COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE MÁSTER

AI-based framework for EV aggregators bidding in service markets

Director: Matteo Troncia

Co-Director: José Pablo Chaves Ávila

Madrid

Junio de 2025

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

AI-based framework for EV aggregators bidding in service markets

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2024/25 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Rafael Gómez-Aparici Vega Fecha: 24/ 06/ 2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Matteo Troncia Matteo Troncia Fecha: 26/08/2025

Fdo.: José Pablo Chaves Ávila José Pablo Chaves Ávila Fecha: 26/08/2025

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE MÁSTER

AI-based framework for EV aggregators bidding in service markets

Director: Matteo Troncia

Co-Director: José Pablo Chaves Ávila

Madrid

AI-BASED FRAMEWORK FOR EV AGGREGATORS BIDDING IN SERVICE MARKETS

Autor: Gómez-Aparici Vega, Rafael.
Director: Troncia, Matteo; Chaves Ávila , José Pablo
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este trabajo desarrolla y valida un marco de predicción a corto plazo para la demanda de carga de un agregador de vehículos eléctricos (VE) a horizonte día-adelantado y resolución horaria. A partir de ~35.000 sesiones reales, se construye una canalización reproducible (limpieza, ingeniería de variables y evaluación temporal) y se compara un Bosque Aleatorio con baselines clásicos. En el año de prueba, el modelo final reduce el MAE en ~44% frente al ingenuo estacional (R^2 = 0,735), aportando evidencia y pautas operativas para su despliegue.

Palabras clave: vehículos eléctricos, agregación, predicción de demanda, series temporales, Rendón Forest, mercado diario, hiperparámetros, importancia de variables.

1. Introducción

La adopción de vehículos eléctricos (VE) convierte la recarga agregada en un recurso de flexibilidad distribuido cuyo comportamiento debe anticiparse para garantizar una operación fiable y eficiente. Las previsiones horarias del mercado diario permiten planificar capacidad, suavizar picos y fijar límites de riesgo, además de servir como entrada estable para módulos posteriores de optimización. Este trabajo se centra en el componente de predicción (excluyendo la puja y la co-optimización del control) mediante una canalización reproducible que integra curación rigurosa de datos, diseño de características y evaluación temporal sin fugas. Con un conjunto de datos reales de sesiones de carga, se comparan modelos base clásicos con un Random Forest y se cuantifican las mejoras con métricas operativamente relevantes (MAE, RMSE, R^2, nMAE). El énfasis en la transparencia y los diagnósticos (importancia de variables y perfiles residuales) posiciona el enfoque para un despliegue práctico.

2. Definición del proyecto

Objetivo: Generar un vector día-adelantado de 24 valores horarios (kWh) para un agregador residencial, empleando únicamente la información disponible en el origen de la previsión.

Alcance: Predicción puntual (sin bidding ni control), prioridad en transparencia, reproducibilidad y gobernanza del modelo.

Datos y proceso: Agregado horario construido a partir de ~35k sesiones reales e información de temperatura y calendario, filtrado de atípicos por IQR (k=3), recorte de meses de borde y uniones estrictamente temporales.

Evaluación y éxito: Validación cronológica con un único año de prueba (2020-08-01 a 2021-07-31), como métricas se utilizarán: MAE, RMSE, R^2, nMAE. El éxito se define como una reducción sustancial del MAE respecto al ingenuo estacional manteniendo interpretabilidad.

Entregables: Producto de previsión automatizado (vector de 24 horas), informe de errores del año de prueba, diagnósticos residuales y una ficha del modelo con hiperparámetros y limitaciones conocidas.

La figura 1 muestra la arquitectura: alineación temporal, pantalla de valores atípicos IQR (k = 3, ~3,7 % de horas excluidas), características con retrasos {1, 24, 168}, estadísticas móviles (24/168 h) y codificaciones cíclicas, bases de referencia clásicas (naive estacional, ETS) y el Random Forest propuesto.

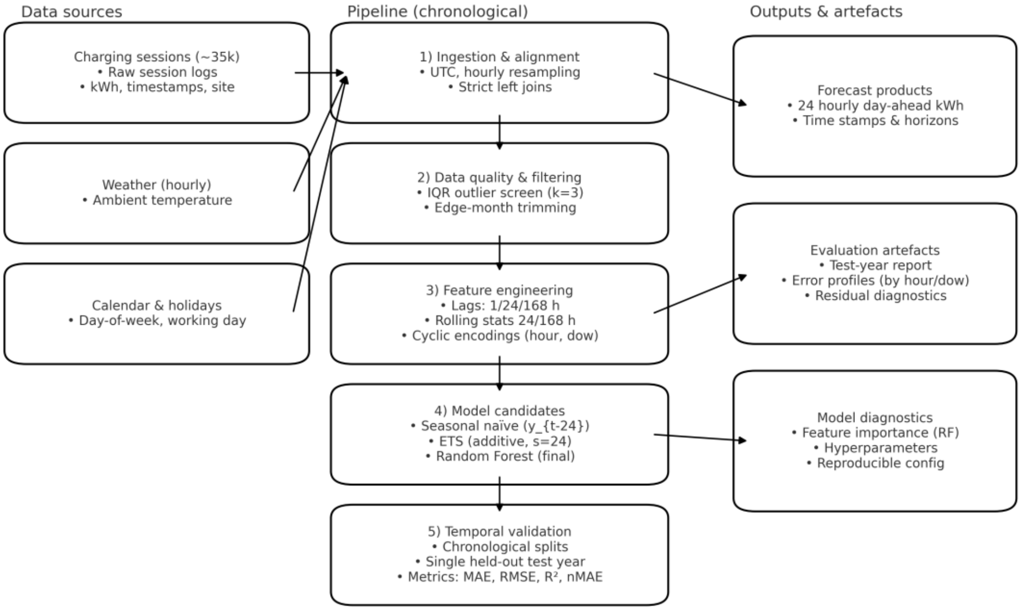


Figura 1 – Arquitectura de predicción (pipeline)

3. Resultados

Protocolo: Baselines y modelo final se entrenan en ventanas pre-comprometidas y se evalúan una única vez en el año de prueba (8.760 horas).

Baselines: Ingenuo estacional (y_{t-24}): MAE = 5,423 kWh; RMSE = 8,600 kWh; R^2 = 0,175; nMAE = 80,88%. ETS (aditivo, s=24) rinde ligeramente peor; la variante SARIMA se degrada en esta serie.

Modelo final (Random Forest): MAE = 3,013 kWh; RMSE = 4,874 kWh; R^2 = 0,735; nMAE = 44,93% (~44% de reducción de MAE frente al ingenuo estacional). Los errores absolutos se concentran en picos matutinos y vespertinos, la nMAE aumenta de madrugada por denominadores bajos. La importancia de variables muestra dominancia del lag de 1 hora y estructura diurna/semanal, temperatura y calendario añaden valor limitado.

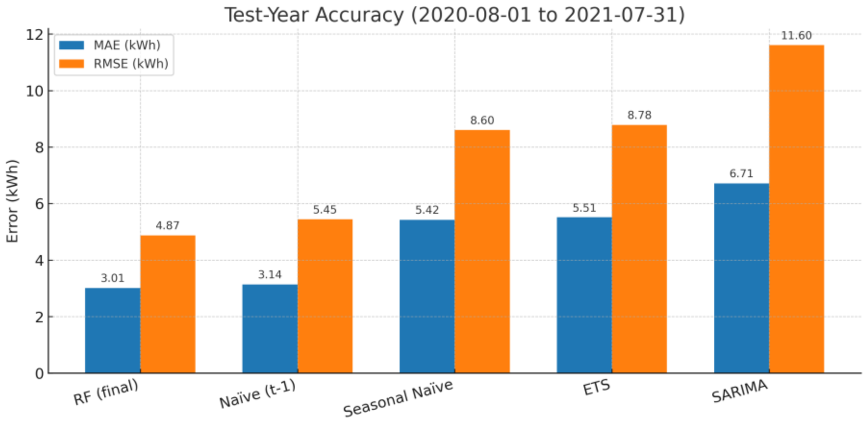


Ilustración 2 - Simulación del bucle completo de la etapa de frecuencias

4. Conclusiones

Un Random Forest, integrado en una evaluación temporal rigurosa y libre de fugas de información, supera a sólidos baselines clásicos y ofrece un comportamiento transparente mediante la importancia de variables. Las mayores ganancias se producen en las horas punta de mayor relevancia operativa. La canalización precomprometida, con limpieza por IQR fija, particiones cronológicas y una única pasada de prueba, constituye un esquema práctico para el despliegue en producción y favorece la auditabilidad. En términos operativos, la precisión alcanzada es suficiente para informar la asignación de capacidad del mercado diario y los límites internos de riesgo, con una carga de mantenimiento moderada y modos de fallo claramente identificables. De cara al futuro, las prioridades incluyen el pronóstico probabilístico y señales exógenas más ricas (llegadas/ocupación, precios). La validación externa en distintas flotas y estaciones permitirá acotar la capacidad de generalización e identificar conjuntos de características dependientes del contexto.

5. Referencias

[1] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[2] Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.

[3] Sørensen, Å., Sartori, I., Lindberg, K., & Andresen, I. (2024). Electric vehicle charging dataset with 35,000 charging sessions from 12 residential locations in Norway, Zenodo. https://zenodo.org/records/13896176.

AI-BASED FRAMEWORK FOR EV AGGREGATORS BIDDING IN SERVICE MARKETS

Author: Gómez-Aparici Vega, Rafael.
Supervisor: Troncia, Matteo; Chaves Ávila , José Pablo
Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

This thesis develops and validates a short-term forecasting framework for an electric-vehicle (EV) aggregator's day-ahead, hourly charging demand. Using ~35,000 real charging sessions, we build a reproducible pipeline (cleaning, feature engineering, time-aware evaluation) and benchmark a Random Forest against classical baselines. On the held-out test year, the final model reduces MAE by ~44% versus the seasonal naive baseline ($R^2 = 0.735$), providing evidence and operational guidance for deployment.

Keywords: Electric vehicles, aggregation, demand forecasting, time series, Random Forest, day-ahead, hyperparameters, feature importance.

1. Introduction

As EV adoption accelerates, aggregated charging loads emerge as a distributed flexibility resource whose behavior must be anticipated to ensure reliable and economical system operation. Accurate day-ahead, hourly forecasts enable capacity planning, peak shaving, and informed risk limits, and they provide a stable input to downstream optimization and scheduling modules. This thesis addresses the forecasting component, explicitly excluding bidding and control co-optimization, by developing a reproducible pipeline that integrates rigorous data curation, parsimonious feature design, and time-ordered evaluation. Using a large, real-world dataset of charging sessions, we benchmark classical baselines against a Random Forest and quantify gains with operationally meaningful metrics (MAE, RMSE, R^2, nMAE). The emphasis on transparency, leakage-free validation, and diagnostic reporting (feature importance and residual profiles) positions the approach for practical deployment and future extensions to probabilistic forecasting and richer exogenous signals.

2. Project definition

Objective: Produce a 24-element day-ahead vector of hourly charging demand (kWh) for a residential EV aggregator, using only information available at the forecast origin.

Scope: Point forecasts only (no bidding or control optimization), priority on transparency, reproducibility, and model governance.

Data & pipeline: Hourly aggregate built from ~35k real charging sessions, enriched with temperature and calendar features, IQR outlier screen (k=3), edge-month trimming, and strict time-ordered joins.

Evaluation & success: Chronological validation with a single held-out test year (2020-08-01–2021-07-31), metrics: MAE, RMSE, R^2, nMAE. Success is defined as a material

MAE reduction versus the seasonal-naïve baseline while preserving interpretability via feature importance.

Deliverables: Automated forecast product (24-hour vector), test-year error report, residual diagnostics, and a model card with hyperparameters and known limitations.

Figure 1 depicts the architecture: time alignment; IQR outlier screen (k = 3, ~3.7% hours excluded); features with lags {1, 24, 168}, rolling statistics (24/168 h), and cyclic encodings; classical baselines (seasonal naive, ETS); and the proposed Random Forest. Chronological splits and a single, final test pass prevent leakage and optimistic bias.

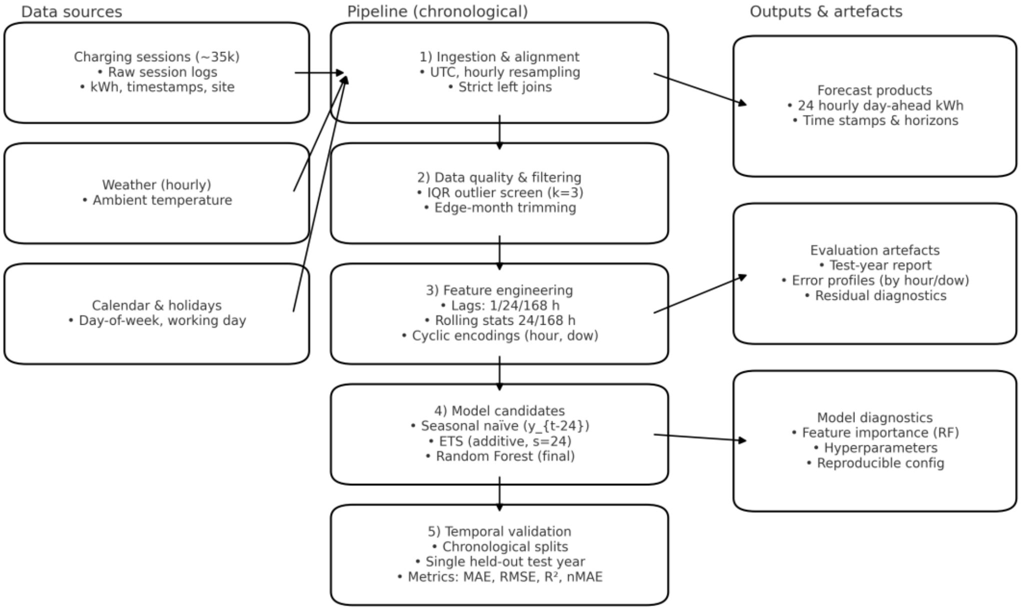

Data sources
Pipeline (chronological)
Outputs & artefacts
Charging sessions (~35k) • Raw session logs • kWh, timestamps, site
Weather (hourly) • Ambient temperature
Calendar & holidays • Day-of-week, working day
1) Ingestion & alignment • UTC, hourly resampling • Strict left joins
2) Data quality & filtering • IQR outlier screen (k=3) • Edge-month trimming
3) Feature engineering • Lags: 1/24/168 h • Rolling stats 24/168 h • Cyclic encodings (hour, dow)
4) Model candidates • Seasonal naïve (y_{t-24}) • ETS (additive, s=24) • Random Forest (final)
5) Temporal validation • Chronological splits • Single held-out test year • Metrics: MAE, RMSE, R², nMAE
Forecast products • 24 hourly day-ahead kWh • Time stamps & horizons
Evaluation artefacts • Test-year report • Error profiles (by hour/dow) • Residual diagnostics
Model diagnostics • Feature importance (RF) • Hyperparameters • Reproducible config

Figure 1 – Forecasting architecture (pipeline)

3. Results

Protocol. Baselines and the final model are trained on pre-committed windows and evaluated once on the 2020-08-01–2021-07-31 test year (8,760 hours).

Baselines. Seasonal naive (y_{t-24}): MAE = 5.423 kWh, RMSE = 8.600 kWh, R^2 = 0.175, nMAE = 80.88%. ETS (additive, s=24) underperforms slightly, the SARIMA variant degrades on this series.

Final model (Random Forest). MAE = 3.013 kWh, RMSE = 4.874 kWh, R^2 = 0.735, nMAE = 44.93% (~44% MAE reduction versus seasonal naive). Absolute errors concentrate at morning/evening peaks, normalized errors inflate overnight due to low denominators. Feature-importance diagnostics show lag-1 dominance with diurnal/weekly structure, temperature and coarse calendar add limited incremental value in this residential setting.

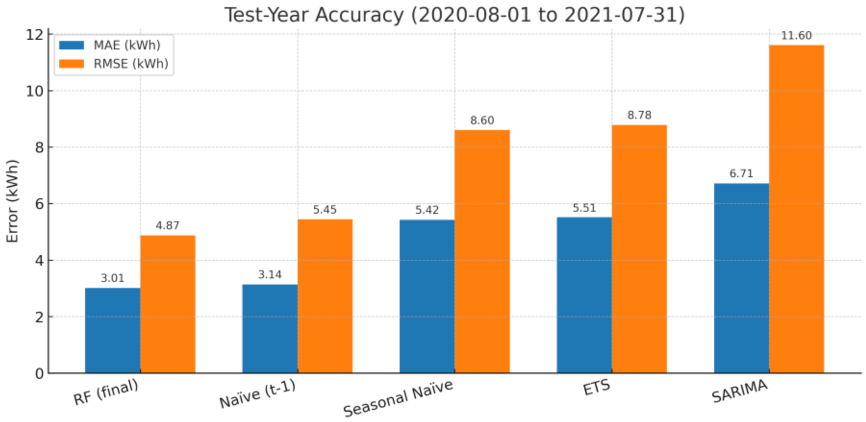


Figure 2 - Test-year accuracy (MAE/RMSE) across models

4. Conclusions

An engineered Random Forest, embedded in a disciplined, leakage-free temporal evaluation, outperforms strong classical baselines and offers transparent behavior via feature importance. Gains are largest at operationally salient peak hours. The pre-committed pipeline, fixed IQR cleaning, chronological splits, single test pass, constitutes a practical blueprint for production deployment and supports auditability. In operational terms, the achieved accuracy is sufficient to inform day-ahead capacity allocation and internal risk limits with modest maintenance overhead and clear failure modes. Looking forward, priorities include probabilistic forecasting and richer exogenous signals (arrival/occupancy, prices, refined weather), together with adaptive learning to handle concept drift, while preserving the governance that guards against ex-post tuning. External validation across fleets and seasons will further quantify generalizability and help identify context-dependent feature sets.

