instituto superior de engenharia de lisboa

# Electric Vehicle X Driving Range Prediction 2
# EV X DRP2


**JOÃO FRANCISCO FIDALGO VALIDO**

(Licenciatura em Engenharia Informática)


Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e Multimédia


**Orientadores:** Doutor Artur Jorge Ferreira

Doutor David Pereira Coutinho


**Júri:**

**Presidente:** Doutor Pedro Miguel Mendes Torres Jorge

**Vogais:** Doutor Gonçalo Nuno De Oliveira Duarte

Doutor David Pereira Coutinho

**Julho 2025**

instituto superior de engenharia de lisboa

# Electric Vehicle X Driving Range Prediction 2
# EV X DRP2


**JOÃO FRANCISCO FIDALGO VALIDO**

(Grau de Licenciatura em Engenharia Informática)


Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e Multimédia


**Orientadores:** Doutor Artur Jorge Ferreira, ISEL/DEI
Doutor David Pereira Coutinho, ISEL/DEETC


**Júri:**

| | |
|---|---|
| **Presidente:** | Doutor Pedro Miguel Mendes Torres Jorge, ISEL/DEI |
| **Vogais:** | Doutor Gonçalo Nuno De Oliveira Duarte, ISEL/DEM |
| | Doutor David Pereira Coutinho, ISEL/DEETC |

**Julho 2025**

# Acknowledgements

I would like to express my deepest gratitude to my supervisors, **Dr. Artur Jorge Ferreira** and **Dr. David Pereira Coutinho**, for their invaluable guidance, constant support, and insightful feedback throughout the development of this work. I am also sincerely thankful to my family for their unconditional encouragement and to my friends for their motivation and companionship during this academic journey. I am also grateful to my classmate and friend **Duarte Valente** for his teamwork and incentive.

This research was supported by Instituto Politécnico de Lisboa (IPL) under Grant `IPL/IDI&CA2024/ML4EP_ISEL`

**Statement of integrity**

I declare that this **project work** is the result of my personal and independent research. Its content is original, and all sources listed in the bibliographic references were consulted and are duly mentioned in the text. I further declare that all scientific and technical references relevant to the development of the work are duly cited and included in the bibliographic references.

The author

---

Lisbon, July 07, 2025

# Abstract

The use of Electric Vehicles (EV) has increased in recent years. The autonomy of the EV, expressed as its Driving Range (DR) is a key factor. This autonomy depends on several variables related to the vehicle itself as well as with external conditions. An accurate estimation of the DR value at each moment is a challenging task. In this thesis, we address the DR estimation problem using machine learning techniques. We build a dataset with 11 features, for DR estimation, using publicly available EV data. Then, we discuss the use of Machine Learning (ML) Regression techniques to estimate DR, with Linear Regression (LR), Multilayer Perceptron (MLP), and Radial Basis Function (RBF) neural networks. Moreover, we assess the effect of unsupervised dimensionality reduction techniques using feature selection and feature reduction approaches. The experimental results show that the use of both feature selection and feature reduction are useful at reducing the dimensionality of the data, keeping or improving the performance for DR estimation. This study also identifies the top features for DR estimation. The best feature selection method was the Mean-Median approach, while Principal Component Analysis yielded the best results in terms of feature reduction. Among the regression techniques evaluated, linear regression achieved the best overall performance. However, in real-world scenarios, where a larger number of variables may be present, methods such as MLP or RBF might offer better adaptability and robustness.

**Keywords:** Dimensionality Reduction; Driving Range Estimation; Electric Vehicle; Feature Reduction; Feature Selection; Machine Learning; Neural Networks; Regression.

# Resumo

A utilização de veículos eléctricos (VE) tem aumentado nos últimos anos. A autonomia do VE, expressa na sua autonomia de condução (DR), é um fator-chave. Esta autonomia depende de diversas variáveis relacionadas com o próprio veículo, bem como com as condições externas. Uma estimativa exacta do valor do DR em cada momento é uma tarefa difícil. Neste artigo, construímos um conjunto de dados com 11 caraterísticas para a estimativa da DR, utilizando dados de VE disponíveis publicamente. Em seguida, discutimos a utilização de técnicas de regressão de Aprendizagem Automática (ML) para estimar a DR, com Regressão Linear (LR), Perceptron Multicamada (MLP) e redes neurais de Função de Base Radial (RBF). Além disso, avaliamos o efeito de técnicas de redução de dimensionalidade não supervisionadas utilizando abordagens de seleção e redução de caraterísticas. Os resultados experimentais mostram que a utilização tanto da seleção como da redução de caraterísticas são úteis para reduzir a dimensionalidade dos dados, mantendo ou melhorando o desempenho da estimativa de DR. Este estudo também identifica as principais caraterísticas para a estimativa de DR. O melhor método de seleção de caraterísticas foi a abordagem ML, enquanto a PCA produziu os melhores resultados em termos de redução de caraterísticas. Entre as técnicas de regressão avaliadas, a regressão linear obteve o melhor desempenho global. No entanto, em cenários do mundo real, onde pode estar presente um maior número de variáveis, métodos como o MLP ou o RBF podem oferecer uma melhor adaptabilidade e robustez.

**Palavras-chave:** Redução da dimensionalidade; Estimativa da autonomia; Veículo elétrico; Redução de caraterísticas; Seleção de caraterísticas; Aprendizagem automática; Redes neuronais; Regressão.

# Contents

# List of Figures

# List of Tables

# Glossary

| | |
|---|---|
| Absolute Cosine (AC) | A metric used to measure geometric similarity between feature vectors, indicating redundancy when values are close to 1 44 |
| Baseline Model | A simple reference model used for performance comparison 13 |
| Bayesian Neural Network (BNN) | A neural network that models uncertainty by learning distributions over weights 24 |
| ChargeCar Database | A public EV dataset containing trip and charging data 7 |
| Emobpy | Tool to generate synthetic EV trip data based on European mobility statistics 10 |
| Ensemble Stacked Generalization | A model that combines several base models to improve predictive performance 14 |
| Evaluation Metrics | Quantitative measures of prediction performance such as MAE, MSE, RMSE, MAPE, and $R^2$ 27 |
| Filter Method | Feature selection method based on statistical measures without model training 19 |
| Gaussian Noise | Random noise following a normal distribution used to simulate uncertainty 4 |
| IEC Power | Standardized electric power metric used in EV telemetry 29 |
| Min-Max Scaling | Normalization technique that scales data between 0 and 1 30 |
| Power (kW) | Instantaneous electrical power measured in kilowatts 29 |

| | |
|---|---|
| Regression | Supervised learning task where the output is a continuous variable ix |
| Stacked Generalization | Modeling approach that combines multiple learners using a meta-model 14 |
| Telemetry Data | Sensor-based, time-stamped data collected from EVs (e.g., speed, power, SoC) 7 |
| Trip Dataset | Collection of trip-level records used for model training and evaluation 2 |
| Trust Calibration | Method to adjust model confidence to reflect real-world reliability 24 |
| Voltage (V) | Electric potential difference, part of EV power calculation 8 |

# Acronyms

| | |
|---|---|
| AEC | Average Energy Consumption 30 |
| AFA | Adaptive Filter Algorithms 10 |
| AI | Artificial Intelligence 6 |
| CM | Conventional Methods 10 |
| CNN | Convolutional Neural Networks 17 |
| CR | Cumulative Relevance 34 |
| DR | Driving Range 1 |
| DT | Decision Trees 15 |
| eRange | Electric Range xv, 1, 7, 14 |
| EV | Electric Vehicle 1, 5, 7 |
| FBD | Full Battery Distance 30 |
| FBE | Full Battery Energy 30 |
| FR | Feature Reduction 3 |
| FS | Feature Selection 3 |
| GNN | Graph Neural Network 5 |
| GPS | Global Positioning System 8 |
| HBA | History-Based Algorithm 29 |
| HM | Hybrid Methods 10 |
| HVAC | Heating, Ventilation, and Air Conditioning 23 |
| ICE | Internal Combustion Engine 9 |
| JARI | Japan Automobile Research Institute 14 |
| KNN | k-Nearest Neighbor 17 |
| LA | Learning Algorithms 10 |

| | |
|---|---|
| LightGBM | Light Gradient-Boosting Machine 14 |
| LIME | Local Interpretable Model-agnostic Explanations 6 |
| LR | Linear Regression 3 |
| LSTM | Long Short-Term Memory 5, 13 |
| MAD | Mean Absolute Difference 24 |
| MAE | Mean Absolute Error 3, 13 |
| MAPE | Mean Absolute Percentage Error 3 |
| ML | Machine Learning 1, 5, 7 |
| MLP | Multilayer Perceptron 3, 5 |
| MM | Mean-Median ix, 24 |
| MSE | Mean Squared Error 3 |
| NDANEV | National Big Data Alliance of New Energy Vehicles 13 |
| NLO | Non-Linear Observers 10 |
| NN | Neural Networks 24 |
| PCA | Principal Component Analysis ix, 24 |
| PINN | Physics-Informed Neural Network 7 |
| $R^2$ | Coefficient of Determination 3 |
| RBF | Radial Basis Function 3 |
| ReLU | rectified linear unit 15 |
| RF | Random Forests 5 |
| RMSE | Root Mean Squared Error 3 |
| RNN | Recurrent Neural Networks 16 |
| SHAP | SHapley Additive exPlanations 6 |
| SoC | State of Charge 1, 5, 14 |
| SVD | Singular Value Decomposition 24 |
| VED | Vehicle Energy Dataset 7 |
| XAI | eXplainable AI 24 |
| XGBoost | Extreme Gradient Boosting 5, 13 |

# 1 Introduction

The growing global concern about climate change has been a major driver of international agreements, such as the Paris Agreement (2015) [1], which encourage governments and industries to adopt sustainable transportation solutions. Electric Vehicle (EV) has emerged as a central component of this transition, offering a clean alternative to reduce greenhouse gas emissions and support a more sustainable future.

