second, as markets become more efficient, simpler indicators such as the last price may become sufficient. We will monitor such developments, particularly as electricity markets transition to full quarter-hourly resolution. Third, the hyperparameters are tuned empirically. Additional hyperparameter tuning and a larger number of trials may further improve performance, potentially enabling tree-based models to match the performance of deep learning models. Fourth, we use trade-count as a simple liquidity proxy; examining additional

liquidity measures (e.g., bid-ask spread and orderbook depth) and their relationship to asymmetric generalization is an interesting direction for future work. Fifth, the features are selected via LQR, which capture only linear effects; extending feature selection to non-linear models is left for future work. Lastly, this work focuses on the central regions in Europe; extending the analysis to Nordic markets is worth exploring.

REFERENCES

- [1] M. Narajewski and F. Ziel, "Ensemble forecasting for intraday electricity prices: Simulating trajectories," *Applied Energy*, vol. 279, p. 115 801, 2020.
- [2] G. Marcjasz, B. Uniejewski, and R. Weron, "Beating the naïve—combining lasso with naïve intraday electricity price forecasts," *Energies*, vol. 13, no. 7, 2020, ISSN: 1996-1073. DOI: 10.3390/en13071667. [Online]. Available: <https://www.mdpi.com/1996-1073/13/7/1667>.
- [3] M. Narajewski and F. Ziel, "Econometric modelling and forecasting of intraday electricity prices," *Journal of Commodity Markets*, vol. 19, p. 100 107, 2020.
- [4] T. Serafin, G. Marcjasz, and R. Weron, "Trading on short-term path forecasts of intraday electricity prices," *Energy Economics*, vol. 112, p. 106 125, 2022.
- [5] E. F. Fama, "Efficient capital markets," *Journal of finance*, vol. 25, no. 2, pp. 383–417, 1970.
- [6] C. Monteiro, I. J. Ramirez-Rosado, L. A. Fernandez-Jimenez, and P. Conde, "Short-term price forecasting models based on artificial neural networks for intraday sessions in the iberian electricity market," *Energies*, vol. 9, no. 9, p. 721, 2016.
- [7] J. R. Andrade, J. Filipe, M. Reis, and R. J. Bessa, "Probabilistic price forecasting for day-ahead and intraday markets: Beyond the statistical model," *Sustainability*, vol. 9, no. 11, p. 1990, 2017.
- [8] T. Janke and F. Steinke, "Forecasting the price distribution of continuous intraday electricity trading," *Energies*, vol. 12, no. 22, p. 4262, 2019.
- [9] B. Uniejewski, G. Marcjasz, and R. Weron, "Understanding intraday electricity markets: Variable selection and very short-term price forecasting using lasso," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1533–1547, 2019, ISSN: 0169-2070. DOI: <https://doi.org/10.1016/j.ijforecast.2019.02.001>.
- [10] S. Hirsch and F. Ziel, "Simulation-based forecasting for intraday power markets: Modelling fundamental drivers for location, shape and scale of the price distribution," *The Energy Journal*, vol. 45, no. 3, pp. 107–144, 2024.
- [11] S. Hirsch and F. Ziel, "Multivariate simulation-based forecasting for intraday power markets: Modeling cross-product price effects," *Applied Stochastic Models in Business and Industry*, 2024.
- [12] E. Cramer, D. Witthaut, A. Mitsos, and M. Dahmen, "Multivariate probabilistic forecasting of intraday electricity prices using normalizing flows," *Applied Energy*, vol. 346, p. 121 370, 2023.

- [13] C. Scholz, M. Lehna, K. Brauns, and A. Baier, “Towards the prediction of electricity prices at the intraday market using shallow and deep-learning methods,” in *Mining Data for Financial Applications*, V. Bitetta, I. Bordino, A. Ferretti, F. Gullo, G. Ponti, and L. Severini, Eds., Cham: Springer International Publishing, 2021, pp. 101–118, ISBN: 978-3-030-66981-2.
- [14] D. K. Kılıç, P. Nielsen, and A. Thibbotuwawa, “Intraday electricity price forecasting via lstm and trading strategy for the power market: A case study of the west denmark dk1 grid region,” *Energies*, vol. 17, no. 12, 2024, ISSN: 1996-1073. DOI: 10.3390/en17122909. [Online]. Available: <https://www.mdpi.com/1996-1073/17/12/2909>.
- [15] R. Yu et al., *Orderfusion: Encoding orderbook for end-to-end probabilistic intraday electricity price forecasting*, 2026. arXiv: 2502.06830 [q-fin.CP]. [Online]. Available: <https://arxiv.org/abs/2502.06830>.
- [16] R. Yu et al., *Deep learning for electricity price forecasting: A review of day-ahead, intraday, and balancing electricity markets*, 2026. arXiv: 2602.10071 [q-fin.CP]. [Online]. Available: <https://arxiv.org/abs/2602.10071>.
- [17] R. Koenker and G. Bassett, “Regression quantiles,” *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.
- [18] P. K. Bhattacharya and A. K. Gangopadhyay, “Kernel and nearest-neighbor estimation of a conditional quantile,” *The Annals of Statistics*, pp. 1400–1415, 1990.
- [19] G. Ke et al., “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
- [20] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16, San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794, ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. [Online]. Available: <https://doi.org/10.1145/2939672.2939785>.
- [21] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- [22] Z. Liu et al., “KAN: Kolmogorov–arnold networks,” in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=Ozo7qJ5vZi>.
- [23] V. Chernozhukov, I. Fernández-Val, and A. Galichon, “Quantile and probability curves without crossing,” *Econometrica*, vol. 78, no. 3, pp. 1093–1125, 2010. DOI: <https://doi.org/10.3982/ECTA7880>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA7880>.
- [24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’19, Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2623–2631, ISBN: 9781450362016. DOI: 10.1145/3292500.3330701. [Online]. Available: <https://doi.org/10.1145/3292500.3330701>.