5. References

[1] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[2] Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.

[3] Sørensen, Å., Sartori, I., Lindberg, K., & Andresen, I. (2024). Electric vehicle charging dataset with 35,000 charging sessions from 12 residential locations in Norway, Zenodo. https://zenodo.org/records/13896176.

Table of Contents

List of Tables

List of Figures

Chapter 1. INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Rapid growth in electric vehicle (EV) adoption is transforming electricity systems by coupling mobility demand with power-sector flexibility. While a single EV has negligible system impact, large EV fleets collectively constitute a significant, highly distributed flexibility resource that can both stress and stabilize the grid depending on how charging is coordinated. This latent flexibility is increasingly mobilized through EV aggregators and intermediaries that coordinate charging and, where enabled, discharging across many vehicles to deliver energy, capacity, and ancillary services to markets and system operators while respecting user constraints. Framed as distributed energy resources (DERs), aggregated EVs can participate in day-ahead, intraday, and balancing markets, as well as in frequency regulation and demand response programs. Recent assessments underscore the material scale of this opportunity as EVs approach one quarter of global light-duty sales and begin to displace meaningful volumes of oil demand (International Energy Agency, 2025).

In Europe, the Clean Energy Package formally recognizes independent aggregators and mandates non-discriminatory market access for demand response and distributed flexibility, creating regulatory space for EV aggregation to compete in energy and ancillary services markets. This framework explicitly calls for market products, including ancillary and capacity, to be defined to enable demand-side participation, thereby lowering institutional barriers for aggregator-led flexibility (European Union, 2019).

1.2 EV AGGREGATORS AND THEIR ROLE IN ENERGY MARKETS

As documented by the (International Energy Agency, 2024) and (Muratori, 2018), accelerating electric-vehicle (EV) adoption introduces both operational challenges and strategic opportunities for electricity markets and power-system stability. While a single EV has only a marginal effect on local feeders, the coordinated behavior of thousands of EVs can materially shape net-load profiles, ramping requirements, and reserve needs. When charging is orchestrated while respecting user mobility constraints, aggregated EV fleets function as a dispatchable, distributed energy resource (DER) capable of delivering services across time scales from seconds for frequency regulation to hours for load shifting.

Under the European Union's Directive (EU) 2019/944 (2019), independent aggregators are formally recognized and granted non-discriminatory market access, establishing the institutional footing for demand-side flexibility to participate in energy and ancillary-service markets. Building on the market-design perspective of (Papadaskalopoulos & Strbac, 2013), an EV aggregator contracts with, coordinates, and manages a portfolio of EVs and its associated charging infrastructure to translate mobility needs and battery constraints into market-facing energy and ancillary-service offerings. In practice, aggregators implement control and forecasting algorithms, manage telemetry and communications, and ensure that service delivery aligns with user convenience, battery health, and regulatory requirements.

Drawing on IEA (2024) and Muratori (2018), aggregators first optimize energy consumption by shifting flexible charging to periods with lower marginal system costs or higher renewable availability. Through price-responsive or model-predictive control, they schedule charging in off-peak hours or during low locational marginal prices, thereby flattening demand profiles and reducing congestion. Where bidirectional power flow is available, aggregators may also schedule discharging during high-price or peak periods to arbitrage temporal price differences, subject to state-of-charge, user departure times, and degradation constraints. This form of smart charging mitigates sharp evening load increases associated with uncoordinated charging and can materially reduce system peaks and network reinforcements.

Aggregators also enable participation in ancillary services by modulating charging power and, in V2G settings, discharging to track grid signals. Fast, symmetric adjustments around a baseline charging set-point allow EV fleets to contribute to frequency containment and regulation, while slower adjustments support secondary reserves, voltage management, and demand-response programs. Evidence from foundational V2G studies and subsequent demonstrations, notably (Kempton & Tomić, 2005), establishes both the technical feasibility and the revenue potential of these services, with aggregate performance hinging on fleet size, communication latency, and the diversity of mobility patterns.

By pooling many small devices into a single market-facing resource, aggregators lower transaction costs and meet minimum bid sizes, enabling participation in wholesale markets. Aggregated portfolios can submit bids in day-ahead, intraday, and real-time markets, as well as capacity and reserve auctions where market design permits independent aggregation. This is codified in (Directive (EU) 2019/944, 2019) and consistent with the market mechanism described by Papadaskalopoulos and Strbac (2013). Effective market interfacing requires accurate short-term forecasting of available flexibility, robust baseline methodologies, and settlement processes that allocate revenues and penalties transparently among participating EV owners.

As argued by (Sundström & Binding, 2012), aggregators further facilitate renewable-energy integration by aligning charging with periods of high variable renewable generation such as midday solar or overnight wind. Temporal shifting of EV demand reduces curtailment, improves renewable utilization, and decreases reliance on thermal units for balancing. In systems targeting high shares of wind and solar, this coupling between transport electrification and electricity supply can lower system costs and emissions while maintaining adequacy and operational security.

At scale, EV fleets can enhance grid resilience by acting as spatially distributed storage that supports critical loads during contingencies and, where regulations and interconnection standards allow, provides black-start or islanded microgrid services. Vehicle-to-grid (V2G) and vehicle-to-building (V2B) capabilities enable back-feeding during peak events, outages, or emergencies, complementing stationary storage and improving the overall flexibility of the distribution network. Early analyses (2005) and more recent assessments (Global EV Outlook 2024) emphasize that the magnitude and reliability of these services depend on participation rates, diversity of plug-in times, and the technological readiness of bidirectional chargers and communication protocols.

EV aggregators participate in electricity markets through several distinct business models shaped by prevailing regulatory frameworks and market structures. One common model frames aggregator-operated fleets as a virtual power plant (VPP). As described by (Pudjianto, Ramsay, & Strbac, 2007) the aggregator co-optimizes many small, heterogeneous EV charging resources and presents them to the market as if they were a single dispatchable unit. By forecasting available flexibility, meeting minimum bid sizes, and ensuring telemetry and verification, the VPP can submit bids in day-ahead, intraday, and real-time markets alongside conventional generators and storage assets. Implementation hinges on accurate short-term forecasting of EV availability, robust baseline methodologies for settlement, and control architectures capable of tracking market or operator set points in near real time.

A second pathway is peer-to-peer (P2P) energy trading, in which EVs act as mobile prosumers transacting directly with other consumers or prosumers. Reviews of pilot projects and architectures, such as (Andoni, 2019), highlight how distributed ledgers and smart contracts can reduce transaction costs, automate settlement, and enhance trust in decentralized coordination, while also surfacing issues of scalability, privacy, cybersecurity, and interoperability that must be addressed before widespread deployment.

A third model emphasizes demand-response (DR) aggregation, wherein EV charging profiles are modulated to reduce peak loads, alleviate network constraints, and provide system balancing. Classic overviews (Siano, 2014) outline the need for reliable baselines, verifiable load adjustments, and incentive schemes that are both economically meaningful and acceptable to users. In practice, EV-based DR complements traditional industrial and commercial DR by adding fast, distributed flexibility with high temporal granularity.

Market design and institutional context strongly influence how these models are implemented. In liberalized markets, aggregators must comply with market-participation rules, grid codes, metering and verification requirements, and settlement procedures, often coordinating with retailers and balance-responsible parties. Recent policies including (Directive (EU) 2019/944, 2019), which recognizes independent aggregators and clarifies market access and compensation mechanisms, and the Federal Energy Regulatory (Federal Energy Regulatory Commission (FERC), 2020) mandating that RTOs/ISOs enable distributed-energy-resource aggregations have lowered barriers to entry and expanded eligible services for aggregated EV flexibility. In vertically integrated systems, by contrast, EV aggregation often proceeds via bilateral arrangements with utilities or system operators, with tariffs and program rules embedded in integrated resource planning and demand-side management portfolios. Across settings, effectiveness depends on technological readiness like bidirectional chargers or secure communications, market incentives (price volatility, reserve clearing prices), and policy frameworks that align incentives with measurable system value.

Given the coordination complexity and the stochastic availability of mobile storage, AI-driven strategies are increasingly central to operational excellence and market performance. Forecasting models predict near-term charging demand, plug-in durations, and available flexibility, price and imbalance-risk models inform bidding and hedging decisions, and optimization/control algorithms translate forecasts into dispatchable schedules. As shown in (Vázquez-Canteli & Nagy, 2019), reinforcement learning and other sequential decision-making techniques can adapt online to non-stationary conditions and heterogeneous user behavior, while stochastic and robust optimization provide risk-aware bids under uncertainty. Data driven characterization of charging patterns from real-world telemetry further improves model fidelity and user-centric service delivery, enabling more sophisticated participation across energy, capacity, and ancillary service markets.

1.3 AI-DRIVEN STRATEGIES AND THEIR SIGNIFICANCE

As the International Energy Agency (2024; 2025) documents, the rapid diffusion of electric vehicles (EVs) is reshaping power systems, intensifying the need for accurate forecasting, real-time balancing, and economically sound market participation. Uncoordinated charging can exacerbate evening peaks and feeder stress (Muratori, 2018), whereas coordinated strategies turn EV fleets into controllable resources. In this setting, Artificial Intelligence with reinforcement learning (RL) and broader machine learning toolkits emerges as a practical enabler for scalable, adaptive EV aggregation (Vázquez-Canteli & Nagy, 2019).

1.3.1 FORECASTING AGGREGATED EV DEMAND (AND RELATED SUPPLY).

Day-ahead scheduling and bidding hinge on robust expectations of plug-in behavior and charging energy. Surveys of demand response and smart-grid operation emphasize the importance of predictive baselines and multi-feature inputs for operational planning (Siano, 2014). Combined with empirical usage patterns from Muratori (2018) and system-level trends reported by the IEA (2024), data-driven models provide the short-term forecasts aggregators need to align charging with market opportunities and renewable output.

1.3.2 REAL-TIME CHARGING AND DISCHARGING OPTIMIZATION.

To respect grid limits and user constraints while reacting to volatile signals, control needs to be adapted online. Flexible-charging optimization under distribution constraints is well established (Sundström & Binding, 2012), and RL offers a natural extension for sequential decision-making in demand response and EV coordination (Vázquez-Canteli & Nagy, 2019). Together, these approaches schedule charging and, where available, vehicle-to-grid (V2G) discharging while honoring state-of-charge and network bounds.

1.3.3 MARKET BIDDING IN DAY-AHEAD, INTRADAY, AND BALANCING MARKETS.

Independent aggregation of flexible demand has viable market mechanisms (Papadaskalopoulos & Strbac, 2013), and aggregator-operated fleets can be organized as virtual power plants that meet telemetry and verification requirements (Pudjianto, Ramsay, & Strbac, 2007). Policy developments further open the door to distributed-resource aggregations Directive (EU) 2019/944 (2019) in Europe and FERC (2020) in the United States clarify participation pathways and coordination with system operators. Within this framework, AI helps translate forecasts into risk-aware bids and continuous re-optimization across sequential markets.

1.3.4 ANCILLARY SERVICES: FREQUENCY REGULATION AND DEMAND RESPONSE.

Foundational V2G analyses show that EV fleets can technically and economically supply fast, short-duration services valuable for frequency regulation (Kempton & Tomić, 2005). On the demand-response side, consolidated reviews highlight control designs, verification needs, and market interfaces for automated load modulation (Siano, 2014; Vázquez-Canteli & Nagy, 2019). AI-enabled coordinators exploit these insights to track grid signals at fine time scales and monetize flexibility without compromising user mobility.

1.3.5 DECENTRALIZED AND PEER-TO-PEER (P2P) TRADING, PRIVACY, AND COLLABORATION.

Where regulations permit, EVs can transact locally as prosumers within community or retail platforms. A systematic review of the energy sector points to blockchain-backed settlement and smart contracts as mechanisms to reduce transaction costs and automate verification, while also surfacing challenges of scalability, interoperability, and privacy (Andoni, 2019). AI complements these architectures by informing price discovery, matching, and strategic interaction capabilities that become more valuable as fleets scale.

1.3.6 OPERATIONAL ASSURANCE AND COMPLIANCE.

Aggregator platforms must satisfy metering, telemetry, and coordination requirements set by market rules and regulators (European Union, 2019; Federal Energy Regulatory Commission, 2020). Embedding AI-based monitoring and forecasting within these constraints improves operational resilience, helps detect abnormal behavior, and supports secure, auditable service delivery alongside traditional controls.

1.4 WHY AI-DRIVEN STRATEGIES ARE ESSENTIAL FOR EV AGGREGATION

AI-driven strategies have become indispensable for electric vehicle (EV) aggregation because they address the complexity and dynamism of large-scale coordination. By enabling scalable data processing, managing uncertainty in volatile environments, adapting decisions in real time, optimizing participation in electricity markets, and fostering both system sustainability and stability, AI provides the foundation for unlocking EV flexibility at scale. The following section details how these capabilities translate into tangible advantages for aggregators and the power system as a whole.

Scalability and efficiency. AI methods process heterogeneous, high-volume data from large EV portfolios and markets, automating decisions beyond manual or rule-based limits (Vázquez-Canteli & Nagy, 2019; Siano, 2014). Uncertainty management. RL and probabilistic learning handle stochastic arrivals, weather, and prices, yielding robust policies under volatility.

Real-time adaptation. Flexible-charging control proven at the device and feeder level (Sundström & Binding, 2012) combines with online learning to react to non-stationary demand and market signals, improving performance over time.

Economic optimization. Market-compatible aggregation mechanisms (Papadaskalopoulos & Strbac, 2013) and VPP coordination (Pudjianto, Ramsay, & Strbac, 2007) provide the structure. AI turns forecasts into bids and dispatch that enhance revenues and reduce costs across day-ahead, intraday, and ancillary-service opportunities.

Sustainability and stability. Coordinated, AI-enabled charging reduces peaks and integrates variable renewables (International Energy Agency, 2024), while V2G capabilities create fast services that support frequency and reserve needs (Kempton & Tomić, 2005).

In sum, AI equips EV aggregators with predictive, adaptive, and market-compatible tools that unlock flexibility at scale, strengthen grid operations, and enable profitable multi-market participation consistent with emerging regulatory frameworks in Europe and the United States.

1.5 OBJECTIVES

This thesis aims to develop and validate an AI-based framework that forecasts an electric-vehicle (EV) aggregator's short-term charging demand and demonstrates how such forecasts can be translated into economically efficient and market-compliant bidding strategies. The motivating hypothesis is that reliable, well-calibrated predictions of aggregated EV load, produced at the temporal resolution and horizons relevant to day-ahead and intraday decisions, enable aggregators to align charging with low-cost, low-carbon supply while meeting the telemetry, verification, and minimum-bid requirements of contemporary market designs.

Methodologically, the work proceeds by constructing an audit-ready dataset from session-level telemetry, applying transparent cleaning and outlier treatment, and aggregating to an hourly series with consistent calendar alignment. Feature engineering focuses on lag structures, rolling statistics, and cyclic encodings to capture diurnal and weekly seasonality, with optional exogenous drivers when available. Competing forecasting models are then trained for 1–24-hour horizons, with classical seasonal naive and exponential-smoothing baselines retained to anchor performance. Evaluation follows a rolling-origin scheme and reports mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and a normalized MAE (nMAE) to enable scale-free comparisons across seasons and load levels.

Analytically, the thesis asks what accuracy and calibration are achievable with operationally tractable AI models, which temporal and exogenous features are most influential and how stable their contributions are across seasons and horizons, how forecast uncertainty propagates to revenues and penalties in service markets and what practical guidance follows for integrating forecasting into day-ahead, participation. Model interpretation via feature-importance diagnostics is used to extract actionable insights for portfolio scheduling and market interfacing.

1.6 SUSTAINABLE DEVELOPMENT GOALS

This thesis contributes to the 2030 Agenda by advancing methods that allow EV fleets to act as reliable, grid-supportive flexibility resources without relying on market bidding. By focusing on accurate, well-calibrated forecasting of aggregated charging demand and on the operational scheduling insights that follow the work supports cleaner energy use, smarter infrastructure, more sustainable cities, and accelerated climate action.

SDG 7: AFFORDABLE AND CLEAN ENERGY

Forecasts of near-term charging demand enable aggregators to time charging to hours with abundant renewable generation (like midday solar or overnight wind) and lower marginal emissions. In residential and depot contexts, this improves on-site renewable self-consumption and reduces reliance on high-carbon imports. At the system level, shifting flexible charging away from peaks lowers dispatch costs and helps integrate variable renewables more efficiently.

SDG 9: INDUSTRY, INNOVATION, AND INFRASTRUCTURE.

The thesis delivers digital building blocks (data pipelines, feature engineering, forecasting services, and monitoring) that modernize distribution-level operations. Better visibility of forthcoming charging demand allows operators to utilize existing network capacity more effectively, reduce congestion incidents, and defer costly reinforcements. The emphasis on transparency, reproducibility, and auditability strengthens data governance and supports interoperability with emerging grid-edge technologies.

SDG 11: SUSTAINABLE CITIES AND COMMUNITIES.

Urban areas experience concentrated plug-in behavior and evening peaks. Forecast-driven scheduling smooths these peaks, mitigates local feeder stress, and improves quality of service at public and workplace charging sites. By aligning charging with local conditions, communities can retain more value from their distributed resources, while user-centric design, such as focusing on respecting mobility needs and battery health, sustains participation and social acceptance.

SDG 13: CLIMATE ACTION.

Aligning EV charging with low-emission hours directly lowers transport-related electricity emissions and reduces renewable curtailment. Peak shaving further decreases technical losses and the need for carbon-intensive peaking generation. The framework's focus on uncertainty quantification supports robust operational decisions under volatility, sustaining emissions reductions across seasons and changing fleet compositions.