With the rising popularity of EV, manufacturers are striving to improve vehicle performance, particularly regarding driving range, or *Electric Range (eRange)*. The *eRange*, expressed in kilometers, represents the estimated distance a vehicle can travel on the remaining battery charge. Accurate range estimates help alleviate driver's anxiety about reaching charging stations and enable better route planning.

Estimating the *eRange* is a complex challenge, a regression problem, since one aims to predict the value of the Driving Range (DR), at each time instant. It depends on numerous interrelated variables, including vehicle design, driving behavior, environmental conditions, road inclination, and the State of Charge (SoC) of the battery. These dependencies are often nonlinear and dynamic, making *eRange* prediction a difficult and relevant research problem.

This thesis focuses on addressing this problem: **estimating the electric range (*eRange*) of electric vehicles**.

- **Primary objective:** To develop a predictive model for estimating the *eRange* using real-world electric vehicle data and Machine Learning (ML) techniques.
- **Secondary objective:** To compare the performance of different regression algorithms in order to identify the most effective model for *eRange* prediction.

## 1.1 Machine Learning

Machine Learning has shown great success in tackling complex problems in areas such as big data analysis, pattern recognition, and data mining. By learning from historical

data, ML models are capable of making increasingly accurate predictions, making them well-suited to address the intricacies of *eRange* estimation.

Previous research has demonstrated that ML methods outperform traditional approaches in estimating the *eRange*. This work builds on these efforts by implementing and comparing ML-based regression models, trained on publicly available datasets, to provide scalable and reproducible solutions for *eRange* prediction.

## 1.2 Neural Networks

Neural networks, a subset of machine learning techniques, are particularly well-suited to modeling nonlinear relationships and working with high-dimensional data. These capabilities make them ideal for predicting *eRange*, where input variables interact in dynamic and unknown ways.

In this project, neural networks were employed alongside traditional regression models to improve prediction accuracy. Techniques such as feedforward architectures, backpropagation, and gradient descent optimization were utilized. These models were benchmarked against conventional algorithms using standardized performance metrics, showcasing their strengths in handling diverse and dynamic datasets.

Moreover, neural networks support real-time prediction and adaptation to changing driving conditions, making them a promising approach for future EV applications.

## 1.3 Proposed Approach

To address the challenge of EV driving range estimation, this thesis proposes a structured approach grounded in ML techniques. The goal is to develop predictive models capable of accurately estimating the remaining range of an EV based on real-world usage data and vehicle characteristics. Figure 1.1 illustrates the overall methodology adopted in this work, encompassing all stages from data acquisition and preprocessing, through regression model training, to performance evaluation. This pipeline is designed not only to achieve high prediction accuracy, but also to ensure computational efficiency and model interpretability.

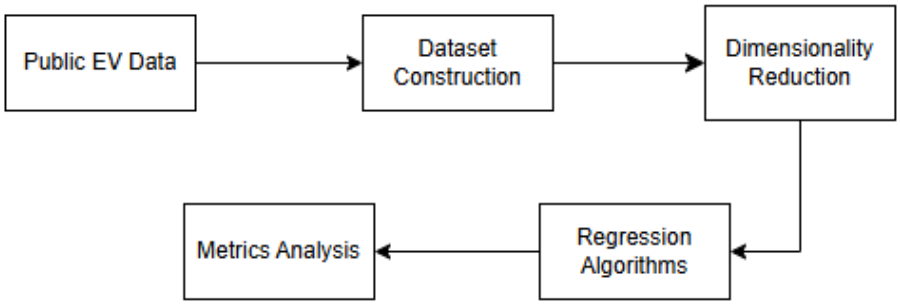


Figure 1.1: *Methodology followed in this thesis to address the eRange estimation problem—from data acquisition and preprocessing, through model training, to evaluation.*

The main approach of this thesis is to analyze and compare machine learning-based techniques for predicting *eRange*, using publicly available EV Trip Dataset. The proposed

methodology follows these steps:

### Data Collection and Processing

The process begins by collecting two types of data:

- EV specifications
- EV trip logs

These are processed and combined into a single dataset that includes both raw and engineered features.

### Target Definition (*eRange*)

The ground-truth eRange definition is then computed and added as the target variable to the dataset, enabling predictive modeling.

### Machine Learning Algorithms

Dimensionality reduction techniques are applied to enhance model interpretability and performance. This step involves the use of Feature Reduction (FR) and Feature Selection (FS) techniques.

### Regression Modeling

Several regression models are trained on the different versions of the dataset, namely:

- Linear Regression (LR)
- Multilayer Perceptron (MLP)
- Radial Basis Function (RBF)

### Model Evaluation

Performance is assessed using various metrics, including:

- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination ($R^2$)

## 1.4 Thesis Contribution

From the work developed in this thesis, the following papers have been published:

- J. Valido, D. Albuquerque, A. Ferreira, and D. Coutinho. "Electric Vehicle Driving Range Prediction with Neural Networks". In: Portuguese Conference on Pattern Recognition (RECPAD). Covilhã, Portugal, 2024 [2].
- J. Valido, D. Albuquerque, A. Ferreira, and D. Coutinho. "Assessing Dimensionality Reduction on Driving Range Estimation". In: Proceedings of the 12th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA). Poster Session 1. Coimbra, Portugal, 2025 [3].
- The code developed in this thesis is available at:

  `https://github.com/JoaoFranciscoValido/TFM_EV-X-DRP2.git`

## 1.5 Thesis Structure

The remainder of this document is structured as follows: Chapter 2 presents a comprehensive review of the current state of the art in electric vehicle (EV) range estimation, covering traditional and machine learning-based approaches, available datasets, and dimensionality reduction techniques. Chapter 3 describes the methodology developed in this thesis, including dataset construction, data preprocessing, application development, model implementation, and the integration of dimensionality reduction techniques. Chapter 4 reports the experimental results obtained with different regression models and dimensionality reduction methods, and evaluates their performance using standard metrics. Chapter 5 investigates the robustness of the History-Based Algorithm (HBA) by introducing Gaussian Noise and analyzing its impact on range prediction accuracy. Finally, Chapter 6 presents the overall conclusions and suggests directions for future work in this domain.

# 2 State of the Art

This chapter presents a comprehensive and structured review of the current approaches to Electric Vehicle (EV) range estimation, emphasizing the evolution of datasets, methodologies, and models used in this field. As EV adoption grows and sensor data becomes more widely available, Machine Learning (ML) and data-driven techniques have become central to developing more accurate and adaptive eRange prediction models.

The chapter begins with a general overview of eRange estimation (Section 2.1), highlighting the complexity of the task and the multitude of variables that influence range prediction, such as battery State of Charge (SoC), driver behavior, environmental factors, and vehicle specifications.

Section 2.2 explores the various datasets used for range prediction, categorizing them into empirical, synthetic, and hybrid datasets. It evaluates their structure, resolution, richness, and availability, all of which play a critical role in the performance and generalizability of prediction models.

Section 2.3 addresses pre-ML range estimation methods, including rule-based systems, lookup tables, history-based algorithms, and physics-driven models. While simple and interpretable, these approaches often fail to capture the nonlinear and context-dependent aspects of real-world driving.

Section 2.4 introduces machine learning-based methods, which have become a cornerstone in range prediction. It covers a variety of supervised learning models such as linear regression, tree-based methods (e.g., Random Forests, Extreme Gradient Boosting), and neural networks (e.g., Multilayer Perceptron, Long Short-Term Memory, Graph Neural Network), highlighting their ability to model complex, nonlinear relationships and adapt to varying driving conditions.

In Section 2.5, the focus shifts to unsupervised learning and dimensionality reduction techniques, which help improve model performance and training efficiency by simplifying high-dimensional datasets. It also addresses feature selection and reduction methods that enhance robustness and interpretability.

Section 2.6 explores real-time and distributed learning, with emphasis on on-board deployment strategies like model compression, federated learning, and online learning, which are critical for maintaining performance in real-world, resource-constrained environments. Discusses hybrid and physics-informed models, which combine domain knowledge with machine learning to improve prediction accuracy, particularly when data is limited or physical constraints must be respected.

Section 2.7 covers the importance of interpretability and uncertainty quantification in safety-critical systems like EV. Techniques such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Bayesian Neural Networks are reviewed as tools to increase user trust, enable model transparency, and manage range prediction risks.

Finally, Section 2.8 summarizes the key takeaways, highlighting emerging trends such as sensor data fusion, transfer learning, explainable Artificial Intelligence (AI), and trustworthy ML models. These innovations aim to make EV range prediction more accurate, robust, and suitable for real-world deployment.

## 2.1 eRange Estimation

Predicting the Electric Range (eRange) of Electric Vehicle (EV) with high accuracy is a complex problem due to the number of influencing variables, including vehicle parameters, environmental conditions, and driver behavior. The growing availability of sensor data, vehicular Telemetry Data, and advances in Machine Learning (ML) has opened new possibilities for building data-driven eRange models. Accurate prediction of EV range needs a comprehensive datasets encompassing both instantaneous driving parameters (e.g., speed, acceleration, state of charge) and aggregated trip-level data (e.g., energy consumption, trip duration). Traditional datasets, such as the Vehicle Energy Dataset (VED) Dataset [4] and ChargeCar Database [5], have provided foundational insights but often suffer from limitations in scale, diversity, or accessibility. Recent efforts have introduced more sophisticated datasets. For instance, the Digital Twin-Based Remaining Driving Range Prediction study utilized a year-long dataset from Beijing, incorporating features like battery state of charge, voltage metrics, and mileage to enhance range prediction accuracy [6]. Similarly, the EV-PINN approach leveraged in-situ battery log data from Tesla models to train physics-informed neural networks for dynamic EV behavior prediction [7]. These advancements underscore the importance of rich, diverse datasets in developing robust EV range prediction models.

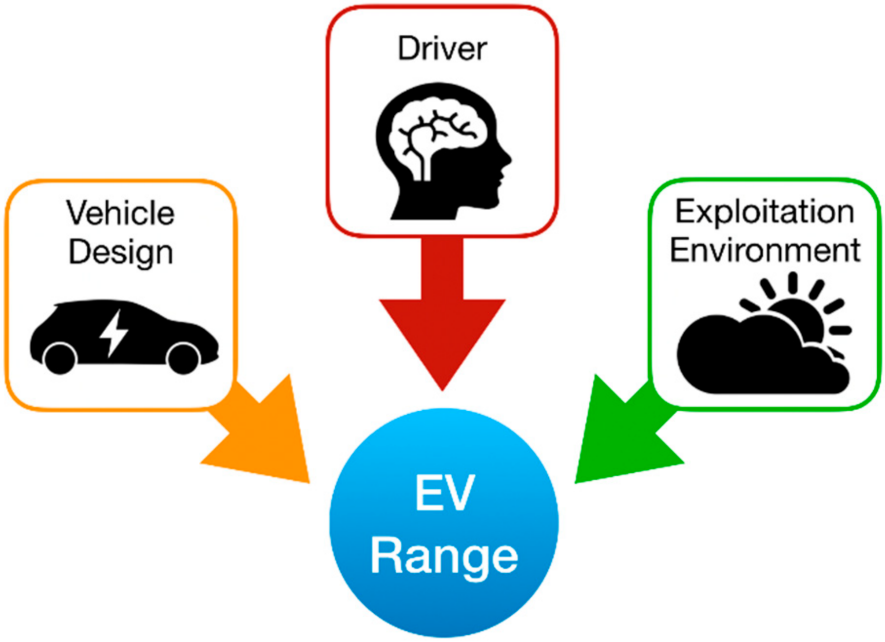


Figure 2.1: *Key factors affecting eRange: The vehicle's design parameters, the driver's behavior, and the external environment (e.g., weather, traffic, road slope) all interact to determine overall energy consumption and range performance (taken from [8]).*

## 2.2 Datasets for eRange Prediction

Before addressing the types of datasets used in eRange prediction, it's important to understand the broader system influences. Figure 2.1 presents a conceptual overview of the

primary domains affecting EV range: vehicle design, driver behavior, and the surrounding environment. These categories encompass the main sources of variability and uncertainty in predicting energy consumption and thus must be carefully considered when designing data-driven models. The foundation of accurate eRange prediction lies in the availability of rich, representative, and high-resolution datasets. These datasets must ideally contain granular, time-stamped records of a wide array of variables, such as battery state of charge (SoC), vehicle speed, instantaneous power consumption, regenerative braking activity, road grade, outside temperature, and Global Positioning System (GPS) coordinates. High-resolution temporal data (e.g., sampling intervals below 1 second) allow for the detection of short-term fluctuations and transient behaviors that significantly affect consumption. Structurally, datasets are often presented in either flat tabular format with fixed intervals or as multi-modal logs incorporating both continuous telemetry and discrete event labels (e.g., charge session starts, braking events). Some datasets include annotated features with physical or statistical metadata (e.g., variance, entropy), which support advanced preprocessing, dimensionality reduction, and model feature selection. The completeness, consistency, and contextual richness of such datasets directly influence the robustness and adaptability of eRange models, particularly in generalizing across vehicle types, driving styles, and environmental conditions. These datasets typically fall into three categories: empirical (collected from real-world driving), synthetic (simulated or generated from models), and hybrid (combinations of both). The foundation of any eRange prediction solution lies in access to robust and representative datasets. Typically, we have instantaneous data such as SOC, speed, acceleration, and road elevation) and trip-level data (e.g., average energy consumption, commute type, and total distance).

### 2.2.1 Empirical Datasets

Empirical datasets are derived from real-world driving data collected through onboard vehicle sensors, telemetry systems, or crowd-sourced platforms. These datasets offer high realism and reflect actual usage patterns, making them essential for developing reliable and generalizable eRange prediction models. However, they often present challenges such as limited access, inconsistent data quality, or lack of standardization.

Figure 2.2 illustrates a typical example of the structure of empirical EV datasets, showcasing both telemetry variables (e.g., SoC, Speed, Voltage (V), and Current) and annotated events (e.g., trip start, charging sessions, and GPS-based location tags). This multi-layered data is crucial for capturing the dynamic and contextual aspects of vehicle behavior.

The following paragraphs present key empirical datasets used in the literature, detailing their scope, features, and limitations in the context of EV range modeling.

The structure presented in Figure 2.2 is representative of the data format found in several empirical datasets, including the Vehicle Energy Dataset (VED). The VED dataset, which collects time-series data from real-world EV trips, aligns closely with this format by providing key telemetry parameters such as state of charge (SoC), speed, current, and distance, all indexed over time. This temporal granularity, illustrated in the figure, is essential for capturing short-term consumption dynamics and forms the basis for developing and

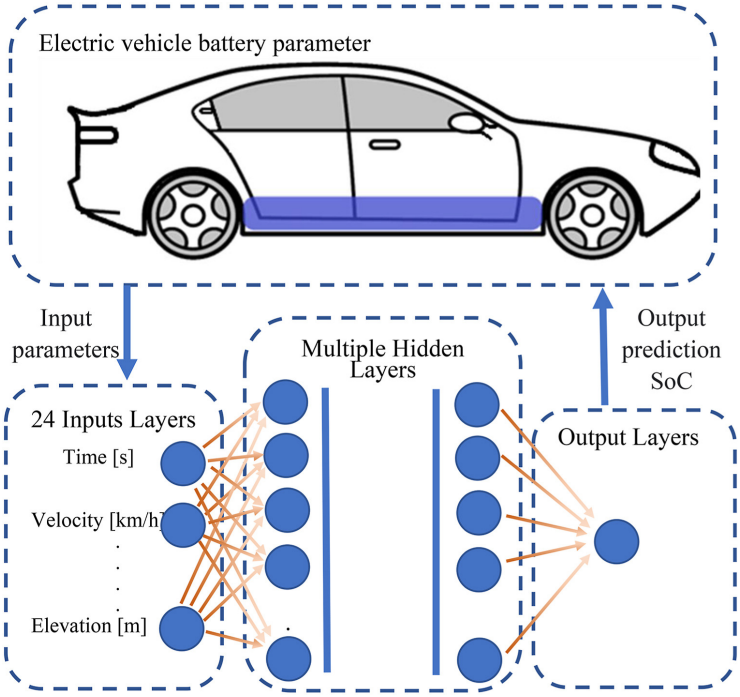


Figure 2.2: *Visualization of telemetry and event-based data collected from electric vehicles. Key features include time-indexed variables like SoC, speed, voltage, and current, as well as higher-level annotations such as trip start, charging sessions, and geographic tags (taken from [9]).*

validating time-dependent range prediction models.

- **VED Dataset -** The Vehicle Energy Dataset (VED) dataset [4], includes data from 54 real-world trips using three Nissan Leaf 2013 vehicles. Features include SoC, speed, current, and distance, with time series resolution. However, the dataset's scale and variety are limited.

- **ChargeCar Database -** A crowd-sourced dataset [5], Developed by Carnegie Mellon University, with data from hundreds of users. While it includes both EV and Internal Combustion Engine (ICE) vehicle traces, the latter must be filtered, and the heterogeneity of data sources introduces challenges in preprocessing and standardization.

- **JARI Dataset -** Collected by the Japan Automobile Research Institute, it includes probe data from over 500 EV with measurements of speed, acceleration, SoC, and geographic coordinates. Despite its scope, data access is limited [10] [11] [12].

- **NDANEV Dataset -** China's National Big Data Alliance of New Energy Vehicles provides a vast dataset of over 10 million kilometers of driving logs. Unique features include temperature at the cell level, GPS-based vehicle location, and detailed SoC information, enabling more robust eRange modeling [13].

- **Tesla Vehicle Logs -** Used in the EV-PINN framework [7], these logs include battery voltage, temperature, and current sampled at high frequency, ideal for deep

learning models requiring fine-grained temporal input.

### 2.2.2 Synthetic and Hybrid Datasets

- **Emobpy Tool -** A synthetic dataset generator based on empirical mobility statistics. It provides simulated trips that replicate common European driving behaviors [14]. However, its lack of real-time contextual factors (e.g., slope, traffic and weather) reduces its fidelity.

- **EV Database -** Though not a time-series dataset, this database aggregates public EV specifications, including nominal range, usable battery capacity, and WLTP/real-world energy consumption. It is useful for benchmarking and as a feature source.[15].

In summary, empirical datasets offer realism but are constrained by availability and coverage. Synthetic datasets offer flexibility and scalability but may lack realism. A hybrid approach that calibrates synthetic models on empirical data offers the best trade-off. While public datasets provide a strong foundation, challenges remain, including access restrictions, incomplete feature sets, and insufficient vehicle diversity.

## 2.3 eRange Prediction

Accurate electric vehicle range prediction is inherently tied to the precision of State of Charge (SOC) estimation, as it directly reflects the energy available for propulsion. Over time, several methods have been developed to estimate SOC, each varying in complexity, accuracy, and adaptability. Before the widespread use of machine learning, these estimation techniques were primarily deterministic, often based on physical principles or heuristic rules.