Real-world impact depends on the cleanliness of the marginal grid mix, the accuracy and calibration of forecasts, user participation rates, and reliable communications. Ethical and operational safeguards, privacy-preserving data handling, transparent model validation, and battery-health-aware scheduling, are integral to ensuring that environmental benefits are achieved without compromising user trust or asset longevity.

In sum, by equipping EV aggregators with credible forecasting and operational scheduling capability, this thesis provides a practical pathway to cleaner energy use, smarter networks, more livable cities, and measurable progress toward climate goals.

Chapter 2. AI FOR DEMAND FORECASTING

Accurate demand forecasting is a prerequisite for economically and operationally viable participation of EV aggregators in energy and ancillary service markets. Unlike traditional load, an aggregator's charging demand is highly stochastic and price-responsive, shaped by heterogeneous driver behavior, spatial dispersion of charging points, weather and calendar effects, fleet composition, and control policies such as smart charging. Forecasts are required across multiple horizons very short-term (minutes to hours), intraday, and day-ahead with granular temporal resolution to inform energy procurement, reserve scheduling, and real-time balancing. In this context, probabilistic forecasting (prediction intervals or quantiles) is often more useful than point forecasts because market bids and risk management depend on the distribution of possible loads rather than a single expected value (Hong & Fan, 2016; Hyndman & Athanasopoulos, 2021).

2.1 TIME-SERIES FORECASTING METHODS

EV-aggregator load exhibits strong autocorrelation, recurring intra-day and weekly seasonality (like commuting patterns), calendar effects (weekends/holidays), and gradual trends as fleet size evolves. Time-series models explicitly exploit these temporal regularities and are well suited for short and medium term horizons from minutes to day-ahead. When market bidding requires risk-aware decisions, probabilistic forecasts (prediction intervals or quantiles) are preferable to point forecasts (2016; 2021).

2.1.1 ARIMA / ARIMAX

Autoregressive integrated moving-average models remain a strong baseline whenever the series can be rendered stationary by differencing. After differencing the series *d* times, an ARIMA(p,d,q) assumes the transformed process is governed by a joint autoregressive moving-average structure:

$$\varphi(B)(1 - B)^d y_t = \theta(B)\varepsilon_t$$

where B is the backshift operator, ϕ(·) and θ(·) are polynomials of orders p and q, and the residual (ε_t) has a distribution of N(0,σ^2) (Box, Jenkins, Reinsel, & Ljung, 2015). When external drivers such as temperature, retail price signals, or operator-announced charging events materially affect demand, an ARIMAX specification augments the conditional mean with exogenous regressors:

$$\Phi(B^s)\varphi(B)(1 - B)^d(1 - B^s)^D y_t = \beta^T x_t + \Theta(B^s)\theta(B)\varepsilon_t$$

In practice, these models are attractive because they are transparent, quick to re-estimate as fresh observations arrive, and yield closed-form prediction intervals under Gaussian errors. Their limitations are equally well understood, differencing must achieve stationarity, the linear form can underfit nonlinear responses, and performance deteriorates when several strong seasonalities coexist unless these are explicitly encoded (Box, Jenkins, Reinsel, & Ljung, 2015; Hyndman & Athanasopoulos, 2021).

2.1.2 SARIMA / SARIMAX

Seasonal ARIMA extends the ARIMA family by multiplying a nonseasonal component with a seasonal component tuned to a dominant period *s* (e.g., 24 hours or 168 hours):

$$\Phi(B^s)\varphi(B)(1 - B)^d(1 - B^s)^D y_t = \Theta(B^s)\theta(B)\varepsilon_t$$

with seasonal order $(P,D,Q)_s$. Incorporating exogenous predictors yields SARIMAX. For EV-aggregator operations, these models are often effective for day-ahead planning when daily or weekly cycles dominate a single timescale. A key caveat is that the classical multiplicative structure targets one seasonal period at a time, where intra-day and weekly cycles coexist at sub-hourly resolution, methods that represent multiple seasonalities within a single framework, such as TBATS, or decompositions with flexible seasonal terms may be preferable (2021).

2.1.3 PROPHET (META/FACEBOOK)

Prophet operationalizes an additive decomposition with a trend term, one or more seasonal terms, and calendar effects, while automating changepoint detection in the trend:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$$

Here, g(t) is a piecewise linear or logistic trend with a sparse set of estimated changepoints, s(t) aggregates user-specified seasonalities represented by Fourier series, naturally supporting multiple seasonal periods and h(t) encodes holiday or event effects and other regressors (Taylor & Letham, 2015). The approach is appealing as a rapidly deployable baseline: it is comparatively robust to missing values and outliers, it simplifies the inclusion of several seasonalities and calendar effects, and it trains quickly. Its main constraints arise when the autocorrelation structure is complex beyond additive components, in which case residual dependence may persist and interval calibration can depend sensitively on prior settings and distributional assumptions.

2.1.4 HYBRID ARIMA–LSTM MODELS

When the load process displays both linear dependence and nonlinear responses, for example threshold-like reactions to price spikes or aggregator control signals, hybrid strategies can combine complementary strengths. A common design first fits an ARIMA/SARIMA to capture linear dynamics and obtains residuals $\hat{e}_t$. A recurrent neural network, typically an LSTM, is then trained on $\{\hat{e}_t\}$ (optionally augmented with exogenous inputs) to learn nonlinear structure not explained by the linear model. Forecasts are combined additively:

$$\hat{y}_{\{t+h\}} = \hat{y}_{\{t+h\}}^{\{ARIMA\}} + \hat{r}_{\{t+h\}}^{\{LSTM\}}$$

This architecture preserves an interpretable linear backbone while allowing flexible function approximation on the remainder. The trade-offs are practical: training and tuning costs rise, the risk of overfitting increases when historical data are limited or nonstationary, and rigorous cross-validation with appropriate regularization becomes essential. Nevertheless, for high-frequency EV-aggregator demand influenced by policy, price, or weather in nonlinear ways, such hybrids frequently yield accuracy gains that translate into more reliable, risk-aware bids.

2.2 REINFORCEMENT LEARNING METHODS FOR EV AGGREGATOR BIDDING

Reinforcement Learning (RL) is a branch of machine learning concerned with sequential decision-making, where an agent learns to interact with an environment in order to maximize cumulative rewards (Sutton & Barto, 2018). Unlike supervised learning, which relies on labeled datasets, RL operates on trial-and-error mechanisms. Agents iteratively observe states, select actions, and receive feedback in the form of rewards or penalties, progressively refining their policies.

The foundations of RL are rooted in behavioral psychology, particularly operant conditioning, where learning emerges from interaction and reinforcement. In computational terms, RL is typically modeled as a Markov Decision Process (MDP), defined by a tuple (S,A,P,R,γ), where:

- S represents the set of states,
- A the set of actions,
- P the transition probabilities,
- R the reward function, and
- γ the discount factor for future rewards.

The ultimate objective is to determine an optimal policy π^*, mapping states to actions, which maximizes the expected discounted return:

$$G_t = \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1})$$

Where G_t is the return at time t.

RL is particularly suitable for problems with high uncertainty, delayed rewards, and dynamic environments, making it highly relevant for energy systems, electric vehicle (EV) aggregation, and service-market bidding strategies.

2.2.1 DEEP Q-NETWORKS (DQN)

DQN is an extension of the classical Q-learning algorithm, introduced by (Mnih, Kavukcuoglu, & Silver, 2015), where a deep neural network approximates the action-value function Q(s,a). Classical Q-learning struggles in high-dimensional or continuous state spaces because it requires a table representation of all state–action pairs. DQN overcomes this by employing deep learning to approximate Q, allowing it to scale to complex environments.

The key mechanism involves using experience replay and target networks to stabilize training. Experience replay stores transitions (s,a,r,s′) in a buffer, which are later sampled randomly to break correlations between successive observations. Target networks, updated less frequently than the main network, provide stable reference values for training.

DQN succeeds because deep networks generalize across large state spaces, and the use of stabilization techniques prevents divergence during training. This allows agents to learn effective policies in environments where tabular methods are infeasible.

DQN is compelling when actions are discrete and observations are high-dimensional. Its implementation is comparatively straightforward, and it has demonstrated strong empirical performance on diverse benchmarks. However, the algorithm can be sample-hungry, particularly in environments with sparse rewards or large action branching factors. Overestimation bias in the max operator may degrade learning stability, variants such as Double DQN address this at the cost of additional bookkeeping. Moreover, extending DQN to continuous control requires auxiliary mechanisms like discretization or actor–critic adaptations, which can erode its simplicity. For EV aggregation, vanilla DQN is attractive for stylized decision layers with discrete bids or tariff choices, but it becomes less convenient when fine-grained power set-points or continuous price-responsive actions are central.

2.2.2 Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO) is a first-order policy-gradient method designed to combine the reliability of trust-region ideas with practical scalability (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017). Instead of solving a constrained optimization as in TRPO, PPO maximizes a clipped surrogate objective,

$$L^{(clip)(\theta)} = E\left[\min\left(r_{t\widehat{(\theta)}} A_t, clip\left(r_{t(\theta)}, 1 \overparen{\, - \, \epsilon}, 1 + \epsilon\right) A_t\right)\right]$$

$$r_t(\theta) = \frac{\pi_\theta(a\_t \mid s\_t)}{\pi_{\theta_{old}}(a\_t \mid s\_t)}$$

Where $r_t(\theta)$ is the probability ratio and A_t an advantage estimate, typically from generalized advantage estimation. The clipping term suppresses excessively large policy updates that would otherwise lead to performance collapse, yielding a monotonic-improvement heuristic in practice. PPO is commonly implemented with an actor–critic architecture, where a value network reduces gradient variance and provides a baseline for A_t.

PPO's success lies in controlling the bias–variance trade-off of policy optimization. By constraining updates implicitly through the clipped ratio, it maintains a "trust region" without the computational expense of second-order constraints. Mini-batch stochastic optimization and multiple epochs over collected trajectories improve sample usage relative to naive REINFORCE, while the critic stabilizes learning by providing lower-variance advantage signals. The method scales gracefully to continuous action spaces via Gaussian policies, making it well suited for control tasks that require smooth actuation.

In applied settings, PPO is valued for its robustness across domains and relatively forgiving hyperparameters compared with earlier policy-gradient algorithms. It accommodates both discrete and continuous actions, integrates naturally with recurrent or attention-based encoders for partial observability, and tends to produce stable learning curves. These benefits come with two caveats. First, performance remains sensitive to the choice of clipping parameter, learning rate, and advantage normalization, poor choices can silently under-update the policy or, conversely, erode the proximal constraint. Second, PPO's on-policy nature increases data requirements relative to off-policy actor–critic methods, especially when environment interaction is expensive. In EV aggregation, PPO is a strong candidate when the action space is continuous (like power set-points or continuous bidding quantities) and when stability under changing market regimes is paramount, provided that trajectory data are sufficiently abundant either from high-fidelity simulators or carefully instrumented field trials.

2.2.3 Multi-Agent Reinforcement Learning (MARL)

Multi-Agent Reinforcement Learning generalizes RL to systems with multiple decision-makers whose interactions shape the dynamics and rewards. A common formalism is the decentralized partially observable MDP (Dec-POMDP), where each agent *i* receives private observations, selects actions under partial information, and contributes to joint rewards. A pragmatic and influential paradigm is centralized training with decentralized execution (CTDE): during training, agents (or a centralized critic) may access global state and other agents' actions to facilitate credit assignment and stabilize learning at execution time, each agent acts using only its local observation and its learned policy.

Algorithmic families include independent learners, each agent learns as if others were part of the environment, value-decomposition methods that factorize a joint action-value function into per-agent terms to enable scalable credit assignment, and centralized critics (such as actor–critic schemes where the critic conditions on joint information). Communication-enhanced policies and opponent-modeling are additional techniques for cooperative or competitive settings.

MARL is effective when the task's structure is inherently distributed and coordination is essential. In EV aggregation, vehicles, chargers, and the aggregator can be framed as interacting agents whose local constraints (state of charge, mobility, feeder limits) and economic objectives (market revenues, penalties) must be reconciled. CTDE allows learning global coordination strategies respecting network and market couplings while preserving decentralized implementability in the field, where communication may be limited or latency constrained.

The principal advantage of MARL is its fidelity to multi-actor realities: it captures strategic interactions, supports cooperative load-shaping, and can internalize network externalities that single-agent formulations must approximate. This expressiveness enables policies that are resilient to local perturbations and that exploit heterogeneity across agents, for instance by prioritizing flexible EVs when system conditions tighten. Against these benefits stand several challenges. From each agent's viewpoint, simultaneous learning by others induces non-stationarity, which can destabilize value estimation and impede convergence. Credit assignment determining which agent's action caused a change in global reward remains difficult at scale, even with value decomposition. Sample and computing demands can be substantial, as joint action spaces grow combinatorially, careful curriculum design, parameter sharing, and hierarchical decompositions are often required. For EV aggregation, MARL is attractive when distribution-network constraints and diverse user preferences are central and when the control architecture must remain decentralized. Nonetheless, it typically requires a carefully engineered simulator and thoughtful regularization to achieve reliable training.

DQN, PPO, and MARL occupy complementary positions on the RL design spectrum. DQN exemplifies value-based learning for discrete decisions and is appealing where action granularity can be coarsened without sacrificing performance like selecting among a small set of standardized bids or tariff responses. PPO represents a robust, implementation-friendly policy-gradient approach that natively handles continuous control and tends to deliver smooth, stable policies well matched to dispatching continuous charging power and shaping intra-day flexibility for energy and ancillary-service markets. MARL extends either value-based or policy-gradient cores to settings with many interacting agents, it is the natural choice when coordination among heterogeneous EVs, chargers, and feeders is integral to the problem definition, such as feeder-aware reserve provision or distribution-level congestion management coupled to market bidding.

In practice, a layered architecture can be effective: MARL or hierarchical MARL governs local coordination under network and mobility constraints, PPO optimizes continuous bidding and scheduling at the aggregator level given forecasts and risk preferences, and a DQN-like module maps discrete market states (e.g., scarcity conditions, price caps) to tactical overrides. Such hybridization leverages each method's strengths while mitigating their individual weaknesses.

Algorithm	**Type**	**Key Strengths**	**Limitations**	**Suitable Applications**
DQN	Value-based	Good for discrete, high-dimensional states; simple implementation	Inefficient in continuous actions; data-hungry	Games, simplified grid models
PPO	Policy-gradient	Stable, robust, works in continuous actions	Requires tuning; computationally heavier	Robotics, energy management, EV scheduling
MARL	Multi-agent	Captures interactions between agents; models cooperation/competition	Training instability, scalability issues	EV aggregation, traffic control, smart grids

Table 1 Reinforcement Learning Types Comparison

2.3 FUZZY LOGIC & EXPERT SYSTEMS

Fuzzy logic and expert systems emerged to address a limitation of classical, binary logic in representing human knowledge and reasoning under uncertainty. In many engineering and decision-making contexts particularly those involving linguistic rules, vague thresholds, heterogeneous data quality, or incomplete measurements crisp true/false propositions and sharp boundaries are ill-suited to capture the gradual, graded structure of human concepts (Klir & Yuan, 1995). Fuzzy logic introduces degrees of membership to model such gradualness: rather than assigning an element strictly to a set, it specifies a membership function (x)∈[0,1] that quantifies how strongly x belongs to the fuzzy set A. Building on this representation, fuzzy inference enables the manipulation of linguistic rules (if charging demand is high and price is low, then charge aggressively) with mathematically rigorous operators that generalize conjunction, disjunction, and implication (Ross, 2010).

Expert systems, in parallel, arose from the desire to encode specialist knowledge in rule-based programs that explain their conclusions. Classical expert systems use crisp predicates and logical inference (like forward or backward chaining). Fuzzy expert systems extend that paradigm by allowing rules whose antecedents and consequents are fuzzy propositions and by propagating partial truth values through inference. This fusion is especially compelling in energy applications where measurements are uncertain, context is dynamic, and meaningful heuristics are naturally expressed in linguistic form. In the context of this thesis, fuzzy systems provide a principled way to embed expert heuristics about mobility patterns, state-of-charge priorities, and tariff regimes, they also support data-driven adaptation when combined with learning methods.

2.3.1 Fuzzy Inference Systems (FIS)

The modern FIS traces to the first fuzzy controller constructed by (Mamdani & Assilian, 1975), which demonstrated that control rules stated by experts in natural language could drive real processes if encoded with fuzzy sets and inference. Two major families subsequently crystallized. Mamdani-type systems treat both antecedents and consequents as fuzzy sets, producing a fuzzy output that is “defuzzified”. Takagi–Sugeno (TS) systems use fuzzy antecedents but crisp, typically affine, consequents of the form $y = a_0 + \sum_i a_i x_i$ (Takagi & Sugeno, 1985). In both families, the computational pipeline comprises fuzzification of inputs, rule evaluation, aggregation across rules, and output synthesis.

FIS operate as universal function approximators under mild conditions, leveraging overlapping linguistic partitions and local rules to reconstruct complex nonlinear mappings. Intuitively, each rule encodes expert knowledge about a region of the input space, and the degree to which the current input matches that region determines the rule's activation. Aggregating the contributions of several partially active rules yields smooth, context-dependent outputs. Mamdani systems excel when interpretability is paramount because their consequents are themselves linguistic terms. TS systems trade some linguistic transparency for computational efficiency and convenient optimization properties, as their crisp consequents support straightforward least-squares estimation of local models and efficient real-time evaluation.

Knowledge-driven design starts from domain expertise: define linguistic variables (time-of-day, state-of-charge, real-time price…), craft membership functions to capture meaningful regimes, like peak hours or low price, and articulate rules consistent with operational policy. Data-driven tuning then adjusts membership function parameters and rule weights to reduce prediction or control error. For TS models, identification often uses clustering (like subtractive or fuzzy c-means) to propose rule antecedents and local linear regression to estimate consequents, for Mamdani systems, gradient-based or evolutionary strategies can shape membership parameters (Ross, 2010).