Figure 2.3 illustrates the five major categories of SOC estimation methods commonly found in the literature: Conventional Methods (CM), Adaptive Filter Algorithms (AFA), Learning Algorithms (LA), Non-Linear Observers (NLO), and Hybrid Methods (HM). Each category reflects a different modeling philosophy, ranging from analytical and empirical approaches to more recent data-driven and integrated techniques.

The remainder of this section reviews these traditional SOC and eRange prediction methods in detail, examining their operational logic, advantages, and the limitations that have motivated the shift toward more flexible and data-driven alternatives.

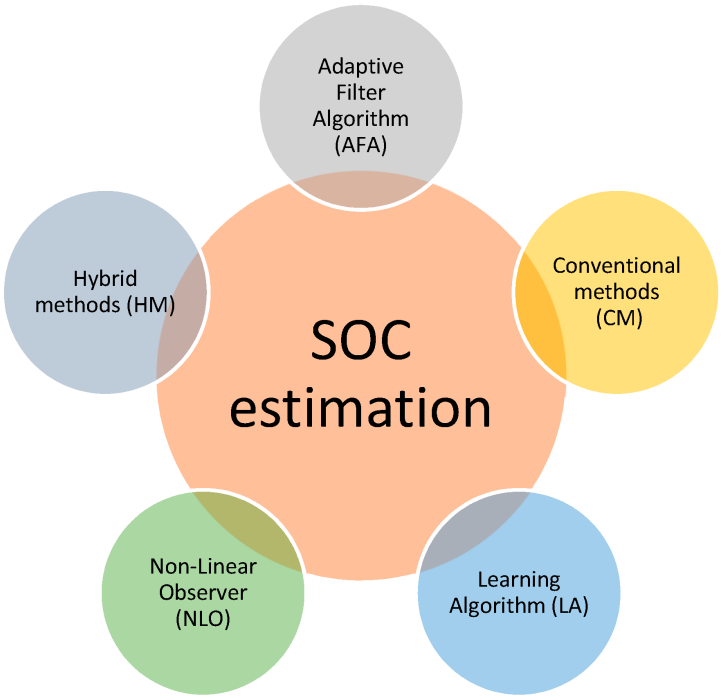


Figure 2.3: *Diagram showing the operational logic of rule-based, historical average, and physics-based models for range prediction. These approaches use simplified consumption models and static assumptions which limit adaptability in real-world contexts (taken from [8]).*

The prediction of range for an EV directly depends on the accuracy of the SOC prediction. This is because it provides primary information about the amount of available energy to be used by the EV's powertrain. Therefore, the accuracy of prediction (the magnitude of prediction errors) is an important factor in choosing and implementing a SOC estimation method within an EV's systems.

Before the emergence of machine learning, range prediction was tackled through deterministic and rule-based approaches, often grounded in physics or historical averages. Traditional eRange prediction approaches often rely on deterministic models using energy consumption rates or lookup tables. Some of these approches examples are:

- **Lookup Tables -** Some EV use predefined tables to estimate range based on remaining SoC and current energy consumption rate. While fast, they fail to generalize under atypical conditions.
- **Rule-Based Models -** These models incorporate simple heuristics, such as averaging the consumption over the last 5–10 km. While effective in steady conditions, they are unreliable in dynamic environments with frequent accelerations or elevation changes.
- **History-Based Algorithms (HBA) -** Adaptive methods proposed in works such as [16] dynamically compute energy consumption based on prior intervals. Despite improvements over static rules, they are sensitive to noise and offer limited generalization.

- **Physics-Based Models -** These include longitudinal dynamic models that compute consumption based on Newtonian equations factoring mass, air drag, rolling resistance, and drivetrain efficiency. They are highly accurate but require parameters that are often unavailable or difficult to calibrate in real-time. Sarrafan et al. [17] proposed a physics-based range estimation model that incorporates dynamically changing environmental conditions (e.g., wind, road slope, weather) and time-varying traction system losses. Unlike conventional deterministic models that assume constant efficiency parameters, this approach captures location-dependent variations in motor and inverter efficiency, resulting in significantly improved accuracy. Their method was validated with real-world driving experiments, showing precise state-of-charge and range predictions that help mitigate range anxiety.

Although simple and interpretable, these methods cannot effectively capture non-linearities or contextual variations in real-world driving. For example, during urban driving with frequent stop-and-go traffic, rule-based models often overestimate the remaining range due to their reliance on recent average consumption, which does not account for acceleration patterns. Similarly, physics-based models may underperform when faced with changing road gradients or weather conditions unless meticulously calibrated. In a study by Zhuo [18], deterministic range estimation approaches showed an average prediction error of over 20% when applied across mixed highway and urban scenarios, compared to under 10% for machine learning-based approaches. These shortcomings become especially evident when external variables such as ambient temperature, auxiliary power usage, or driver-specific behavior significantly impact consumption. A noteworthy hybrid approach combining physical modeling and adaptive filtering is presented by Sangeetha et al. [19]. The authors propose an Extended Kalman Filter (EKF)-based State of Charge (SoC) estimation algorithm, validated against a detailed vehicle dynamic model implemented in MATLAB Simulink. Their approach accounts for real-world driving factors such as road gradient, vehicle mass, rolling resistance, and environmental conditions. The model achieves exceptionally low Root Mean Square Error (RMSE) values ($<$0.03%) under different standard driving cycles (LA92, FTP-72, NEDC), demonstrating high reliability. This integration of a physics-based vehicle model enhances the robustness and realism of SoC estimation, making it a dependable basis for range prediction in electric vehicles. A notable hybrid approach was proposed by Hong et al. [20], who developed a high-fidelity remaining range estimation method by integrating a physics-based EV power consumption model with empirical regression techniques. Their method separates the estimation task into two phases: prediction of future driving profiles (velocity and acceleration) and subsequent energy consumption estimation. By incorporating real-time road slope, velocity, and vehicle-specific parameters, and enhancing the power model with dynamic motor efficiency and quadratic speed terms, the authors achieved a prediction error as low as 2.52%, significantly outperforming traditional model-based estimators. This hybrid framework highlights the importance of both model accuracy and fine-grained telemetry for reliable range prediction.

## 2.4 Machine Learning Models for eRange Prediction

With the increasing complexity of real-world driving environments and the availability of high-resolution vehicular data, traditional methods for eRange prediction have become insufficient in capturing nonlinear interactions and temporal dependencies. In response, machine learning (ML) has emerged as a powerful alternative, enabling data-driven models that adapt to various driving patterns, road conditions, and vehicle behaviors with significantly improved predictive accuracy.

Figure 2.4 provides an overview of a typical end-to-end machine learning pipeline for EV range prediction. The process begins with data ingestion, followed by preprocessing and feature engineering, which are critical for ensuring input quality. This is followed by model training, validation, and testing using appropriate performance metrics. In deployment scenarios, the final model is integrated into real-time systems through online inference mechanisms.

The remainder of this section explores different classes of machine learning models—ranging from simple linear regressors to complex neural networks—highlighting their architecture, strengths, limitations, and reported performance in the context of eRange prediction.

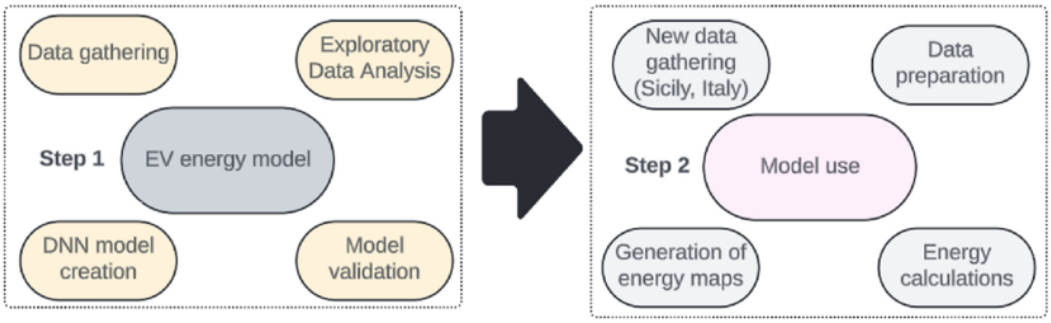


Figure 2.4: *End-to-end pipeline of a machine learning system for EV range prediction, including stages like data ingestion, preprocessing, feature selection, model training, evaluation, and online inference for real-time scenarios (taken from [21]).*

ML has become a cornerstone for addressing the complexities of eRange prediction. Supervised learning models, unsupervised clustering, and reinforcement learning have shown to achieve accurate predictions by leveraging large datasets. The shift to ML-based methods has enabled more adaptive and accurate range estimations by allowing models to learn complex, nonlinear relationships among input features such as speed, road elevation, traffic conditions, and driver behavior. Traditional rule-based and physics-driven models, while interpretable and fast, often fail under real-world variability, such as sudden weather changes or route deviations. For instance, ML models like Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) have shown superior performance in public benchmarks like the NDANEV dataset, with reductions in Mean Absolute Error (MAE) by up to 30% compared to Baseline Model algorithms.

Wei et al. [22] developed an online range estimation model using real-world BEV driving

data collected in Beijing. Their method segments the driving cycle based on 1% SOC intervals and applies Principal Component Analysis (PCA) combined with Fuzzy c-means clustering to classify driving behavior. An econometric model is then used to estimate the energy consumption rate for each segment type. The model demonstrated high accuracy in capturing seasonal and behavioral variations, and supports real-time range estimation under diverse operating conditions. Pan et al. [23] proposed a hybrid range estimation model that combines driving cycle identification with predictive modeling. Their method utilizes Kernel Principal Component Analysis (KPCA) and a fuzzy C-means clustering algorithm to classify driving conditions, followed by fuzzy rule-based reasoning and a Markov-BP neural network to forecast future conditions. This approach significantly improves range prediction accuracy by accounting for the dynamic nature of real-world driving patterns.In practical deployments, Tesla's use of neural networks for battery range estimation illustrates the capacity of ML systems to adapt to usage patterns over time, recalibrating predictions based on historical and contextual data. These advantages allow ML-based methods to offer not only better generalization but also the flexibility to be refined continuously as new data becomes available. These models can capture complex, non-linear relationships between input features and energy consumption. Beyond purely data-driven methods, optimization-based models that incorporate road network topology have emerged as effective alternatives for electric range prediction. Chkalov and Dropa [24] introduce a graph-based framework that models road segments as weighted directed edges, where weights correspond to energy consumption. Their approach integrates vehicle parameters, traffic data, road slope, intersections, and acceleration dynamics into an adjacency graph. A modified Bellman-Ford algorithm is used to compute energy-optimal eco-routes, enabling precise real-time mileage estimation. Unlike traditional average-consumption approaches, this method generates less conservative and more accurate driving range estimates while enabling spatial visualization on digital maps—an essential feature for navigation systems in electric vehicles.

### 2.4.1 Linear Models

Supervised learning techniques, such as decision trees, random forests, and Ensemble Stacked Generalization methods, have been widely applied. Ullah et al. [25] used an ensemble Stacked Generalization model with decision trees and K-nearest neighbors on the JARI dataset, demonstrating its effectiveness in minimizing overfitting. Zhao et al. [13] combined Extreme Gradient Boosting (XGBoost) and Light Gradient-Boosting Machine (LightGBM) to classify driving patterns, achieving high accuracy with the NDANEV dataset.

- **Linear Regression (LR)** remains widely used due to its simplicity and interpretability [26] [27]. It assumes a linear relationship between input variables (e.g., speed, power and State of Charge (SoC)) and output (Electric Range (eRange)), which limits its accuracy in diverse driving conditions. Despite this, LR often performs surprisingly well when inputs are preprocessed and normalized. It serves as a fundamental statistical method for modeling the relationship between a dependent variable and one or more independent variables. Its simplicity and interpretability

make it a common baseline in predictive modeling. In the context of EV range prediction, LR can model straightforward relationships between factors like speed and energy consumption. However, its linear nature limits its ability to capture complex, nonlinear interactions inherent in EV dynamics.

### 2.4.2 Tree-Based Models

- **Decision Trees (DT)** offer interpretable non-linear modeling but may suffer from overfitting. Random Forests (RF) mitigate this through ensembling, improving generalization. Gradient Boosted Trees, such as XGBoost and LightGBM, further enhance accuracy and speed and have shown strong performance on structured telemetry data [13].

### 2.4.3 Neural Networks

Neural networks, particularly multilayer perceptrons (MLP), as shown on Figure 2.5, have also been employed for non-linear relationships in range prediction. These methods benefit from hyperparameter tuning, including architecture optimization, activation functions, and learning rate adjustments [13], combined Extreme Gradient Boosting (XGBoost) and LightGBM to classify driving patterns, achieving high accuracy with the NDANEV dataset.

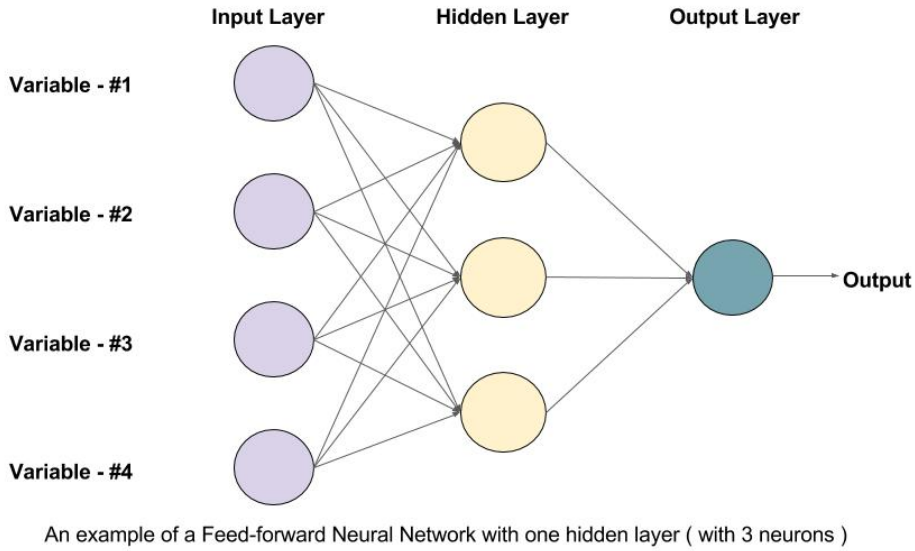


Figure 2.5: *Example of a feedforward Multilayer Perceptron (MLP) (taken from [28]).*

Neural networks have different topologies. Here are the most common topologies:

- **Multilayer Perceptrons (MLP) -** Fully connected neural networks that can approximate any continuous function. Are feedforward neural networks composed of multiple fully connected layers. They are particularly effective when combined with techniques such as batch normalization, dropout regularization, and activation functions like ReLU or tanh. MLP are widely used for range regression tasks due to their

ability to capture nonlinear interactions among variables such as energy consumption, road gradient, and average speed. When applied to EV telemetry data, they require tuning of architecture (number of layers, neurons) and regularization (dropout, weight decay) to prevent overfitting.

- **Radial Basis Function (RBF) Networks -** are particularly suited for problems with localized or region-specific behaviors. These networks use radial basis functions—typically Gaussian—in the hidden layer, which enables them to learn localized approximations. This architecture is useful in modeling abrupt behavioral shifts, such as regenerative braking events or sharp transitions in driving patterns. Due to their simplicity and efficiency, RBF perform well in smaller datasets and offer smooth interpolation capabilities [29] [30].

- **Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks -** These are designed for sequence modeling. LSTM architectures have been employed to model trip progressions and traffic conditions [31] [32]. They capture temporal dependencies, making them suitable for real-time applications.

- **Graph Neural Networks (GNN) -** Recently explored for modeling road networks and vehicle states jointly. GNN can generalize route-level consumption using topological data [18] [33].

Among the neural network architectures explored for eRange prediction, Radial Basis Function (RBF) networks stand out for their ability to approximate complex, non-linear functions using localized activation responses. Unlike multilayer perceptrons (MLP), which rely on global weight adjustments, RBF networks use radial basis functions—typically Gaussian kernels—in their hidden layer, making them particularly effective in scenarios with discrete or region-specific behavior patterns, such as regenerative braking events or sharp driving condition changes.

Figure 2.6 illustrates the architecture of a typical feedforward RBF network, where input features are mapped to a hidden layer of radial basis units, followed by a linear output layer. This structure allows for efficient learning in smaller datasets while maintaining good approximation capabilities.

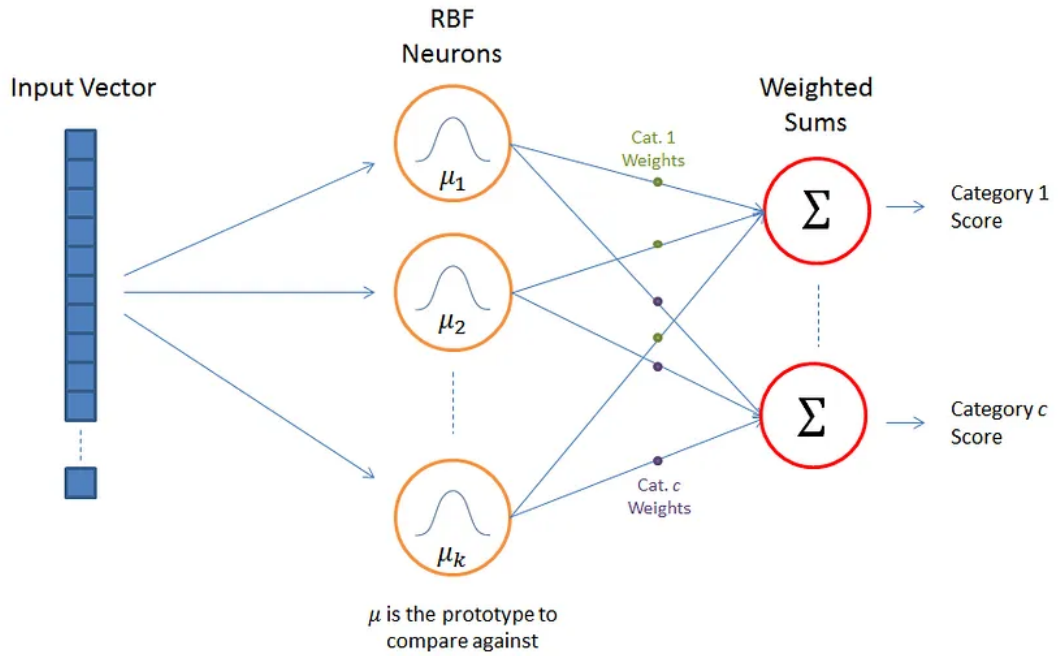


Figure 2.6: *Example of a Radial Basis Function (RBF) Network (taken from [34]).*

### 2.4.4 Relevant Related Work

Several recent studies have proposed advanced models for EV range estimation:

- Ullah et al. [35] explored the use of ensemble learning methods, specifically employing stacked generalization to combine algorithms such as k-Nearest Neighbor (KNN) and decision trees for predicting the energy consumption of electric vehicles in urban environments. Their study emphasized the importance of driver profiling, demonstrating that incorporating individual driving behaviors significantly improves prediction accuracy and enhances the model's adaptability to real-world usage scenarios.