FIS offer a rare combination of interpretability and nonlinearity. They can encode qualitative expertise in a form auditable by stakeholders, and their graded reasoning naturally accommodates noisy sensors and incomplete information. Compared with many black-box models, they provide transparent rationales, rule activations and linguistic labels, that support explainable decision making in regulated energy settings. In practical deployments, TS systems achieve low latency because inference reduces to evaluating a small number of basic functions and affine consequents.

These advantages come with trade-offs. Hand-crafted rule bases can grow quickly with the number of inputs, leading to a combinatorial "rule explosion" that complicates maintenance and may erode interpretability. Parameter tuning is nontrivial when many overlapping membership functions interact, and without careful validation or regularization, overfitting can arise especially when rules are adapted to limited data. Furthermore, although FIS can approximate complex dynamics, their extrapolation beyond the range of training or expert knowledge is not guaranteed, and designing robust membership partitions in high-dimensional spaces remains a challenge (Klir & Yuan, 1995; Ross, 2010).

For EV aggregators, FIS enable the explicit encoding of heuristics that experts use daily: for example, if it is a weekday evening and public chargers near workplaces are saturated, expected charging demand at home hubs is high, or if real-time prices are low and state-of-charge is below a mobility threshold, prioritize charging despite moderate feeder loading. A Mamdani FIS can generate qualitative risk scores, like charging urgency, while a TS FIS can map inputs such as time-of-day, weather, mobility patterns, and tariff signals to continuous forecasts of charging demand or to bids in reserve markets. Because rule activations can be inspected, operators may trace why a specific forecast or bid was produced, improving trust and facilitating compliance reviews.

2.3.2 NEURO-FUZZY MODELS

Neuro-fuzzy systems integrate fuzzy inference with neural-network learning to achieve adaptive, data-driven tuning while retaining a rule-based structure. The canonical architecture is ANFIS (Adaptive-Network-Based Fuzzy Inference System), introduced by (Jang, 1993). ANFIS realizes a first-order TS FIS as a layered network: layer 1 parameterizes input membership functions; layer 2 computes rule firing strengths; layer 3 normalizes these strengths; layer 4 evaluates rule consequents (affine functions); and layer 5 aggregates outputs. This representation enables gradient-based learning of premise (membership) parameters and least-squares updates of consequent coefficients via a hybrid algorithm that alternates between the two steps.

The core rationale is to combine the universal approximation and interpretability of fuzzy rules with the powerful optimization machinery of neural networks. Membership functions serve as adaptive basis functions that carve the input space into soft regions, local linear consequents provide low-bias approximations within each region. Learning adjusts both the location and shape of these regions and the local models to minimize prediction error. Because rule activations are normalized, the overall mapping is a smooth convex combination of local models, which lends itself to stable training and efficient inference (Pedrycz & Gomide, 2007).

A central design issue is determining the number of rules. Clustering methods can initialize rule antecedents, after which training prunes or refines rules based on contribution to accuracy. Regularization of consequent parameters, like ridge penalties, and constraints on membership spreads help mitigate overfitting. Early stopping on a validation set and rule merging based on similarity can further control complexity. For nonstationary environments such as EV fleets whose behavior evolves with seasons or tariffs online or recursive learning variants update parameters as new data arrive, preserving adaptability without catastrophic drift.

Neuro-fuzzy models typically deliver higher predictive accuracy than static FIS because they learn both structure and parameters from data. They are sample-efficient compared with deep neural networks of similar capacity, and the resulting rule base remains, at least partially, interpretable: one can still inspect linguistic antecedents and local consequents. The training process is relatively stable thanks to convex subproblems for consequents and smooth membership functions, and the final models are fast enough for real-time forecasting and bidding.

However, the interpretability advantage can diminish as the number of rules increases or as membership functions become highly tuned to idiosyncratic patterns. Gradient-based adaptation may drift from the initial, semantically meaningful partitions, complicating human validation. Moreover, hybrid training introduces several hyperparameters (learning rates, regularization strengths, clustering radii) whose selection materially affects performance, suboptimal choices can yield overfitting or poor generalization. Finally, while ANFIS scales are better than hand-crafted FIS, very high-dimensional inputs can still provoke rule proliferation unless dimensionality-reduction or sparse rule induction is applied (Jang, 1993; Pedrycz & Gomide, 2007).

In the thesis context, neuro-fuzzy models are natural candidates for forecasting the charging demand of an EV aggregator across intraday horizons. Inputs may include temporal covariates (hour, weekday/holiday), mobility and fleet composition indicators, weather proxies, electricity prices, and historical charger occupancy. An ANFIS-style model can learn localized regimes such as commute-driven spikes or price-induced shifting and produce smooth, quickly computable forecasts needed for market bidding. Because the learned rules remain auditable linguistically (if hour is late evening and temperature is low then home-charging demand is high), the model's decisions can be justified to operators and regulators, while its adaptive training captures evolving usage patterns.

2.4 MACHINE LEARNING BASED APPROACHES

Machine Learning (ML) refers to a family of computational methods that infer patterns from data and use those patterns to make predictions or discover structure. At its core, ML operationalizes statistical learning: it posits a hypothesis class (a set of candidate functions), selects a loss function that quantifies predictive error, and searches for the function that minimizes expected loss with respect to the data-generating process. Generalization, the ability to perform well on unseen data, arises when the hypothesis class embodies suitable inductive biases, the training data are representative, and optimization is regularized to prevent overfitting (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2009).

In the context of this thesis the principal supervised task is to predict short-term charging demand at various temporal granularities. These forecasts inform bidding and scheduling decisions under market and network constraints. Complementing this predictive task, unsupervised learning assists in discovering latent structure: typical charging session archetypes, user segments, station-level clusters, and anomalous behavior that may degrade forecast accuracy or violate service commitments. Together, supervised and unsupervised approaches form a coherent toolkit for both operational forecasting and strategic portfolio management.

2.4.1 Supervised versus Unsupervised Learning

Supervised learning addresses problems where each observation pairs inputs with a target, ($\mathbf{x}_t$,y_t), and the goal is to learn a mapping $f : \mathrm{R}^p \to \mathrm{R}$ (regression) or $f : \mathrm{R}^p \to \{1,\ldots,K\}$ (classification) that minimizes an expected loss, often approximated by the empirical risk $(1/n)\sum_{t=1}^{n} \mathrm{l}(y_t, f(x_t))$. In EV demand forecasting, y_t may represent aggregate charging power for an aggregator at time t, and $\mathbf{x}_t$ can include calendar indicators, electricity prices, weather, fleet availability, mobility proxies, and lagged load features. The training objective is to maximize predictive accuracy while ensuring temporal robustness, which typically requires time-aware cross-validation, leakage prevention, and drift monitoring.

Unsupervised learning, by contrast, operates without labeled targets. Its aim is to discover structure in $\{\mathbf{x}_t\}$: clusters, manifolds, or low-dimensional embeddings. In this thesis, clustering can reveal distinct usage patterns across depots or customer cohorts, while dimensionality reduction and reconstruction-based methods support anomaly detection and feature learning. Although unsupervised outputs are not directly evaluated by target error, they are invaluable for model design like creating stratified training regimes, for feature engineering such as cluster memberships as predictors, and for operational insights.

The choice between supervised and unsupervised learning is determined by problem formulation and data availability. When reliable target labels exist and business value is tied to predictive accuracy, supervised learning is primary. When the goal is exploration, segmentation, or data quality control, unsupervised learning is complementary. In practice, both paradigms are often combined: unsupervised structure informs supervised models, and supervised residuals guide unsupervised anomaly screens.

2.4.2 Supervised Learning Models

2.4.2.1 Linear and Polynomial Regression

Classical linear regression assumes a linear relationship between predictors and the response,

$$\hat{y}_t = \beta_0 + \sum_{j=1}^{p} \beta_j x_{t,j}$$

with parameters estimated by Ordinary Least Squares (OLS),

$$\min_{\beta} \sum_{t=1}^{n} (y_t - \beta_0 - x_t^T \beta)^2$$

Regularized variants such as ridge and lasso add ℓ_2 or ℓ_1 penalties to stabilize estimates under multicollinearity and high-dimensionality. Polynomial regression augments $\mathbf{x}_t$ with nonlinear basis functions (squared and interaction terms) to capture curvature while retaining linear-in-parameters estimation.

Historically rooted in Gauss-Legendre least squares and later generalized through modern statistical learning theory (Hastie, Tibshirani, & Friedman, 2009), these models remain a robust baseline for EV load forecasting. They are computationally efficient, interpretable via coefficients and partial dependence, and well-suited to incorporating domain structure through engineered features. However, they rely on restrictive assumptions: linear or low-degree polynomial relations, homoscedastic errors, and limited interactions unless explicitly modeled. When charging dynamics are highly nonlinear, driven by thresholds (tariff blocks), saturations (charger capacity), or unobserved heterogeneity (user schedules), linear models may underfit even with careful feature design.

2.4.2.2 Random Forests (RF)

Random Forests are ensemble learners that aggregate predictions from many decision trees trained on bootstrap samples, with feature subsampling at each split to decorrelate trees (Breiman, 2001). For regression, the forest prediction is the average across B trees,

$$f(x) = (1/B) \sum_{b=1}^{B} T_b(x)$$

where each T_b is a regression tree. The bootstrapping and random feature selection reduce variance relative to a single deep tree while maintaining the ability to model nonlinearities and interactions automatically.

For EV demand prediction, Random Forests often perform strongly with minimal preprocessing: they handle mixed data types, accommodate non-additive effects, like interactions between weather and calendar, and provide measures of variable importance. Out-of-bag (OOB) error offers efficient internal validation, and quantile variants yield probabilistic forecasts. Their limitations stem from piecewise-constant fits that struggle to extrapolate beyond observed ranges and from reduced interpretability at the ensemble level. In very large datasets or with many trees, training and inference may also become computationally heavy, although parallelization alleviates this burden.

2.4.2.3 Gradient Boosting Machines (GBMs)

Gradient Boosting constructs an additive model by sequentially fitting weak learners (typically shallow trees) to the negative gradients of the loss function (Friedman, 2001). After M boosting rounds,

$$f_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

where each h_m is a tree fitted to current residuals, and γ_m is a learning-rate-scaled step size. Modern implementations, like XGBoost, LightGBM or CatBoost, introduce system-level and algorithmic optimizations that deliver state-of-the-art accuracy on tabular data.

In EV aggregation, GBMs are attractive for capturing complex, nonlinear dependencies among exogenous drivers (prices, temperature, calendar effects, mobility signals) and lags of the target series. With careful regularization such as, shrinkage, subsampling or tree depth constraints, they balance bias and variance and can produce calibrated probabilistic outputs via quantile or distributional loss functions. Their drawbacks include a sensitivity to hyperparameters, potential overfitting when boosting too many rounds without early stopping, and limited extrapolation beyond the convex hull of the training data. Model explainability requires post hoc tools, which, while powerful, add analytical overhead.

2.4.2.4 Neural Networks (ANNs and Deep Learning)

Artificial Neural Networks approximate functions by composing linear transformations with nonlinear activations. A feed-forward network with L hidden layers computes

$$a^{(l)} = \sigma\left(W^{(l)}a^{(l-1)} + b^{(l)}\right)$$

with $a^{(0)} = x$, and learns parameters $\{W^{(l)},b^{(l)}\}$ by minimizing a loss function using backpropagation and stochastic gradient descent (Goodfellow, Bengio, & Courville, 2016). Deep learning extends this paradigm with many layers and specialized architectures. For temporal EV demand, recurrent networks (LSTM/GRU), temporal convolutional networks (TCN), and attention-based models can represent long-range temporal dependencies and regime shifts.

Neural networks excel when relationships are highly nonlinear, interactions are ubiquitous, and rich auxiliary data are available. They implicitly learn features, reducing reliance on manual engineering, and can produce multi-horizon forecasts in a single forward pass. Their disadvantages are well known, they require substantial data and careful regularization, training can be compute-intensive, and interpretability is limited without dedicated explainability methods. Sensitivity to dataset shift is also a concern, concept drift in charging behavior can degrade performance unless models are updated or adapted.

2.4.3 UNSUPERVISED LEARNING MODELS

2.4.3.1 Clustering: K-Means, DBSCAN, and Hierarchical Clustering

Clustering aims to partition observations into groups that are internally coherent and externally distinct. K-Means seeks K centroids {μk} minimizing within-cluster sum of squares,

$$\min_{\{C_k\},\{\mu_k\}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|\left|x_i - \mu_k\right|\right|^2$$

usually optimized by Lloyd's iterative assignment-and-update procedure (Lloyd, 1982). It is efficient and effective for roughly spherical, similarly sized clusters after appropriate scaling. In EV analytics, K-Means can reveal canonical daily load shapes or typical charging-session profiles that later serve as features or priors in forecasting.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) defines clusters as regions of high point density separated by sparse regions (Ester, Kriegel, Sander, & Xu, 1996). With parameters ε (neighborhood radius) and minPts (minimum neighbors), it discovers arbitrarily shaped clusters and identifies outliers as noise. This makes DBSCAN suitable for identifying unusual charging events, rare station behaviors, or spatial-temporal hotspots without pre-specifying K.

Hierarchical clustering builds a tree (dendrogram) by iteratively merging (agglomerative) or splitting clusters according to a linkage criterion (single, complete, average or Ward) (Murtagh & Contreras, 2012). Its multiscale perspective is valuable for EV portfolios spanning heterogeneous sites, analysts can cut the dendrogram at different levels to obtain coarse or fine segmentations, and the tree structure aids interpretability. Through these methods, careful preprocessing (scaling, transformation, and choice of distance) materially influences results. Cluster stability checks and external validation are recommended before downstream use.

2.4.3.2 Autoencoders

Autoencoders learn compact representations by training an encoder f_ϑ and decoder g_ϕ to minimize reconstruction error,

$$\min_{\theta,\phi} \sum_{i=1}^{n} | x_i - g_\phi\big(f_\theta(x_i)\big)|^2$$

Undercomplete architectures, bottleneck latent dimension smaller than input, or regularization (sparsity, denoising) force the model to capture salient structure rather than memorize inputs. Variational Autoencoders (VAE) introduce a probabilistic latent variable with a Kullback–Leibler regularization term to encourage a well-structured latent space (Kingma & Welling, 2014).

In EV aggregation, autoencoders can denoise metering data, embed high-dimensional temporal features into compact vectors for downstream forecasting, and support anomaly detection by flagging observations with large reconstruction errors. Their strengths lie in flexibility and scalability to complex, nonlinear manifolds. Limitations include sensitivity to architectural choices and training objectives, interpretability is indirect, and reconstructions may be overly smooth, potentially masking sharp events unless the loss and architecture are tailored to the application.

2.5 AI-POWERED SOFTWARE TOOLS FOR DEMAND FORECASTING

This section reviews the principal software tools used to operationalize AI for short-term and day-ahead demand forecasting, focusing on their modeling scope, practical workflows, and suitability for an EV-aggregator context. The emphasis is on capabilities that matter for aggregator portfolios: handling strong seasonality (hour-of-day/day-of-week), incorporating exogenous drivers (prices, weather, mobility), scaling to many chargers, and deploying forecasts reliably into market-bidding pipelines.

2.5.1 TENSORFLOW AND PYTORCH (DEEP LEARNING FRAMEWORKS)

What they are and why they work. TensorFlow and PyTorch are general-purpose deep learning frameworks that provide automatic differentiation, GPU/TPU acceleration, and high-level APIs to implement sequence models (RNN/LSTM/GRU), temporal convolutional networks (TCN), transformers, and hybrids. For forecasting, they enable end-to-end learning of nonlinear temporal dependencies and interactions with exogenous signals, which is advantageous when EV charging demand exhibits regime shifts, complex calendar effects, or weather sensitivities that are hard to encode in linear models. Official tutorials illustrate time-series workflows (windowing, multi-step forecasts) and model training patterns (TensorFlow Team, 2025).

Their main advantages are modeling flexibility, hardware acceleration, and rich ecosystems, which facilitate large-scale training and real-time inference. Limitations include higher engineering overhead (feature pipelines, hyperparameter tuning, and monitoring) and the need for careful regularization to avoid overfitting when training data are limited at the charger level.

2.5.2 SCIKIT-LEARN (CLASSICAL ML TOOLKIT)

Scikit-learn offers a mature suite of classical ML estimators (regularized linear models, tree ensembles, gradient boosting) plus preprocessing and model-selection utilities. Although it does not provide native forecasting estimators, lagged-feature formulations like autoregression via feature engineering, allow one to use robust tabular learners as strong baselines. Its consistent API for pipelines and cross-validation makes it ideal for rapid experimentation and interpretable benchmarks.

When station-level histories are short or when interpretability is paramount, tree-based models with engineered lags and calendar features offer competitive accuracy with low operational complexity (Scikit-Learn Developers, 2025).

2.5.3 STATSMODELS (STATISTICAL FORECASTING)

Statsmodels implements classical time-series models ARIMA/SARIMA/SARIMAX for seasonal dynamics with exogenous regressors, exponential smoothing/ETS for trend-seasonality decomposition, and VAR/VECM for multivariate dependencies. These models embody well-understood stochastic assumptions and likelihood-based estimation, enabling inference (confidence intervals, diagnostics) valuable for operators and regulators (Statsmodels Developers, 2025).

Strengths include interpretability, principled uncertainty quantification, and strong performance on stable seasonal patterns. Limitations include reduced flexibility for nonlinear effects, challenges with large cross-sectional hierarchies, and sensitivity to structural breaks which is common when fleet composition or charging policies change.

2.5.4 H2O.AI (AUTOML FOR FORECASTING)

H2O AutoML (open-source) and H2O Driverless AI (commercial) automate feature engineering, model selection, and ensembling for tabular/temporal data. Driverless AI includes dedicated time-series workflows (rolling-window training, time-aware cross-validation, test-time augmentation), yielding strong baselines with minimal hand-tuning and built-in explanations.

AutoML is effective for rapidly establishing portfolio-wide baselines and for sites with heterogeneous data quality. It reduces manual iteration while providing reproducible pipelines suitable for MLOps integration (H2O.ai, 2025).

2.5.5 CLOUD PLATFORMS: AZURE ML, GOOGLE VERTEX AI, AND AWS SAGEMAKER

Managed cloud platforms provide experiment tracking, scalable training, AutoML for time series, reproducible pipelines, and online/batch serving. Azure Machine Learning includes time-series-specific featurization, model sweeping, and evaluation components. Google's Vertex AI offers a forecasting workflow that integrates dataset preparation, training, evaluation, and deployment with MLOps primitives. AWS SageMaker supplies built-in forecasting algorithms such as DeepAR and CNN-based quantile regression, and it integrates with Amazon Forecast for turnkey pipelines.

For production bidding, these platforms facilitate secure data ingestion (telemetry, prices, and weather), automated retraining, A/B testing across sites, and low-latency inference endpoints. Trade-offs include cost management, data-sovereignty constraints, and potential vendor lock-in (Microsoft, 2025; Google Cloud, 2025; Amazon Web Services, 2025).

2.6 CHALLENGES IN AI-BASED DEMAND FORECASTING

Accurate forecasting of electric-vehicle (EV) charging demand is essential for aggregators to construct profitable and reliable bids in energy and ancillary-service markets. Three intertwined challenges data availability and quality, uncertainty in charging behavior, and scalability, constrain the performance and deployment of state-of-the-art AI models.

2.6.1 DATA AVAILABILITY AND QUALITY

Effective AI models require granular, well-labeled, and representative datasets. In EV aggregation, inputs span charging-session telemetry, site and feeder constraints, fleet attributes, prices/tariffs, and exogenous variables such as weather and events. In practice, these data are fragmented across stakeholders and systems, with heterogeneous schemas and sampling intervals, frequent missingness, and misaligned timestamps, all of which degrade feature engineering, bias models toward data-rich sites, and hinder transferability across regions (International Energy Agency, 2024). Privacy and commercial-sensitivity constraints further limit data access, underscoring the need for privacy-preserving data governance and standardized data contracts to support robust preprocessing, temporal alignment, and bias-aware validation.

2.6.2 UNCERTAINTY IN EV CHARGING BEHAVIOR

Aggregator-level demand exhibits pronounced variability driven by heterogeneous mobility patterns, state-of-charge on arrival, tariff design, real-time prices, weather, holidays, and the evolving spatial distribution of chargers. High-resolution, bottom-up analyses show that both the magnitude and timing of EV loads can shift substantially across locations and seasons, with local peaks that matter for distribution operations and market. Consequently, forecast products should quantify uncertainty, not just provide point estimates, through probabilistic methods and be evaluated with proper scoring rules and calibration diagnostics common in the probabilistic load-forecasting literature (Hong & Fan, 2016).

A further complication is non-stationarity (concept drift): policy changes, adoption growth, infrastructure expansion, and user responses to prices alter the data-generating process over time. Models must therefore adapt online or via periodic updates, with drift detection and hierarchical reconciliation to maintain accuracy across portfolio levels (Xiang, Zhen, Peng, Zhang, & Pu, 2023).

2.6.3 SCALABILITY FOR REAL-TIME, PORTFOLIO-SCALE OPERATIONS

Production forecasting is a streaming, near-real-time task: features and predictions must update at minute-to-hour horizons across thousands of connectors and multiple market zones, under strict latency budgets. This creates three scaling pressures. First, computational scaling: multivariate spatiotemporal models can be costly to retrain frequently, adaptive learners that detect and accommodate drift mitigate retraining overhead while sustaining accuracy (2023). Second, data-pipeline scaling: feature engineering must handle late-arriving data and schema evolution without information leakage, favoring event-time processing, feature stores, and idempotent backfills (International Energy Agency, 2024). Third, organizational scaling: as infrastructure and fleets expand, the number of forecasted series and the pace of change rise, necessitating automated monitoring (data/feature/forecast drift), alerting, and safe rollbacks integrated into MLOps.

Chapter 3. STATE OF THE ART

Short-term electricity demand forecasting (STLF) underpins secure system operation, congestion management, and market bidding. Over the last decade, tree-based ensemble methods especially Random Forests (RF) have emerged as robust competitors to classical time-series models and to more complex neural architectures. This chapter reviews the evidence, progressing from early comparisons against statistical baselines to broad, cross-country studies versus modern machine-learning (ML) and deep-learning (DL) approaches, and finally to applications directly relevant to EV charging and, by extension, EV-aggregator demand forecasting.

3.1 WHY RANDOM FOREST IS A STRONG BASELINE FOR LOAD FORECASTING

RF is an ensemble of decision trees built on bootstrap samples with random feature subsetting at each split, averaging across many decorrelated trees reduces variance and yields robust generalization with minimal tuning (typically the number of trees and the number of features per split) (Breiman, 2001). These properties make RF attractive for electric-load and EV-charging forecasting, where nonlinearities, interactions among weather, calendar, and behavioral variables, and distributional shifts are common. RF natively handles mixed predictors, is resilient to outliers, and offers diagnostics such as permutation-based variable importance that support interpretability, valuable in market and operations settings where feature attributions inform decisions.

Early comparative evidence found that RF achieved STLF accuracy on par with feed-forward neural networks and clearly superior to ARIMA and Holt–Winters exponential smoothing, while being easier to tune and less prone to overfitting due to variance reduction from tree averaging. This result, together with the method's limited hyperparameter burden, helped establish RF as a practical baseline for day-ahead system-load prediction.

3.2 EVIDENCE FROM POWER-SYSTEM SHORT-TERM LOAD FORECASTING

Building on early results, a comprehensive evaluation by (Dudek, 2022) compared local versus global RF training and multiple input encodings across four national STLF problems (Poland, Great Britain, France, Germany). With careful input design that explicitly encodes intra-day and intra-week seasonal patterns and a global training regime enhanced by calendar features, the proposed RF delivered the lowest MAPE in three of the four countries and remained highly competitive in France. Pairwise Giacomini–White tests confirmed that many of the improvements were statistically significant.

Two methodological insights are particularly relevant for EV-demand contexts. First, encoding multiple seasonalities directly in the features simplifies the learning problem for nonparametric models like RF. Second, training a single global model across many days augmented with calendar variables improves data efficiency and generalization. Despite its simplicity and few hyperparameters, RF competes credibly with sophisticated DL variants when supplied with informative, well-engineered inputs (Dudek, 2022).

In sum, tree ensembles remain state-of-the-art contenders for STLF: when inputs exploit calendar and seasonal structure, training leverages cross-series regularities via global modeling and the objective is reliable accuracy with modest engineering effort, conditions common to EV-aggregator operations that must scale models across many feeders or depots.

3.3 APPLICATIONS TO EV-RELATED DEMAND AND CHARGING

RF has also been evaluated specifically for EV-charging demand. (Khan, et al., 2023) studied 15-minute EV load across multiple spatial resolutions and found that RF outperformed a multilayer ANN at several aggregation levels, the authors emphasized the importance of calendar and user-presence features at small scales, while RF’s performance improved with aggregation, consistent with the variance-reduction benefits of ensembles.

From an energy-procurement and risk perspective, (Ostermann & Haug, 2024) compared probabilistic models for EV charging using hundreds of thousands of sessions at more than 500 sites in Germany. Ensemble trees (Bagging, AdaBoost, RF) delivered strong point and quantile accuracy at higher aggregation levels (such as portfolio or TSO zones), with RF achieving low error and high R^2 alongside narrow prediction intervals, evidence that RF scales well and supports risk-aware planning via quantile forecasts in operational settings.

At single-station and fleet level, results are mixed but informative. (Deb, Kalam, & Agalgaonkar, 2022) built an RF-based framework to forecast charging demand for an electric-bus fleet in Helsinki, showing effective short-term predictions for operational planning. Conversely, in a day-ahead charging-station case study, a thorough comparison reported gradient-boosted trees slightly outperforming other regressors on power consumption, while RF still tracked the target closely illustrating that boosted trees may dominate at highly granular horizons, though RF remains competitive and easier to tune (Amezquita, Rojas, & Arango, 2024).

In weather-sensitive contexts similar to EV-charging demand, RF remains highly competitive while enabling post-hoc explanation through permutation importance or SHAP. Empirical analyses highlight irradiance and daylight duration as dominant drivers of hourly consumption, aligning with operational intuition and facilitating transparent communication of forecast drivers to market stakeholders (Qu, Kou, & Zhang, 2025).

3.4 STRENGTHS, LIMITATIONS, AND BEST PRACTICES FOR EV-AGGREGATOR USE

The main strengths of RF for EV and power-system demand forecasting are robustness to noise and outliers, the ability to capture nonlinear interactions without bespoke feature transforms, scalability to large predictor sets (calendar, weather, event flags, mobility proxies) and modest hyperparameter sensitivity, often many trees and small leaves, the most consequential choice is the number of features per split. Empirically, RF excels as aggregation increases like station to depot to portfolio, and when calendar and seasonal pattern encodings are used. It is also comparatively data-efficient when per-site history is short, a common situation for new charging assets (Breiman, 2001; Dudek, 2022; Ostermann & Haug, 2024).

Limitations are equally clear. RF produces step-wise predictions and cannot extrapolate beyond the convex hull of observed features, temporal dynamics are captured indirectly through lagged, seasonal, and calendar features rather than explicit sequence modeling. At very fine granularity, like single charging stations with volatile sessions, boosted trees often edge out RF in point accuracy, and specialized deep models can win when long sequences and rich exogenous context are available. Finally, native RF is a point forecaster, quantile or interval forecasts require adaptations, although practice shows that tree ensembles still yield competitive probabilistic accuracy at higher aggregation levels (Amezquita, Rojas, & Arango, 2024; Dudek, 2022; Ostermann & Haug, 2024).

For an EV-aggregator bidding framework, a strong and pragmatic setup consists of: a global or extended-global RF trained across sites or depots to leverage cross-sectional regularities, a feature design that encodes intra-day and weekly seasonality, holiday effects, recent lags, and weather, a portfolio-level and node-level models to exploit aggregation benefits while retaining local signal and a probabilistic outputs via quantile RF or ensemble bootstrapping, enabling risk-aware bids and reserve offers. RF's interpretability supports model governance and market compliance by explaining forecast changes due to weather or calendar shifts (Dudek, 2022; Ostermann & Haug, 2024).

Chapter 4. RESEARCH PROBLEM AND APPROACH

4.1 PROBLEM STATEMENT AND USE CASE

This thesis investigates short-term forecasting of the aggregated charging demand at an electric-vehicle (EV) aggregation charging station. The practical aim is to generate accurate day-ahead predictions of hourly energy (kWh) so that operators can plan capacity, schedule maintenance and load management, coordinate with counterparties, and monitor operational risks in advance. As EV adoption grows, the ability to anticipate the next day's load profile at an hourly resolution becomes increasingly consequential for local operations and for the wider system, given the contribution of EV charging to net-load patterns and flexibility provision (International Energy Agency, 2024; Muratori, 2018).

4.1.1 FORMAL FORECASTING TASK

Formally, the task is defined as producing a 24-dimensional vector of point forecasts $(\widehat{y_{t+1}}, \dots, \widehat{y_{t+24}})$ each day, where $\widehat{y_{t+h}}$ denotes the predicted total energy demanded (kWh) by the aggregated station in hour $t + h$. The temporal granularity is hourly, and the horizon is day-ahead. Forecasts are issued once per day (point forecasts only) at a fixed cut-off that aligns with the operator's internal planning cycle. The information set comprises historical aggregated load, calendar structure capturing intraday and intraweek regularities, and exogenous temperature signals represented through lagged transformations. The model therefore exploits short, daily, and weekly-scale dependence in the series while allowing weather-sensitive variation to be reflected in a parsimonious way.

4.1.2 OPERATIONAL CONTEXT

Within the station's day-ahead planning workflow, the forecast plays several roles. First, it informs capacity management by indicating expected peak hours and troughs, which supports scheduling of controllable charging policies and preventive load-smoothing actions. Second, it provides a forward view to coordinate routine activities, such as maintenance windows, software updates, and staffing, away from projected peaks. Third, the predicted profile is used as an early-warning signal: large deviations from typical patterns prompt diagnostic checks on data feeds, metering, or unusual user behavior. Finally, although this dissertation does not address bidding, the same forecast stream is readily consumable by downstream decision tools for energy planning in other contexts. The thesis, however, limits itself to the construction and validation of the forecasting component.

4.2 MARKET RELEVANCE AND STAKEHOLDER VALUE

Reliable day-ahead forecasts at the station-portfolio level create value across the ecosystem. For aggregators and charging-station operators, they reduce operational uncertainty, enabling better scheduling of charging policies and targeted peak reduction. Retailers and balance-responsible parties, where they interact with the station, benefit from improved predictability of load that facilitates hedging and internal risk management. Transmission and distribution system operators gain indirect advantages from smoother net-load trajectories and enhanced visibility of flexible demand, factors that contribute to secure system operation and more efficient planning of network resources (International Energy Agency, 2024; Muratori, 2018). In sum, even though this work concentrates on the forecasting layer, the resulting accuracy gains translate into tangible operational and system-level benefits.

4.3 RESEARCH QUESTIONS AND HYPOTHESES

The study is guided by two questions. **RQ1:** To what extent can a Random Forest (RF) model, built on lagged load, calendar structure, and temperature-based exogenous signals, accurately forecast day-ahead hourly demand for an EV aggregation station? The corresponding hypothesis is that a RF model using lagged load, calendar structure, and temperature-based features will be able to achieve lower error than a classical statistical baseline (SARIMA) for day-ahead hourly load. **RQ2:** Which components of the information set contribute most to predictive performance? The expectation is that short, daily, and weekly-lag features dominate, with calendar encodings and temperature lags providing incremental improvements during weather-sensitive periods.

4.4 OBJECTIVES

The scientific objective is to develop and validate a transparent, data-driven forecasting model for day-ahead hourly demand that demonstrates measurable improvement over a classical statistical benchmark. Equally important is to quantify where that improvement comes from by examining feature contributions and error profiles. The practical objective is to deliver an operationally simple pipeline that can be executed daily with modest computational effort, integrates external temperature information without heavy preprocessing, and exposes model outputs that are interpretable for practitioners responsible for station operations.

4.5 METHODOLOGICAL APPROACH

4.5.1 OVERALL PIPELINE

The methodological pipeline proceeds from data acquisition to evaluation as a connected process. It begins with real-world historical charging-demand data aggregated at the station/portfolio level and supplemented with hourly ambient temperature as an exogenous regressor. From these inputs, the study constructs a feature set that encodes temporal dependence and regularity: lagged demand values represent short-term dynamics as well as daily and weekly recurrences (calendar variables capture hour-of-day, day-of-week, month, and weekend effects) and lagged temperature values allow weather sensitivity to enter the model in a stable manner. On this foundation, the primary model is a Random Forest regressor trained to map the engineered features to next-day hourly energy. To contextualize performance, a baseline SARIMA model is fitted to the same target series, following standard practice in short-term load forecasting. The training process is strictly chronological to prevent leakage, and hyperparameters for the RF are selected on a validation slice carved from the training period.

The pipeline culminates with the error metrics, forecast accuracy is assessed with three complementary measures: Mean Absolute Error (MAE) to quantify the typical magnitude of errors in the same physical units (kWh), the coefficient of determination (R^2) to indicate variance explained relative to a mean-only benchmark, and the Normalized MAE (NMAE) to express error as a unitless fraction of average load, enabling scale-free comparison across periods or sites.

Mean Absolute Error (MAE). Measures average absolute deviation between forecasts and observations (lower is better). Units: kWh.

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

Coefficient of determination (R^2). Proportion of variance in the observations explained by the forecasts relative to using the sample mean. $R^2 = 1$ is perfect, $R^2 = 0$ matches the mean predictor, and it can be negative if forecasts are worse than predicting $\bar{y}$.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

Normalized MAE (NMAE). MAE scaled by the average observed load. Unitless and often reported as a percentage by multiplying by 100.

$$\mathrm{NMAE} = \frac{\frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\frac{1}{n}\sum_{t=1}^{n}y_t}$$

Notation: y_t is the observed load, $\bar{y}_t$ the forecast, $\bar{y}$ the sample mean of y_t, and n the number of forecasted hours in the test set.

4.5.2 VALIDATION DESIGN

Model development follows a blocked chronological train/test split that mirrors deployment. The time series is partitioned into an initial training window and a subsequent, contiguous test window, only observations up to the training end are visible during model selection. All hyperparameter tuning is conducted strictly within the training window, using time-ordered resampling to assess candidate settings without leaking future information. After selecting hyperparameters, the Random Forest is refit on the full training period, and forecasts are then generated for the entire test window, respecting the operational cut-off (one 24-hour vector of hourly predictions issued per day).

To prevent leakage, every transformation is estimated on training data only and applied forward: lagged features are constructed using past values exclusively, any scaling or imputation parameters are fitted on the training window and then carried to the test window unchanged. Performance on the held-out test window is reported once, using MAE, R^2, and NMAE, and the test set is never used for tuning or model selection. This design provides an unbiased estimate of out-of-sample accuracy under stable operating conditions.