- Yong et al. [36] applied neural networks for real-time EV range estimation. They demonstrated that multilayer perceptrons (MLP) and deep learning architectures with regularization techniques significantly improve prediction performance, especially under dynamic driving conditions.

- Modi et al. [37] proposed a hybrid predictive model that combines a Convolutional Neural Networks (CNN) with a bagged decision tree to optimize energy management and improve range estimation for electric vehicles in real-time. Their approach adapts to real-world variability in driving patterns and effectively profiles driver behavior, delivering more accurate and robust energy consumption predictions under diverse urban driving conditions

- Sarrafan et al. [38] investigated the impact of environmental conditions — including ambient temperature, road slope, and traffic congestion — on electric vehicle range estimation models. They incorporated ensemble learning techniques and accounted for dynamic traction system efficiency to improve both accuracy and explainability

of the predictions. Their method offers enhanced adaptability to real-world driving variability through careful modelling of external and drivetrain factors.

- Sayed et al. [39] introduced a deep learning model based on Long Short-Term Memory (LSTM) networks for fine-grained, time-series prediction of EV range. Their model leveraged multi-source sensor data and demonstrated high accuracy in capturing short-term fluctuations in energy consumption.

- Bai et al. [40] conducted a comparative study between Artificial Neural Networks (ANN) and hybrid models for predicting the remaining driving range of battery electric vehicles. Their findings indicated that Radial Basis Function (RBF) neural networks achieved superior accuracy, particularly when modeling localized behavioral patterns such as acceleration spikes or regenerative braking events. The RBF approach demonstrated robust performance in capturing dynamic driving behaviors under real-world conditions.

Despite the wide variety of methods explored, many models perform poorly when exposed to noisy or heterogeneous data. Moreover, the lack of standardization in public datasets complicates direct comparisons and realistic performance evaluations. Therefore, there is a clear need for more robust approaches and representative, high-quality datasets.

## 2.5 Unsupervised Learning and Dimensionality Reduction

Unsupervised clustering methods, such as Self-Organizing Maps (SOM), have been used to group similar driving patterns before applying regression models. For example, Zheng et al [41] demonstrated that hybrid SOM-regression tree models improve knowledge extraction while avoiding overfitting. High-dimensional EV data is prone to overfitting, redundancy, and increased computational cost. Dimensionality reduction (DR) addresses these issues by simplifying feature spaces, avoiding the curse of dimensionality

### 2.5.1 Reinforcement Learning

Reinforcement learning has been explored for its adaptability in dynamic environments. De Cauwer et al. [42] combined reinforcement learning with multiple linear regression to address external energy disturbances, achieving reduced prediction errors with the EVteclab dataset.

### 2.5.2 Emerging Trends

Recent trends include integrating physics-informed neural networks (PINN) [7] with ML to incorporate domain knowledge into models. Advances in transfer learning and federated learning have also enabled cross-domain model training without sharing sensitive data, addressing privacy concerns in large-scale EV data collection.

### 2.5.3 Feature Selection (FS)

High-dimensional EV data is prone to overfitting, redundancy, and increased computational cost. Dimensionality reduction (DR) addresses these issues by simplifying feature spaces, primarily through two approaches: Feature Selection (FS), which identifies and retains the most informative variables, through the process presented in Figure 2.7, and Feature Reduction (FR), which transforms the original features into a lower-dimensional representation while preserving essential information.

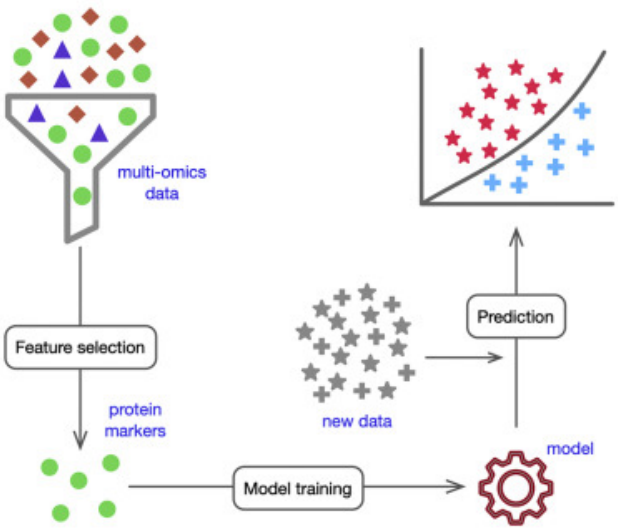


Figure 2.7: *Diagram representing feature reduction methods like Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and autoencoders. These techniques condense feature spaces while preserving variance or nonlinear structure, enhancing model robustness (taken from [43]).*

Feature selection methods are organised into three types of method:

- **Filter Method -** Metrics like Mean Absolute Deviation (MAD) and Mean-Median (MM) rank features based on variability and asymmetry [44].
- **Wrapper Methods -** These use a model to evaluate feature subsets iteratively. While accurate, they are computationally expensive.
- **Embedded Methods -** Techniques like Lasso or Tree-based feature importance embed selection into model training.
- **Maximal Margin (MM) -** This method selects features that maximize the margin between different classes, enhancing the discriminative power of the model. MM is particularly useful in classification tasks where clear separation between classes is desired.

- **Mean Absolute Deviation (MAD) -** Measures the average absolute deviation of each feature from its mean, identifying features with significant variability. Features with higher MAD values are often more informative and can improve model performance.

A comprehensive analysis of various feature selection methods, including their stability and performance across different datasets, was conducted recently [45].

### 2.5.4 Feature Reduction (FR)

Feature reduction (FR) techniques aim to transform high-dimensional feature spaces into lower-dimensional representations while preserving the most relevant information for prediction. Unlike feature selection, which chooses a subset of existing features, FR creates new composite features—often uncorrelated or compressed—that capture the underlying structure or variance in the data. These methods are particularly useful in EV applications, where telemetry datasets can include dozens of correlated or redundant variables, increasing the risk of overfitting and computational overhead.

Figure 2.8 illustrates common approaches to feature reduction used in eRange prediction pipelines. Techniques such as **Principal Component Analysis (PCA)** and **Singular Value Decomposition (SVD)** perform linear transformations to extract the most informative components, whereas autoencoders apply deep learning to learn nonlinear, compressed representations through unsupervised training.

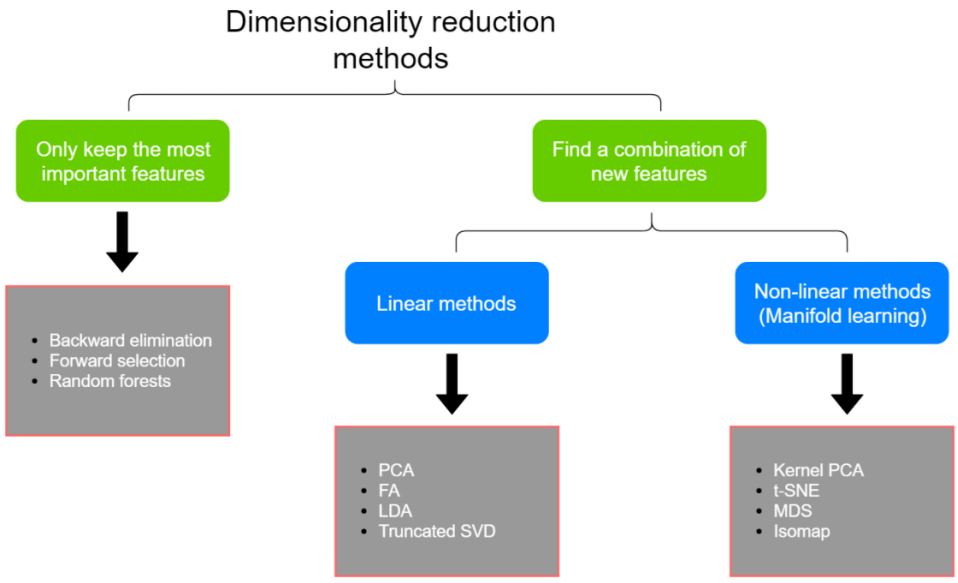


Figure 2.8: *Illustration of common feature reduction techniques used in EV range prediction. The figure presents three widely adopted methods: Principal Component Analysis (PCA), which projects data onto orthogonal components that maximize variance; Singular Value Decomposition (SVD), a matrix factorization approach effective in denoising and compression; and Autoencoders, deep neural networks trained to reconstruct input data through a lower-dimensional latent space, capturing nonlinear relationships among features (taken from [46]).*

- **PCA -** Projects data into orthogonal components maximizing variance [47]. Effective for linear correlations but can lose interpretability.
- **SVD -** Matrix decomposition approach closely related to PCA, effective in denoising and compression.
- **Autoencoders -** Deep networks trained to reconstruct input data. Latent layers serve as compressed representations that preserve non-linear relationships.

DR improves generalization, training efficiency, and in some cases, accuracy, especially when input features are correlated or noisy. For instance, in a study using the NDANEV dataset, applying PCA to reduce the feature space from 30 to 6 components resulted in a 25% reduction in training time and a 12% improvement in $R^2$ score for a LightGBM regressor. Another experiment conducted by Ferreira and Figueiredo [44] demonstrated that MAD-based feature selection significantly reduced overfitting and improved prediction accuracy in high-dimensional data scenarios by isolating features with the most variability. These case studies underscore the practical benefits of dimensionality reduction in enhancing model performance, particularly in noisy or redundant data environments.

## 2.6 Real-Time, Distributed Learning and Hybrid Approaches

As electric vehicle (EV) range prediction systems transition from research to real-world deployment, they must address practical challenges such as limited onboard computational resources, real-time processing constraints, and privacy-sensitive data usage. To meet these requirements, recent advances have focused on **real-time inference**, **distributed learning architectures**, and **hybrid modeling strategies** that combine data-driven techniques with physical system knowledge.