Chapter 5. DATA AND MODEL DEVELOPMENT

The forecasting task is defined as short-term prediction of the aggregated electric-vehicle (EV) charging demand at an hourly cadence. The target variable y_t is the total energy charged during hour t (kWh) across all active sessions in the fleet. This choice keeps the target consistent with the construction of the source dataset, which attributes energy to hourly buckets per session. The forecasting horizon is day-ahead, $h = 1,\ldots,24$ hours ahead. The spatial granularity is the aggregated fleet formed by all sessions and there is no vehicle-to-grid (V2G) discharging occurs in the dataset.

5.1 DATASETS AND SOURCES

5.1.1 AGGREGATOR TELEMETRY AND CHARGING SESSIONS

This study employs the open dataset “Electric vehicle charging dataset with 35,000 charging sessions from 12 residential locations in Norway”, which documents residential charging behavior over the period 6 February 2018 to 5 August 2021. The release contains more than 35,000 sessions across twelve locations in a mature EV market and was curated to support reproducible analyses of charging demand (Sørensen, Sartori, Lindberg, & Andresen, 2024). The accompanying data article explains the rationale for the release and situates it within a methodological effort to generate “complete” charging traces suitable for system studies.

The data is organized as a progression from observed session logs to derived per-user attributes and, finally, to per-hour attributions. The starting point is a session-level table exported from charge-point operators that records the location and user identifier, a unique session identifier, the plug-in and plug-out timestamps, the total connection time, and the energy delivered during the session. Building on these observations, the authors estimate two latent technical parameters for each user, typical charging power and effective battery capacity, by fitting to historical behavior so that subsequent reconstructions reflect realistic user-specific constraints. Using those per-user parameters, each observed session is then translated into an internal timeline that is consistent with the connection window and the empirically inferred power/capacity limits. For every session, the reconstruction yields the active charging time, the change in state of charge (SoC) and its starting level, the length of idle periods while connected, and a flag that marks sessions with negligible flexibility. In the final step, the reconstructed session trajectories are downscaled onto an hourly grid, allocating the delivered energy to the hours in which the vehicle was connected and actively charging, the same hourly panel also contains SoC diagnostics for each hour.

The present analysis aggregates this per-session, per-hour panel into a single fleet-level hourly series, y_t, representing the total energy charged in hour t. In other words, for each clock hour all concurrent session attributions are summed to form the target used for forecasting. The dataset contains charging only, vehicle-to-grid discharging does not occur so y_t is non-negative by construction. Timestamps are retained in local Norwegian time to remain consistent with the source and to facilitate alignment with hourly exogenous variables (temperature and calendar effects) introduced later in this chapter.

5.1.2 EXOGENOUS DATA

Hourly 2-m air temperature observations were obtained from the national meteorological service via the FROST application programming interface and merged with the load series by timestamp. Only historical values were used, no forecasted weather products were incorporated. All meteorological timestamps were retained in Norwegian local time (CET/CEST), consistent with the load data, so that hourly alignment required no temporal interpolation beyond the native resolution. Temperature subsequently enters the model as an exogenous regressor, both contemporaneously and through short seasonal lags, reflecting its role as a proxy for weather-driven variation in charging behavior (Norwegian Meteorological Institute).

Calendar structure was encoded through deterministic indicators that are known at prediction time. Specifically, the feature set includes hour-of-day, day-of-week, a weekend flag, the month of year, and a binary indicator for official Norwegian public holidays. These variables capture recurring intra-day and weekly patterns and isolate holiday effects without introducing look-ahead, since the calendar is fully specified ex ante.

5.2 DATA PREPARATION

5.2.1 CLEANING, MISSING DATA, AND OUTLIER TREATMENT

No missing timestamps are present after aggregation to an hourly cadence. Outliers in y_t are screened using an interquartile-range (IQR) rule. Sensitivity to the multiplier k in [0, 10] is explored (results reported in Chapter 6), and the final setting is fixed prior to model selection to avoid bias.

All sources are hourly and time-stamped in Norwegian local time (CET/CEST). Weather is aligned to load via an inner time join at the hour mark. Because both series are observational and synchronized hourly, no temporal interpolation is required.

5.2.2 FEATURE ENGINEERING

To capture autocorrelation and seasonality, the following features are used:

- Lag features of the target y_t: y_{t-1}, y_{t-24}, y_{t-168}.
- Rolling statistics: 24-hour and 168-hour rolling mean and standard deviation of y_t (trailing windows, past-only).
- Weather features: air temperature at t, and lagged values at $t-1$ and $t-24$.
- Calendar features: hour-of-day (0–23), day-of-week (0–6), weekend flag, holiday flag, and month.

All engineered features are strictly constructed only from information available up to time t.

5.2.3 LEAKAGE CHECKS

Strict temporal joins (no look-ahead) are enforced and all rolling windows are constructed in a one-sided (past-only) fashion. Timestamps are kept in Norwegian local time (CET/CEST), daylight-saving transitions are handled by preserving the platform's native treatment of repeated/missing hours, ensuring features and target remain aligned within local time.

5.3 EXPERIMENTAL DESIGN

5.3.1 TRAIN/VALIDATION/TEST SEGMENTATION (TIME-AWARE)

Data is split into chronological blocks:

- Training block: 2018-02-06 to 2020-02-05
- Validation block: 2020-02-06 to 2020-08-05
- Test block: 2020-08-06 to 2021-08-05

Hyperparameters are selected once using the validation block, the test block remains untouched until the final evaluation to avoid optimistic bias.

5.3.2 BASELINE MODEL (SARIMA)

A seasonal ARIMA model was fitted to the aggregated hourly series as a classical baseline, using a daily seasonal period s = 24. Orders $(p,d,q)\times(P,D,Q)_{24}$ are selected by information criteria (AIC) over a modest grid with $d,D \in \{0,1\}$, $p,q \in \{0,1,2\}$, and $P,Q \in \{0,1\}$, applied only on the training + validation data, and forecast on the test block. This establishes a transparent statistical benchmark (2021).

5.3.3 EVALUATION METRICS (POINT FORECASTS)

Evaluation metrics. Evaluation follows the definitions provided in 4.5.1. Performance is summarized by MAE, R^2, and normalized MAE (nMAE) on the held-out test block (2020-08-06–2021-08-05). To prevent instability of percentage errors at very low load, percentage metrics are not emphasized, near-zero hours are handled as described in 4.5.1. All model selection uses MAE on the validation block.

5.4 RANDOM FOREST SPECIFICATION

5.4.1 HYPERPARAMETERS AND SEARCH STRATEGY

The RandomForestRegressor implementation from scikit-learn (version 1.6.1) was employed with random_state = 42 and n_jobs = −1. The base configuration is:

Hyperparameter	**Value**	**Explanation**
n_estimators	100	Number of decision trees in the forest. A larger number reduces variance but increases computation time.
max_depth	10	Maximum depth of each tree. Limits complexity to prevent overfitting.
max_features	1.0	Fraction of features considered at each split (1.0 = all features).
min_samples_leaf	1	Minimum number of samples required in a terminal leaf node.
min_samples_split	2	Minimum number of samples required to split an internal node.
bootstrap	True	Whether bootstrap samples (random sampling **with replacement**) are used when building trees.
max_samples	None	If bootstrap = True, specifies the fraction of the training set to draw per tree (None = full sample).

Table 2 Random Forest base hyperparameters

Hyperparameters are tuned via random search, scoring MAE on the validation block. The search space is:

- n_estimators ∈ {300, 500, 800, 1000}
- max_depth ∈ {None, 12, 16, 24, 32}
- min_samples_leaf ∈ {1, 2, 5, 10, 20}
- max_features ∈ {'sqrt', 0.3, 0.5, 0.7, 1.0}
- bootstrap ∈ {True, False}

When bootstrap = True, the hyperparameter max_samples was allowed to take values in {None, 0.7, 0.9}. A total of 50 configurations were drawn at random, the best-performing model was refitted on the combined training and validation blocks, and its performance was subsequently evaluated once on the held-out test block.

5.5 IMPLEMENTATION DETAILS

All experiments were conducted in Python (version 3.10.12) using standard scientific libraries: NumPy (2.2.4), pandas (2.2.3), scikit-learn (1.6.1), and Matplotlib (3.10.3). Computations were executed on a conventional CPU workstation, and both the Random Forest and SARIMA models were trained in CPU-bound mode without the need for specialized hardware acceleration. To ensure reproducibility, random seeds were fixed at 42 for all stochastic elements of the pipeline. The project followed a modular organization: raw and processed datasets were stored separately, feature engineering routines were isolated in dedicated scripts, models were implemented and persisted within a structured module, and evaluation metrics and visualizations were maintained in a distinct evaluation layer. Configuration files specifying data splits and hyperparameter ranges were externalized to human-readable formats (YAML), and the full software environment was captured in a requirements.txt file. Trained models, processed data, and evaluation outputs were versioned with timestamped identifiers to allow experiments to be replicated exactly.

5.6 MODEL INTERPRETATION

5.6.1 FEATURE IMPORTANCE

Feature importance was examined to interpret the behavior of the Random Forest model. In addition to the model's impurity-based importance measures (mean decrease in impurity), permutation importance was computed on the validation block. The latter quantifies the increase in forecast error when the values of a given feature are randomly permuted and is less prone to bias toward high-cardinality or high-variance predictors. Both perspectives are reported, although greater emphasis is placed on permutation importance in the subsequent discussion.

5.6.2 ERROR DECOMPOSITION BY REGIMES

Although not central to the reported results, a standard diagnostic can be applied in which errors are disaggregated by context. In particular, MAE and nMAE can be calculated separately by hour-of-day, weekday versus weekend, holiday versus non-holiday, and by terciles of temperature. Such stratification highlights systematic under- or over-prediction patterns, for example around morning or evening charging peaks or during cold spells, and can therefore inform future refinements to the feature set.

Chapter 6. RESULTS AND DISCUSSION

6.1 SENSITIVITY TO THE IQR OUTLIER AND SELECTION OF K

Before reporting headline forecasting accuracy, the effect of the IQR-based outlier policy (Section 5.2.1) on sample size and performance is quantified. Because the filter changes the estimation and evaluation samples, the choice of the multiplier k can materially affect both the learning signal and the reported metrics. Establishing and freezing k prior to model comparison avoids ex-post selection bias and ensures that all subsequent results are comparable and replicable.

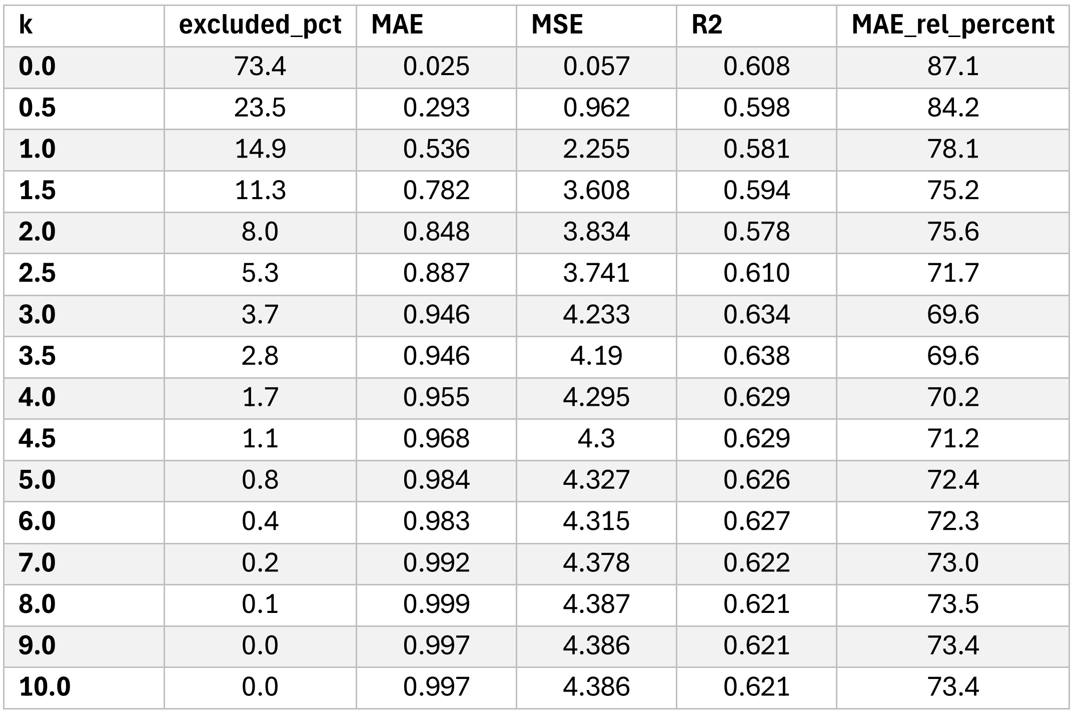
6.1.1 EXPERIMENTAL SETUP

Quartiles Q_1 and Q_3 and IQR = Q_3 - Q_1 were computed on the training block only (2018-02-06 to 2020-02-05). For each k∈{0,0.5,1,…,10}, hours with $y_t \in [Q_1 - k \cdot \text{IQR},\ Q_3 + k \cdot \text{IQR}]$ were removed consistently from the training and validation blocks. A baseline RandomForestRegressor with the default hyperparameters defined in Section 5.4 was fitted on the cleaned training data and evaluated on the cleaned validation block. The following metrics were recorded: mean absolute error (MAE), mean squared error (MSE), R^2, and normalized MAE (denoted MAE$_{\text{rel}}$, i.e., MAE/$\bar{y}$ x 100%). The percentage of excluded hours was also computed.

A compact summary of the grid appears in Table 3 Sensitivity of validation accuracy to the IQR outlier multiplier, and the principal relationships are visualized in Figure 1 Excluded observations (%) as a function of k and on Figure 2 Normalized MAE (%) on the validation block vs k. Monthly diagnostics for the selected policy are summarized later in this section and detailed in Table 4 Monthly metrics for k = 3.

6.1.2 RESULTS

Here are the results for each value of *k* on the model:



k	excluded_pct	MAE	MSE	R2	MAE_rel_percent
0.0	73.4	0.025	0.057	0.608	87.1
0.5	23.5	0.293	0.962	0.598	84.2
1.0	14.9	0.536	2.255	0.581	78.1
1.5	11.3	0.782	3.608	0.594	75.2
2.0	8.0	0.848	3.834	0.578	75.6
2.5	5.3	0.887	3.741	0.610	71.7
3.0	3.7	0.946	4.233	0.634	69.6
3.5	2.8	0.946	4.19	0.638	69.6
4.0	1.7	0.955	4.295	0.629	70.2
4.5	1.1	0.968	4.3	0.629	71.2
5.0	0.8	0.984	4.327	0.626	72.4
6.0	0.4	0.983	4.315	0.627	72.3
7.0	0.2	0.992	4.378	0.622	73.0
8.0	0.1	0.999	4.387	0.621	73.5
9.0	0.0	0.997	4.386	0.621	73.4
10.0	0.0	0.997	4.386	0.621	73.4

Table 3 Sensitivity of validation accuracy to the IQR outlier multiplier

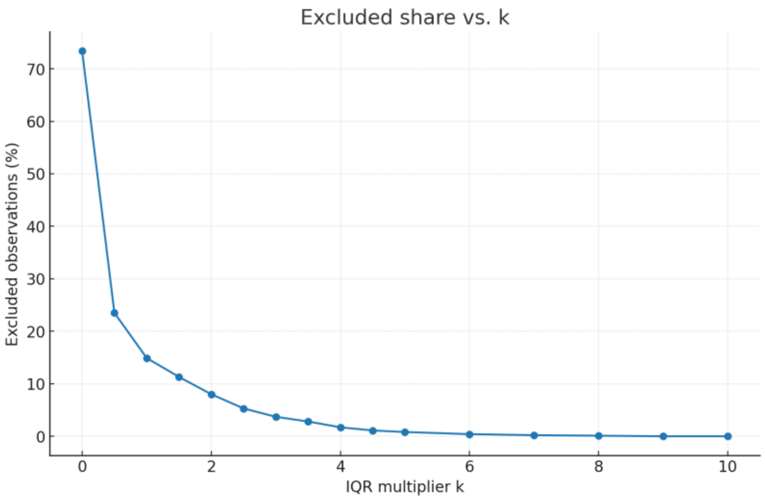


Figure 1 Excluded observations (%) as a function of k

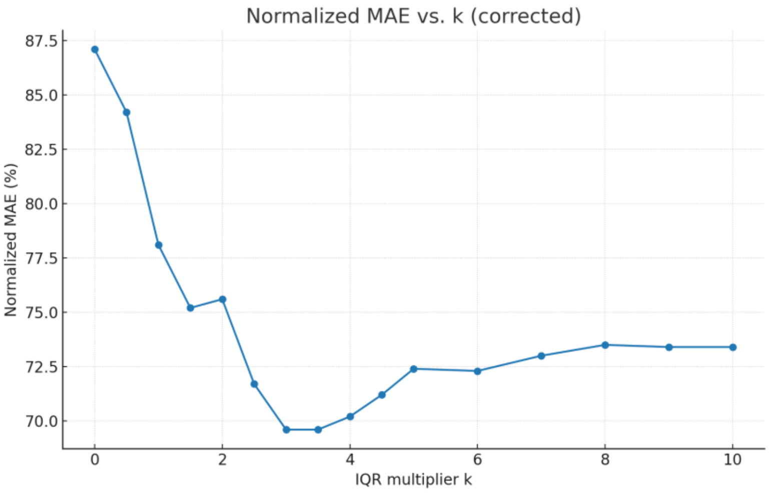


Figure 2 Normalized MAE (%) on the validation block vs k

The exclusion share declines steeply with k, by $k = 3$ only 3.7% of hours are removed, and for $k \geq 4$ the share becomes negligible. See Figure 1 and Table 3.

Performance improves as extreme hours are gradually reincluded, reaches a broad optimum around k = 3–3.5, and then drifts slightly downward as k increases further. At $k = 3$ the baseline model attains $\mathrm{MAE}_{\mathrm{rel}} = 69.6\%$, $R^2 = 0.634$, and MAE = 0.946 (MSE = 4.233). At $k = 3.5$ the results are essentially tied on $\mathrm{MAE}_{\mathrm{rel}}$ (69.6%) with a marginally higher R^2 (0.638) and slightly less trimming (2.8%). For $k \geq 4$, $\mathrm{MAE}_{\mathrm{rel}}$ increases (71.2% at $k = 4.5$, 73.5% at $k = 8$) without compensating gains in R^2.

Across all k, the model is dominated by $y_{t\text{-}1}$ (lag_1) with stable, secondary roles for $y_{t\text{-}24}$, $y_{t\text{-}168}$, and hour-of-day; temperature lags remain modest. This stability indicates that the outlier policy does not artifactually alter the learned structure.

6.1.3 SELECTION AND FREEZE OF *K*

On the basis of these results, $k = 3$ is selected and fixed for all subsequent. The decision is justified by three considerations:

- Accuracy at minimal trimming. $k = 3$ sits on the empirical optimum (tied with $k = 3.5$ on MAE_{rel}) while excluding only 3.7% of hours, preserving nearly the entire dataset and, crucially, peak-demand periods of operational interest.
- Comparability and pre-commitment. Freezing k before model selection, ablations, robustness checks, and bidding back-tests removes a degree of freedom that could bias results ex post. All headline metrics in this chapter are therefore conditional on a fixed data-cleaning policy.
- Robust learning signal. The near-constancy of feature salience across k supports the view that the choice of k does not induce spurious drivers. Selecting $k = 3$ provides additional protection against high-leverage anomalies without materially changing conclusions relative to $k = 3.5$.

The remainder of the chapter proceeds under the fixed policy $k = 3$.

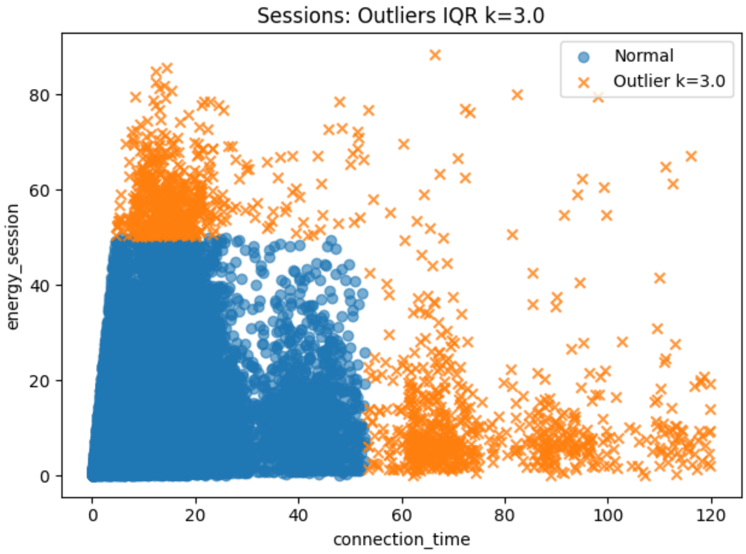


Figure 3 Session outliers with k = 3

To interpret the effect of the $k = 3$ IQR policy, Figure 3 plots per-session energy against connection time, highlighting observations flagged as outliers. The retained cloud (blue) forms a wedge bounded by physical limits, charging power and battery capacity, consistent with residential AC charging. In contrast, excluded points (orange) cluster in two regions:

First very long connection times with little energy (vehicles parked while plugged in), and second unusually high energy given short connection times, implying power rates that are implausible in the residential context. This pattern indicates that the policy primarily removes atypical, low-information, or physically inconsistent sessions rather than structurally relevant charging behavior.

For the selected policy ($k = 3$), monthly normalized errors are lower during high-load winter months (35.7% in 2021-02; 41.8–42.97% in 2020-12/2021-01) and higher during lower-load summer months (69.6–84.5% in 2021-05/2021-06). This pattern is consistent with normalization by the monthly mean, relative errors expand when the denominator is small. A full month-by-month table is provided in Table 4.

Month	MAE	MSE	R2	MAE_rel_percent
2020-08	2.740279	16.289791	0.56729	51.491045
2020-09	2.930573	18.427366	0.632598	49.566059
2020-10	3.845166	31.563427	0.581171	47.072319
2020-11	3.476104	24.201655	0.600156	49.137035
2020-12	4.214996	35.457987	0.612844	42.967323
2021-01	5.057799	55.035279	0.635774	41.819726
2021-02	5.010782	49.270329	0.760781	35.651367
2021-03	4.597326	43.687403	0.688587	42.168079
2021-04	2.160324	12.580135	0.568495	63.653654
2021-05	1.146511	5.003932	0.634391	69.886974
2021-06	0.962261	3.859659	0.515999	84.452034
2021-07	0.945719	4.233109	0.634455	69.56981

Table 4 Monthly metrics for k = 3

6.2 BASELINE MODEL

Baselines serve to contextualize the performance of the proposed Random Forest model by providing simple, transparent forecasting rules that are difficult to outperform on short horizons. Given the strong daily regularity of residential charging, the seasonal naive forecaster, defined by $\bar{y}_t = y_{t\text{-}24}$, is an appropriate primary comparator for hourly data (2021). Two additional classical baselines were assessed without any order or parameter search: a seasonal ARIMA with daily seasonality and a Holt–Winters Exponential Smoothing (ETS) model with additive seasonality and seasonal period 24. These models were fitted in univariate form to the aggregated hourly energy series to maintain a fair comparison and to avoid the complexities of exogenous regressors.

6.2.1 EVALUATION PROTOCOL

All baselines were estimated on the training + validation window and evaluated once on the held-out test window spanning 2020-08-01 00:00 to 2021-07-31 23:00 (8,760 hours). The split is month-aligned and strictly chronological. Accuracy is reported using MAE, RMSE, R^2, and nMAE, following the definitions in 4.5.1 The seasonal naive forecaster requires no estimation, its predictions are formed by lagging the series by 24 hours, with the first 24 test hours bridged using the last 24 hours of the estimation window to avoid undefined values.

6.2.2 TEST-YEAR RESULTS

Table 5 summarizes the test-year results. The seasonal naive baseline attains the best scores among the classical comparators considered, with MAE = 5.423 kWh, RMSE = 8.600 kWh, R^2 = 0.175, and nMAE = 80.88%. The ETS model (additive seasonality, s=24) performs slightly worse on all metrics. The seasonal ARIMA attempted here underperforms markedly, diagnostic inspection indicates the combination of seasonal differencing and sparsity drives forecasts toward zero, yielding poor fit despite correct alignment.

Baseline	MAE (kWh)	RMSE (kWh)	R^2	nMAE (%)
Seasonal naive ($\bar{\boldsymbol{y}}_t = y_{t-24}$)	5.423	8.600	0.175	80.88
ETS (seasonal additive, s = 24)	5.512	8.783	0.139	82.20
SARIMA $(1,0,0) \times (0,1,1)_{24}$, const	6.712	11.603	−0.502	100.10

Table 5 Test-year accuracy of classical baselines.

The test-year evidence indicates that a 24-hour seasonal naive constitutes a strong and defensible baseline for residential EV charging at hourly cadence. Its advantage stems from structural daily repetition in charging behavior and robustness to long runs of zeros. The ETS variant considered here slightly underperforms the naive benchmark, which is consistent with the difficulty of fitting additive seasonal components in zero-inflated series without extensive tuning. The seasonal ARIMA specification attempted performs substantially worse, on this dataset, seasonal differencing removes level information and combined with sparsity, drives forecasts toward near-zero means. The seasonal naive is therefore retained as the primary statistical comparator against which the Random Forest's gains are interpreted.

6.3 RANDOM FOREST RESULTS

6.3.1 MODEL SELECTION AND FINAL SPECIFICATION

With the IQR multiplier fixed at $k = 3$, the Random Forest was selected by minimizing MAE on the validation window under the month-aligned, strictly chronological splits like shown in Table 6. The search identified a variance-reduced configuration: $\mathrm{n}_{\mathrm{estimators}}$ = 300, min_samples_leaf = 20, bootstrap = True, max_samples = 0.9, max_features = 1.0, max_depth = None (random_state = 42). This model achieved a validation MAE of 3.144 kWh and was then refitted on train + validation for the final evaluation on the untouched test year.

parameter	value
n_estimators	300
min_samples_leaf	20
bootstrap	True
max_samples	0.9
max_features	1
max_depth	None
random_state	42
validation_MAE_kWh	3.144
k_IQR	3

Table 6 Hyperparameters

6.3.2 TEST-YEAR POINT-FORECAST ACCURACY

On the held-out test year (2020-08-01 to 2021-07-31; 8,760 hours), the selected Random Forest attains MAE = 3.013 kWh, RMSE = 4.874 kWh, R^2 = 0.735, and nMAE = 44.93%.

Relative to the baselines in Section 6.2, this corresponds to a ≈44% MAE reduction versus the seasonal naive (y_{t-24}) and a ≈10.5% RMSE reduction versus the lag-1 naive, with R^2 improving from 0.669 (lag-1) to 0.735. These improvements are operationally meaningful because RMSE penalizes peak-hour deviations that drive bidding risk.

Model	MAE_kWh	RMSE_kWh	R2	nMAE_percent
Random Forest (final)	3.0129	4.8738	0.7349	44.9317
Naive lag-1	3.1385	5.4461	0.6690	46.8001
Seasonal naive (y_{t-24})	5.4291	8.6072	0.1743	80.8880

Table 7 Aggregate Test Metrics

6.3.3 DAY-AHEAD OPERATIONAL PROFILE

To examine when errors matter across a bidding day, FIGURE 6.3 plots MAE(h) and nMAE(h) by hour of day h ∈ {0,…,23} on the test block. As expected, absolute errors are largest around morning and evening charging peaks, whereas normalized errors inflate overnight when denominators are small as shown in Figure 4 MAE by hour and Figure 5 nMAE by hour.

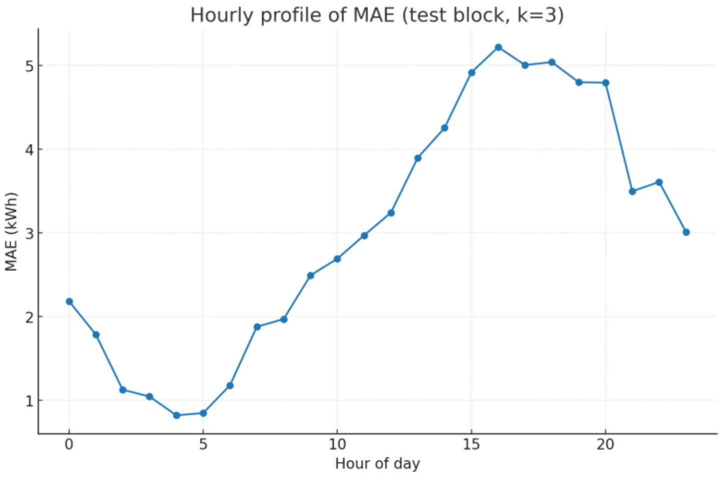


Figure 4 MAE by hour

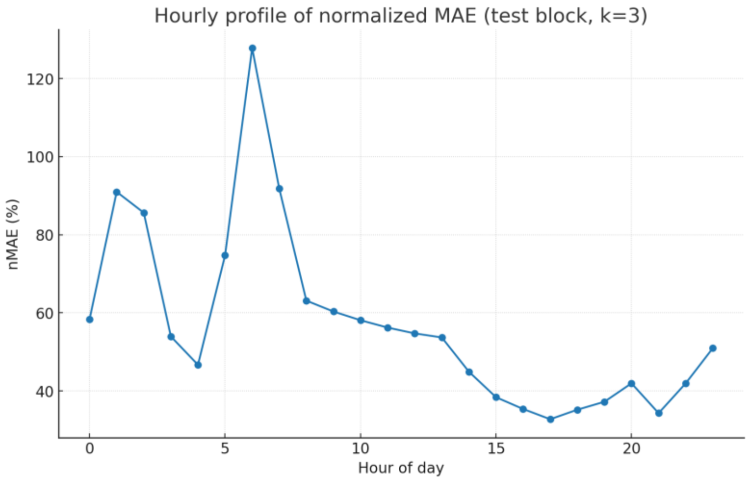


Figure 5 nMAE by hour

6.3.4 INTERPRETATION VIA FEATURE IMPORTANCE

Two complementary diagnostics are reported. Permutation importance on the validation window quantifies the increase in MAE when a feature is permuted, holding others fixed. The Random Forest’s impurity-based importance summarizes average reductions in node impurity across trees.

Both views agree that short-lag persistence is dominant. Permuting lag_1 alone increases MAE by ≈1.14 kWh, while most other predictors have near-zero or slightly negative permutation importances, evidence of redundancy once lag_1 and the diurnal structure are present. The impurity ranking assigns ≈0.933 of total importance to lag, followed by a second tier comprising weekly recurrence (lag_{168}), diurnal encodings (hour_cos, hour_sin, hour), and daily recurrence (lag_{24}), rolling variability measures contribute modestly, calendar and temperature features are negligible. The model thus behaves like a non-parametric with diurnal/weekly modulation, rather than a strongly exogenous forecaster.

For residential aggregations at hourly cadence, further accuracy gains likely require richer exogenous signals (like arrival/occupancy or price exposure) or sequence models designed to capture rare peak patterns.

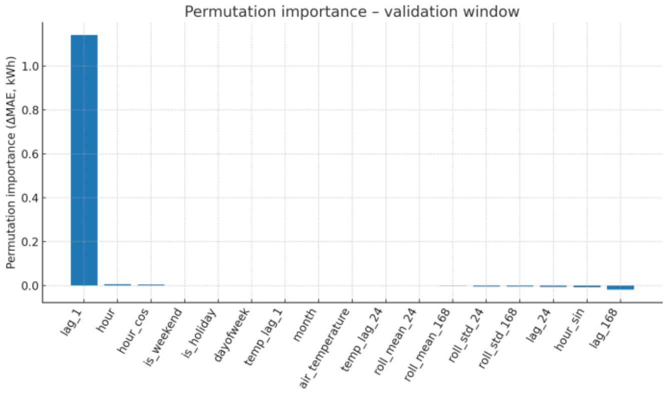


Figure 6 Permutation importantce

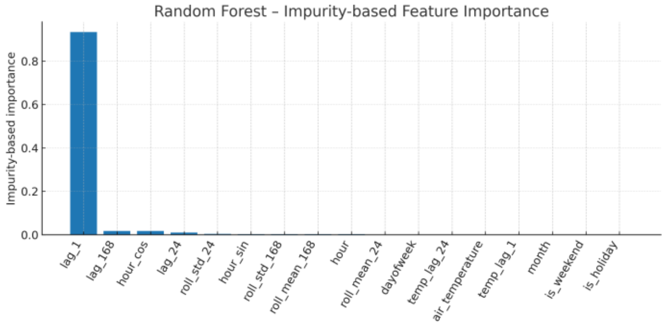


Figure 7 Impurity-based Feature Importance

6.3.5 VISUAL FACE VALIDITY

To provide face validity, Figure 8 Winter Real vs Predicted Load and Figure 9 Summer Real vs Predicted Load overlay realized and predicted hourly energy for two representative weeks (winter and summer). The winter week shows accurate tracking of higher, more volatile demand, the summer week shows small absolute deviations even when normalized errors appear large due to low denominators. A monthly view (February 2021) is provided as a supplementary exhibit to illustrate sustained performance over a full billing cycle.