One prominent approach is the deployment of **compressed machine learning models** [48] -using methods such as pruning and quantization—on embedded hardware platforms (e.g., ECUs and onboard GPUs). This enables *low-latency, on-device predictions*, facilitating real-time assistance for drivers and vehicle subsystems.

Simultaneously, **federated learning (FL)** [49] [50] has gained traction as a privacy-preserving strategy for collaborative model training across multiple EVs. FL avoids centralizing raw data by aggregating model updates locally, thereby enhancing generalization through exposure to diverse driving scenarios without compromising user privacy. Complementing this, **online learning** [51] [52] allows models to adapt incrementally based on recent streaming data, ensuring sustained accuracy in dynamic operational environments. Dong et al. [53] developed a real-time energy consumption prediction framework for electric buses using integrated machine learning models. Their approach combines high-resolution GPS and OBD data with weather and road conditions to extract kinematic and environmental features. By integrating LSTM and fully connected neural networks for kinematic forecasting with an XGBoost-based consumption estimator, their model achieves prediction errors as low as 7.5% over 16km, demonstrating superior performance in distance-based

energy prediction. Additionally, SHAP analysis is employed to interpret feature importance dynamically across different driving segments.

Beyond infrastructure strategies, **hybrid models** [54] have emerged that integrate *physics-based insights* into machine learning frameworks. For example, **Physics-Informed Neural Networks (PINNs)** [7] incorporate physical laws—such as energy conservation—into their loss functions, guiding learning towards physically consistent outcomes. Other hybrid systems combine simulation-based outputs (e.g., from Simulink) with machine-learned residuals to improve predictive precision in data-scarce conditions.

An advanced implementation of this concept is found in **digital twin systems** [55], which create synchronized virtual replicas of physical EVs. These systems assimilate real-time sensor data to simulate consumption patterns, battery dynamics, and environmental interactions, thereby enhancing range estimation, predictive maintenance, and system optimization.

Together, these approaches enable the design of *scalable, adaptive, and interpretable* EV range prediction models. They not only improve prediction accuracy but also ensure operational feasibility, positioning EV technologies for widespread, reliable, and efficient deployment.

In addition to hybrid learning techniques, commercial tools such as the GTI EV Simulator [56] have been used to predict EV range by integrating contextual factors like weather, traffic, terrain, and vehicle dynamics. These simulators provide a robust basis for generating high-fidelity synthetic datasets used for training and validating prediction models.

Furthermore, industry-oriented solutions are represented by patents such as US20120109408A1 [57], which describe embedded systems for dynamic range estimation using real-time sensor data combined with neural networks. These systems exemplify how machine learning can be practically integrated into EV onboard architectures for operational deployment.

Several recent works have explored the integration of model-based strategies to enhance the accuracy of electric vehicle (EV) range estimation. Hong et al. [20] proposed a hybrid modeling framework that combines empirical measurements with dynamic vehicle modeling, introducing a hybrid power model that improves real-time range prediction. Their approach incorporates regenerative braking and drivetrain losses, and was validated on a custom EV testbed, achieving significantly lower range estimation errors compared to traditional dynamic models.

Complementing this, De Nunzio and Thibault [58] introduced a model-based predictive strategy using macroscopic road and traffic data to compute energy-optimal driving ranges. Their method leverages an adjoint graph representation of the road network and a modified Bellman-Ford algorithm to compute reachable destinations via eco-routes. This allows a more accurate and less conservative range prediction compared to distance-based estimators, especially in urban environments where elevation, traffic lights, and auxiliary power demands (e.g., air conditioning) can significantly affect energy consumption.

These approaches highlight the value of incorporating physical and topological context into driving range prediction, moving beyond purely statistical or history-based methods.

Such techniques align with the goals of this thesis, which aims to enhance range prediction robustness through machine learning, while acknowledging the relevance of physical constraints and system dynamics.

### Hybrid and AI-Augmented EV Range Estimation

In recent years, several hybrid approaches have emerged that combine machine learning with physics-based modeling to enhance the accuracy, robustness, and interpretability of driving range predictions in electric vehicles (EVs).

Bustos et al. [59] proposed a novel framework for predicting the Maximum Driving Range (MDR), introducing a spatially-aware concept of "hazard zones" where battery disconnection is more likely. Their method integrates stochastic LSTM networks for velocity prediction and LightGBM models for power and energy forecasting. Additionally, a Thévenin-equivalent circuit was used to approximate battery behavior, making the system robust to a wide range of operating conditions. This approach was validated through real-world case studies in Costa Rica, offering new insights into proactive route planning and battery risk mitigation.

Complementing this, Cavus et al. [60] provided a comprehensive review of AI-driven Battery Management Systems (BMS). They highlighted the role of deep learning, reinforcement learning, and fuzzy logic in predicting battery State of Charge (SoC), State of Health (SoH), and enabling predictive maintenance. Their work supports the idea that next-generation BMSs, enhanced with AI and Internet of Things (IoT) integration, are essential for maximizing EV efficiency, battery lifespan, and safety, particularly under highly variable environmental and operational conditions.

These contributions, along with earlier model-based approaches like those of Hong et al. [20] and De Nunzio et al. [58], demonstrate the clear evolution of EV range estimation from empirical and deterministic models to hybrid and intelligent predictive systems. The growing integration of probabilistic modeling, route-aware segmentation, and real-time AI processing reflects an industry-wide shift toward more adaptive and personalized mobility forecasting solutions.

## 2.7 Interpretability and Uncertainty Quantification

For safety-critical applications like EVs, models must be explainable and transparent. Several explainability frameworks exist, each with distinct advantages and trade-offs. For instance, SHAP (SHapley Additive exPlanations) [61] provides global and local interpretability through additive feature attributions, but can be computationally expensive for large datasets or complex models. LIME (Local Interpretable Model-agnostic Explanations) [62] offers localized linear approximations and is more efficient but may yield unstable explanations depending on input perturbations. In EV contexts, SHAP has been used to highlight which trip features (e.g., slope, speed variability and Heating, Ventilation, and Air Conditioning (HVAC) usage) most affect battery consumption predictions, providing useful insights for eco-driving assistance systems. Meanwhile, feature permutation

importance methods are simpler and faster but lack instance-level detail, limiting their usefulness in debugging anomalous predictions. The selection of an explainability tool thus depends on deployment goals-whether the emphasis is on real-time decision support, post hoc diagnostics, or regulatory transparency. The main properties of these methods are as follows:

- **SHAP and LIME -** Post-hoc explanation tools highlight the contribution of each feature to a given prediction [63].
- **Bayesian Neural Network (BNN) -** Model uncertainty by learning distributions over weights, producing predictive intervals [64].
- **Trust Calibration -** Tools to adjust confidence scores to reflect actual model reliability help mitigate "range anxiety"[65].

These tools are crucial for user trust, regulatory compliance, and debugging of eRange systems.

## 2.8 Summary

While NN offer unparalleled flexibility and predictive power, they come with challenges, including high computational requirements, potential overfitting with limited data, and the need for extensive hyperparameter tuning. Future research aims to address these issues by integrating domain knowledge into Neural Networks (NN) [7] architectures, employing transfer learning for improved generalization, and developing eXplainable AI (XAI) techniques to enhance the interpretability of NN-based eRange models [63]. Neural networks have revolutionized eRange prediction by providing scalable and highly accurate solutions. Their adaptability to diverse data types and driving conditions makes them a vital component of modern EV range estimation systems [66]. As computational resources and dataset quality continue to improve, the role of NN in this field is expected to grow further. Advancements in machine learning and data availability have significantly improved electric vehicle range prediction. Techniques like Linear Regression, MLP, and RBF networks offer varying degrees of complexity and interpretability, catering to different modeling needs. Dimensionality reduction methods, including MM, MAD, PCA, and SVD, play a crucial role in managing high-dimensional data, enhancing model performance and efficiency. The integration of physics-based insights with machine learning models represents a promising direction, offering improved accuracy and robustness in EV range prediction [47].

EV range prediction has evolved from deterministic rules to sophisticated, hybrid, and context-aware machine learning systems. As data availability and computing power improve, future research will likely emphasize:

- Fusion of sensor modalities (traffic, weather and map data)
- **Bayesian Neural Networks -** Model uncertainty by learning distributions over weights, producing predictive intervals [64].

- Transfer learning for vehicle adaptation.
- Real-time, onboard deployment.
- Trustworthy AI via explainable and uncertainty-aware models.

These directions will ensure that predictive systems are not only accurate but also scalable, secure, and trustworthy.

# 3 Proposed Approach

This chapter presents the methodological framework developed to estimate the electric vehicle driving range (eRange) using machine learning techniques. The proposed approach is organized into seven sections, each addressing a key component of the overall pipeline.

Section 3.1 introduces a block diagram that outlines the structure of the proposed methodology. It highlights the three main stages of the process: data collection and preparation, definition of Evaluation Metrics, and the development of an application for model training and analysis that integrates regression models and dimensionality reduction techniques.

Section 3.2 describes the construction of the dataset used in this study. It details the selection of eleven relevant features such as state of charge (SoC), power consumption, speed, and distance, as well as vehicle-specific attributes. Since the eRange target variable is not directly available in the raw data, it was estimated using the History-Based Algorithm (HBA), enabling the dataset to be used for supervised regression tasks. Also focuses on data collection and preparation. It explains how data from public sources such as the Vehicle Energy Dataset (VED), ChargeCar, and Classic EV X Project were integrated to ensure diversity and robustness. This section also includes a justification for the use of dimensionality reduction techniques, which were applied to eliminate redundant or noisy features, reduce computational cost, and improve the generalization ability of the models.