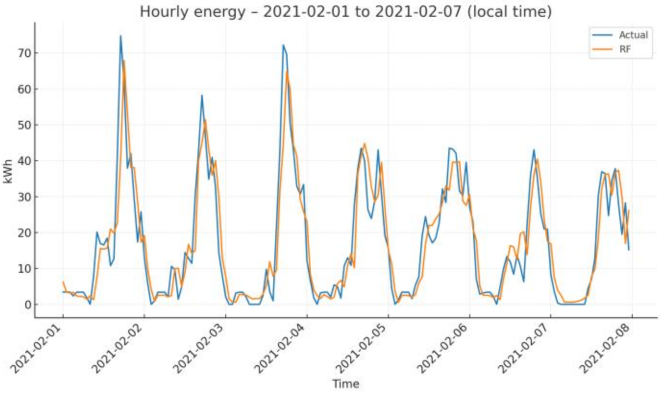


Figure 8 Winter Real vs Predicted Load

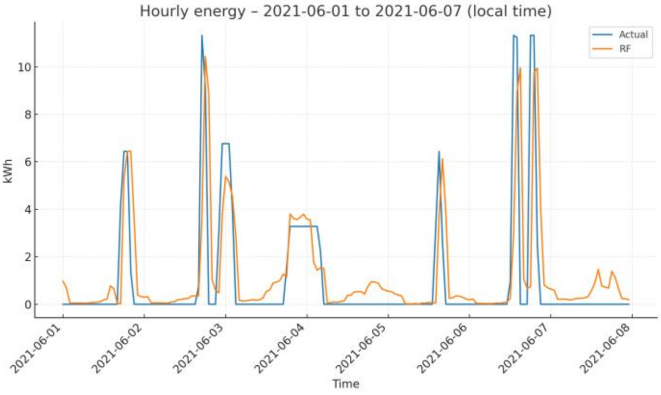


Figure 9 Summer Real vs Predicted Load

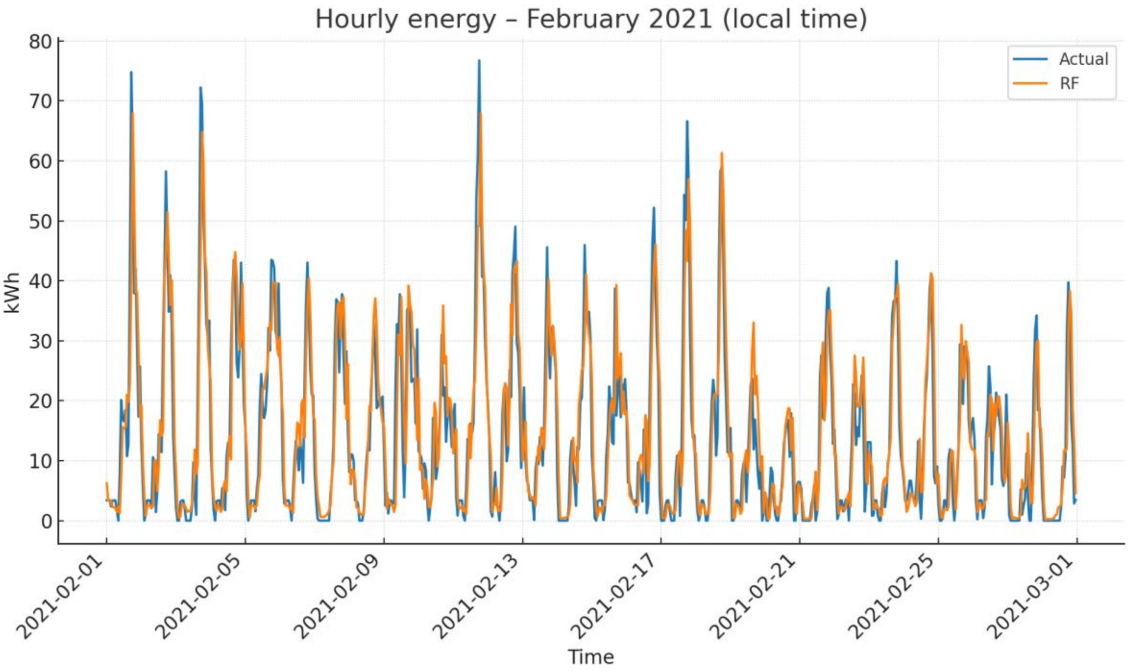


Figure 10 February monthly Real vs Predicted Load

Fixing $k = 3$ and selecting a variance-reduced Random Forest delivers substantial improvements over classical baselines: approximately 44% lower MAE than the seasonal naive and ≈10.5% lower RMSE than the lag-1 naive on the test year, with $R^2 = 0.735$. Errors concentrate around peak hours that matter for bidding, precisely where the Random Forest offers the largest RMSE gains. Feature-importance analyses indicate that the model's predictive power is driven primarily by short-lag persistence plus diurnal/weekly regularities, with weather and calendar signals playing only a minor role for this dataset.

Chapter 7. CONCLUSIONS

7.1 SUMMARY OF CONTRIBUTIONS

This thesis proposed and validated an AI-based framework to forecast the aggregated hourly charging demand of an EV aggregator and to translate forecast behavior into operational guidance for market bidding. The main contributions are the following.

1. A disciplined, time-aware evaluation pipeline. The study enforced strict chronological splits (train/validation/test) with complete-month boundaries and a single final test pass. Model selection was conducted on the validation block using MAE, and the chosen configuration was refit on train + validation before test evaluation, thereby reducing optimistic bias.
2. A transparent data-cleaning policy anchored in robustness. An interquartile-range (IQR) outlier screen was tuned empirically and frozen at $k = 3$ before any headline evaluation. This removes 3.7% of hourly observations, balances fidelity to rare but relevant peaks against undue leverage from implausible points and keeps downstream comparisons defensible. A session-level diagnostic showed that the policy primarily excludes long-idle/low-energy sessions or combinations inconsistent with residential charging power limits.
3. A competitive yet interpretable forecasting model. A Random Forest selected on validation ($n_{estimators}$ = 300, min_samples_leaf = 20, bootstrap = True, max_samples = 0.9, max_features = 1.0, max_depth = None) achieved on the held-out test year MAE = 3.013 kWh, RMSE = 4.874 kWh, $R^2 = 0.735$, and nMAE = 44.93%. Against a strong seasonal naive baseline (y_{t-24}), the model reduced MAE by ~44% and RMSE by ~43%, versus a lag-1 naive, it improved MAE by ~4% and RMSE by ~10.5%.
4. Actionable insights for operations. Errors concentrate around morning/evening peaks, while normalized errors inflate overnight due to small denominators. Feature-importance analyses (permutation and impurity) showed that short-lag persistence

(y_{t-1}) and diurnal/weekly structure provide most of the signal, temperature and coarse calendar variables add little incremental value for this residential dataset. These findings were translated into pragmatic bidding guidance (hour-specific safety margins, rolling updates) and managerial recommendations.

Collectively, the framework delivers measurable accuracy gains over classical baselines while remaining simple enough for reliable deployment and audit.

7.2 ANSWERS TO THE RESEARCH QUESTIONS

To organize the evidence, the thesis addressed three questions stated in the introduction.

RQ1. Can an AI model improve short-term EV-aggregator load forecasting over standard statistical baselines?

The final Random Forest outperformed both seasonal naive and ETS/SARIMA baselines on the held-out year. Relative to the seasonal naive, MAE fell from 5.423 to 3.013 kWh and RMSE from 8.600 to 4.874 kWh, R^2 rose from 0.175 to 0.735. Even against the lag-1 naive, MAE and RMSE improved (from 3.138 to 3.013 kWh and from 5.446 to 4.874 kWh). These gains are largest at peak hours, where operational risk is greatest.

RQ2. Which predictors contribute materially to forecast accuracy for residential EV aggregation?

Short-lag persistence dominates. Permuting y_{t-1} increases validation MAE by ~1.14 kWh, while most other variables exhibit near-zero (or slightly negative) permutation importance, indicating redundancy once y_{t-1} and hour-of-day are present. Impurity-based importance assigns ~0.93 of total weight to y_{t-1}, with a second tier comprising weekly recurrence (y_{t-168}), diurnal encodings (hour, hour_sin, hour_cos, hour, hour_sin hour_cos), and daily recurrence (y_{t-24}). Weather and coarse calendar effects contribute marginally in this setting.

RQ3. How sensitive are results to the outlier-screening policy, and what policy is appropriate?

Moderate trimming is best. A systematic sweep of the IQR multiplier found a broad optimum around k = 3–3.5. Fixing achieves near-minimal normalized MAE with only 3.7% of hours excluded and stable feature salience. This pre-commitment prevents ex-post tuning and ensures that all reported comparisons share an identical sample.

7.3 Practical recommendations for EV aggregators

The empirical results point to a deployment strategy that privileges reliability and operational discipline over model complexity. Because very short-lag persistence (y_{t-1}) and the diurnal/weekly structure carry most of the predictive signal, aggregators will gain more from ensuring timely, high-quality metering than from adding weak exogenous feeds. In practice, this means prioritizing low-latency data ingestion, automated checks for missing or duplicated intervals, and immediate remediation of telemetry gaps. A clean and promptly updated y_{t-1} stream is the single most valuable input to the model identified here.

A second pillar is pre-committing the data-cleaning policy. The IQR screen fixed at k=3 in this study strikes an effective balance between robustness and fidelity to genuine peaks. Freezing that policy ex ante, not revisiting it when results fluctuate, prevents hidden degrees of freedom from seeping into operations and keeps weekly performance reports auditable. In alignment with the thesis design, the screen should be applied to the aggregated hourly target rather than to feature-dependent transforms, so that the forecasting pipeline remains free of look-ahead and selection biases.

Translating point forecasts into bids benefits from hour-specific safety margins. The error profile by hour of day shows that absolute deviations concentrate in morning and evening peaks, precisely when imbalance penalties are most consequential. A pragmatic rule is to deduct, from the raw point forecast, the empirical 75th percentile of the hour-specific absolute residuals computed on a rolling validation window. This simple transformation widens the margin when risk is high and narrows it overnight without requiring full probabilistic modeling. Because the model's signal is largely local in time, rolling updates re-forecasting intraday as new hours realize can further reduce exposure in intraday or balancing markets.

Sustained performance also depends on lightweight model governance. A quarterly refit on a recent window, coupled with routine residual monitoring, will detect drift arising from tariff changes, holidays, or infrastructure upgrades. When control limits are breached, the remedy should be procedural rather than ad hoc: refresh the training window, keep the cleaning policy and evaluation protocol unchanged, and re-issue a versioned model artifact. Finally, feature expansion should be targeted by portfolio context. For residential fleets similar to the data studied here, weather and coarse calendar signals add limited value. In workplace or price-responsive fleets, however, arrival/occupancy indicators, connector ratings, and explicit price exposure are natural candidates, when such signals are available, upgrading to probabilistic outputs can align the forecasting layer with penalty-aware bidding without abandoning the governance practices just described.

7.4 LIMITATIONS

The conclusions of this thesis are bound by several contextual and methodological constraints. The underlying dataset reflects residential charging in Norway without vehicle-to-grid discharging, a setting with specific infrastructure, tariff structures, and climatic conditions. While the modeling approach is general, its quantitative gains may not transfer one-for-one to commercial or mixed portfolios, to regions with different connector power levels, or to fleets whose demand is tightly coupled to workplace schedules or dynamic prices.

Methodologically, the primary model is a Random Forest optimized for point accuracy (MAE) and evaluated on a single held-out year. Although the evaluation protocol is strictly chronological and pre-committed, stronger statistical assurances would come from multi-year external tests and formal uncertainty quantification. Moreover, markets with asymmetric penalties or explicit capacity obligations would benefit from probabilistic forecasts, quantiles or predictive intervals, so that bids can be derived from risk tolerances rather than from heuristic safety margins. These extensions are outside the present scope.

The feature scope is intentionally lean (calendar and temperature only) to preserve auditability and focus on what can be deployed reliably. That several of these variables showed limited incremental value here should not be read as evidence against richer exogenous information, it simply reflects the residential context and available data. Similarly, the IQR filtering choice (fixed at $k = 3$) is justified empirically and supported by session-level diagnostics, yet any filter can attenuate rare but operationally meaningful extremes. The small exclusion rate mitigates this risk but does not eliminate it.

Finally, the framework assumes a degree of temporal stability in user behavior within the test year. Structural breaks, policy changes, rapid EV adoption shifts, or infrastructure upgrades, could alter demand patterns in ways that a point-forecasting Random Forest does not anticipate. The recommended governance measures (rolling updates, drift monitoring, versioned configurations) are intended to reduce this exposure, but they cannot substitute for truly exogenous signals when behavior is driven by factors absent from the data.

Overall, the thesis demonstrates that a carefully engineered yet simple AI model, embedded in a robust, pre-committed evaluation pipeline, can deliver substantial accuracy gains over classical baselines for day-ahead EV-aggregator demand forecasting. The results translate into clear operational guidance for bidding and highlight targeted avenues, richer exogenous data and probabilistic decision-alignment, where further improvements are both feasible and worthwhile.

Chapter 8. REFERENCES

Amazon Web Services. (2025). *Amazon SageMaker*. Retrieved from https://aws.amazon.com/sagemaker/

Amezquita, H., Rojas, M., & Arango, H. (2024). Forecasting electric vehicles' charging behavior at charging stations: A data-science-based approach. *Energies, 17(14)*. doi:10.3390/en17143396

Andoni, M. (2019). Blockchain technology in the energy sector: A systematic review of challenges and opportunities. *Renewable and Sustainable Energy Reviews, 100*, 143–174. doi:10.1016/j.rser.2018.10.014

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer. doi:10.1007/978-0-387-45528-0

Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2015). *Time Series Analysis: Forecasting and Control.* Wiley.

Breiman, L. (2001). *Random forests* (Vol. 45(1)). Machine Learning. doi:10.1023/A:1010933404324

Deb, S., Kalam, A., & Agalgaonkar, A. (2022). Prediction of charging demand of electric city buses of Helsinki, Finland by random forest. *Energies, 15(10)*. doi:10.3390/en15103679

Dudek, G. (2022). A comprehensive study of random forest for short-term load forecasting. *Energies, 15(20)*. doi:10.3390/en15207547

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second*

International Conference on Knowledge Discovery and Data Mining (KDD-96), (pp. 226–231). doi:10.5555/3001460.3001507

European Union. (2019). *Directive (EU) 2019/944.*

Federal Energy Regulatory Commission (FERC). (2020). *Order No. 2222: Participation of Distributed Energy Resource Aggregations in Markets Operated by Regional Transmission Organizations and Independent System Operators.* Washington, DC: Federal Energy Regulatory Commission. Retrieved from https://www.ferc.gov/sites/default/files/2021-03/E-1.pdf

Friedman, J. (2001). *Greedy function approximation: A gradient boosting machine* (Vol. 29(5)). Annals of Statistics. doi:10.1214/aos/1013203451

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Google Cloud. (2025). *Vertex AI.* Retrieved from https://cloud.google.com/vertex-ai

H2O.ai. (2025). *H2O-3*. Retrieved from https://docs.h2o.ai/h2o/latest-stable/h2o-docs/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. doi:10.1007/978-0-387-84858-7

Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting, 32(3)*, 914–938. doi:10.1016/j.ijforecast.2015.11.011

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice.* OTexts.

International Energy Agency. (2024). *Global EV Outlook 2024.* Retrieved from https://www.iea.org/reports/global-ev-outlook-2024

International Energy Agency. (2025). *Global EV Outlook 2025.* Retrieved from https://www.iea.org/reports/global-ev-outlook-2025

Jang, J.-S. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics, 23(3)*, 665–685. doi:10.1109/21.256541

Kempton, W., & Tomić, J. (2005). Vehicle-to-grid power fundamentals: Calculating capacity and net revenue. *Journal of Power Sources*, 268–279. doi:10.1016/j.jpowsour.2004.12.025

Khan, W., Sommers, W., Walker, S., Bont, K., van der Velden, J., & Zeiler, W. (2023). Comparison of electric vehicle load forecasting across different spatial levels with incorporated uncertainty estimation. *Energy, 283*. doi:10.1016/j.energy.2023.129213

Kingma, D., & Welling, M. (2014). Auto-Encoding Variational Bayes. doi:10.48550/arXiv.1312.6114

Klir, G., & Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Prentice Hall (Prentice Hall PTR).

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28(2)*, 129–137. doi:10.1109/TIT.1982.1056489

Mamdani, E., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies, 7(1)*, 1–13. doi:10.1016/S0020-7373(75)80002-2

Microsoft. (2025). *Azure Machine Learning*. Retrieved from https://learn.microsoft.com/azure/machine-learning/

Mnih, V., Kavukcuoglu, K., & Silver, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 529–533. doi:10.1038/nature14236

Muratori, M. (2018). Impact of uncoordinated plug-in electric vehicle charging on residential power demand. *Nature Energy, 3*(3), 193–201. doi:10.1038/s41560-017-0074-z

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1)*, 86–97. doi:10.1002/widm.53

Norwegian Meteorological Institute. (n.d.). FROST API. Retrieved from https://frost.met.no/

Ostermann, A., & Haug, T. (2024). Probabilistic forecast of electric vehicle charging demand: Analysis of different aggregation levels and energy procurement. *Energy Informatics, 7*. doi:10.1186/s42162-024-00319-1

Papadaskalopoulos, D., & Strbac, G. (2013). Decentralized Participation of Flexible Demand in Electricity Markets - Part I: Market Mechanism. *IEEE Transactions on Power Systems, 28*(4), 3658–3666. doi:10.1109/TPWRS.2013.2245686

Pedrycz, W., & Gomide, F. (2007). *Fuzzy Systems Engineering: Toward Human-Centric Computing*. Wiley–IEEE Press. doi:10.1002/9780470168967

Pudjianto, D., Ramsay, C., & Strbac, G. (2007). Virtual power plant and system integration of distributed energy resources. *IET Renewable Power Generation, 1*, 10–16. doi:10.1049/iet-rpg:20060023

Qu, R., Kou, R., & Zhang, T. (2025). The impact of weather variability on renewable energy consumption: Insights from explainable machine learning models. *Sustainability, 17(1)*. doi:10.3390/su17010087

Ross, T. (2010). *Fuzzy Logic with Engineering Applications.* Wiley. doi:10.1002/9781119994374

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. doi:10.48550/arXiv.1707.06347

Scikit-Learn Developers. (2025). *scikit-learn: Machine Learning in Python*. Retrieved from https://scikit-learn.org/

Siano, P. (2014). Demand response and smart grids - A survey. *Renewable and Sustainable Energy Reviews*, 461–478. doi:10.1016/j.rser.2013.10.022

Sørensen, Å., Sartori, I., Lindberg, K., & Andresen, I. (2024). Electric vehicle charging dataset with 35,000 charging sessions from 12 residential locations in Norway. Zenodo. doi:https://doi.org/10.5281/zenodo.13896176

Statsmodels Developers. (2025). *statsmodels: Statistical modeling and econometrics in Python*. Retrieved from https://www.statsmodels.org/

Sundström, O., & Binding, C. (2012). Flexible charging optimization for electric vehicles considering distribution grid constraints. *IEEE Transactions on Smart Grid, 3(1)*, 26–37. doi:10.1109/TSG.2011.2168431

Sutton, R., & Barto, A. (2018). *Reinforcement Learning: An Introduction.* MIT Press.

Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics, 15(1)*, 116–132. doi:10.1109/TSMC.1985.6313399

Taylor, S., & Letham, B. (2015). Forecasting at scale. *The American Statistician, 72(1)*, 37–45. doi:10.1080/00031305.2017.1380080

TensorFlow Team. (2025). *TensorFlow*. Retrieved from https://www.tensorflow.org/

Vázquez-Canteli, J., & Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy, 235*, 1072–1089. doi:10.1016/j.apenergy.2018.11.002

Xiang, S., Zhen, C., Peng, J., Zhang, L., & Pu, Z. (2023). Power load prediction of smart grid based on deep learning. *Procedia Computer Science, 228(C)*, 762–773. doi:10.1016/j.procs.2023.11.090

APPENDIX I