Section 3.3 provides an overview of the final dataset composition. It confirms that the dataset contains 2,176 instances and 11 features, and includes a table with detailed descriptions of each feature used in the predictive models.

Section 3.4 describes the development of a software application that supports the experimental process. This application enables efficient training, evaluation, and comparison of the models under various configurations.

Section 3.5 presents the implementation of the machine learning models. It includes traditional methods like Linear Regression, ensemble-based methods such as Random Forest and Stacked Generalization, and neural network architectures including Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF), each chosen to explore different modeling capabilities.

Finally, Section 3.6 outlines the application of dimensionality reduction (through feature selection and feature reduction), the training of the regression models, and their evaluation using standard metrics such as MAE, MSE, RMSE, MAPE, and $R^2$. The section also compares model performance with and without dimensionality reduction to assess the impact of feature space optimization.

## 3.1 Block Diagram of the Proposed Approach

The methodology adopted in this study was carefully structured to create, train, and evaluate machine learning models focused on predicting the driving range of electric vehicles (eRange). The approach is centered on three main stages: data collection and preparation, definition of evaluation metrics, and the development of an application for experimentation and analysis, with regression models and dimensionality reduction. The block diagram of the proposed approach is depicted in Figure 3.1.

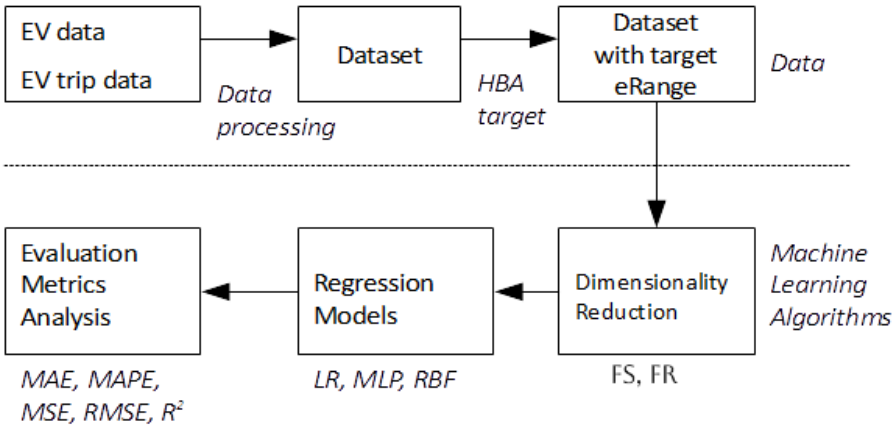


Figure 3.1: *Proposed approach to the eRange estimation problem.*

**General considerations about the learning problem.** Following Figure 3.1, we highlight four key aspects that distinguish this work from conventional machine learning pipelines:

1. **Regression formulation and model spectrum** – The task of estimating the driving range at each instant is framed as a *supervised regression* problem. We explore both simple statistical models (e.g., Linear Regression) and more flexible nonlinear models such as MLP and RBF networks, covering a broad bias–variance trade-off.

2. **Absence of an off-the-shelf dataset** – Unlike popular benchmarks in domains like image classification or NLP, there is no public dataset that directly links raw EV telemetry to ground-truth driving range. Thus, we constructed our dataset by merging and processing heterogeneous logs (VED, ChargeCar, Classic EV X) tailored for this purpose.

3. **Target definition** – Since real-world logs do not provide the range directly, we synthetically defined a target variable using the History-Based Algorithm (HBA). Its plausibility was validated against existing energy consumption formulas.

4. **Redundancy and dimensionality reduction** – Telemetry data shows a high degree of redundancy (e.g., power vs. current × voltage). We addressed this with explicit dimensionality-reduction techniques, including unsupervised feature selection (MM/MAD filters) and transformation-based methods (PCA, SVD).

These considerations underscore the specificity of the problem and the engineering effort required before any learning model can be applied.

## 3.2 Dataset Construction

The dataset is constructed from time-series EV trip data, incorporating key features such as SoC, power consumption, speed, and distance, along with vehicle-specific attributes. A total of 11 features are used, as detailed in Table 3.1.

Table 3.1: Feature Description

| **Feature** | **Description** |
|---|---|
| FBD (Full Battery Distance) | Total distance traveled with a full battery |
| FBE (Full Battery Energy) | Total energy available with a full battery |
| AEC (Average Energy Consumption) | Average energy consumption |
| SoC (State of Charge) | Battery charge status |
| timestamp | Measurement timestamp |
| ac_power | Alternating current power |
| speed | Vehicle speed |
| current | Measured electrical current |
| iec_power | Power according to IEC Power standard |
| power | Total Power (kW) |
| distance | Distance traveled |

Since the original datasets do not include a direct *eRange* value, the History-Based Algorithm (HBA) method is applied to compute the target variable. This allows the construction of a labeled dataset for regression tasks.

The study uses publicly available data from VED and ChargeCar, resulting in a dataset with 2,176 instances. Three regression models are trained: Linear Regression (LR) as a traditional baseline, Multi-Layer Perceptron (MLP) as a neural network, and Radial Basis Function (RBF) as an alternative topology. Implementation is based on the Scikit-Learn library to ensure flexibility and reproducibility.

Model performance is evaluated using MAE, MSE, MAPE, RMSE, and $R^2$, aiming to minimize errors and maximize predictive accuracy.

# Data Collection and Preparation

The performance of machine learning models depends directly on the quality and diversity of the data used during training. In this context, a dataset was compiled from publicly available sources, such as the Vehicle Eneergy Dataset (VED) [4], ChargeCar [5], and Classic EV X Project datasets [67]. These datasets provide essential information, including time-series data of power consumption, battery state of charge (SOC), speed, and traveled distances. Additionally, metadata specific to vehicles, such as Full Battery Energy (FBE), Full Battery Distance (FBD), and Average Energy Consumption (AEC) were integrated.

To ensure model robustness and avoid overfitting, data from different electric vehicle models and driving scenarios were combined [68]. A detailed preprocessing step was undertaken to fill gaps in the data, estimating missing variables from external sources like the Electric Vehicle Database [15]. This effort ensured a diverse and complete dataset, which is critical for model generalization. Furthermore, dimensionality reduction techniques were applied to improve model efficiency and performance. High-dimensional data often contain redundant or irrelevant features that can negatively impact learning algorithms, leading to overfitting, increased computational cost, and reduced interpretability. By applying unsupervised feature selection and feature reduction methods, we aim to retain the most informative attributes, minimize noise, and enhance the generalization capability of the regression models. This step is particularly important when working with real-world data, which can exhibit high variability and noise.

### Data Normalization

The heterogeneity in feature scales within the dataset can adversely affect the performance of machine learning algorithms, particularly those that rely on gradient-based optimization, such as artificial neural networks. To address this issue, feature normalization was applied using the Min-Max Scaling technique, which ensures that all numerical features are rescaled to the range $[0, 1]$. This transformation was performed using the `MinMaxScaler` class provided by the `scikit-learn` library.

Formally, the normalization process is defined as:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.1}$$

where $x$ represents the original value of a given feature, and $x_{\min}$ and $x_{\max}$ correspond to the minimum and maximum values of that feature, respectively, computed over the training set.

It is important to emphasize that $x_{\min}$ and $x_{\max}$ were calculated solely from the training data and subsequently applied to the test set. This approach prevents *data leakage*, thereby ensuring that the model evaluation reflects its generalization capability and does not inadvertently incorporate future information during training.

The choice of Min-Max scaling was deemed appropriate for the characteristics of the dataset, as it preserves the original distribution of the features and does not assume statistical

normality. Moreover, it contributes to more stable and efficient model convergence during the optimization process.

## 3.3 Dataset Composition and Target Definition

First, we integrated time-series EV trip data, with features such as SoC, power consumption, distance and speed, as well as vehicle-specific attributes, like energy consumption and power output. We have gathered data from the VED and ChargeCar datasets, to construct our dataset with $n = 2176$ instances and $d = 11$ features. The details of these features are described in Table 3.2.

The original dataset lacks a target attribute with eRange (not provided by car manufacturers). Thus, the History-Based Algorithm (HBA) eRange estimation [67], generates this target value for the evaluated regression techniques. Optionally, to this HBA-derived target value we may add a controllable Gaussian perturbation, in order to simulate a more realistic environment and enable the models to learn under noisy conditions. This approach is further analyzed in Chapter 5.

Table 3.2: Dataset feature description (n=2176 instances and d=11 features)

| **Feature** | **Description** |
|---|---|
| 1.AEC (Average Energy Consumption) | Average energy consumption |
| 2.AC Power | Alternating current power |
| 3.Current | Measured electrical current |
| 4.Distance | Distance traveled |
| 5.FBD (Full Battery Distance) | Total distance traveled with a full battery |
| 6.FBE (Full Battery Energy) | Total energy available with a full battery |
| 7.IEC (Instant Energy Consumption) Power | Power according to IEC standard |
| 8.Power | Total power |
| 9.Speed | Vehicle speed |
| 10.SoC (State of Charge) | Battery charge status |
| 11.Timestamp | Measurement timestamp |

## 3.4 Application Development

A Python application was developed to manage experiments and enable customization of parameters. The application includes:

- **Dataset configuration -** Allows the selection of specific datasets and the definition of training conditions, such as minimum driving time to avoid biases caused by short or stationary trips.

- **Algorithm comparison -** Integrates various prediction models, including traditional approaches like the basic and history-based models, as well as advanced machine learning techniques such as linear regression, Random Forest, and Ensemble Stacked Generalization.

- **Results visualization -** Displays graphical comparisons and tables with prediction results, enabling a clear analysis of each algorithm's performance.

## 3.5 Model Implementation

The machine learning models used were trained based on the supervised learning paradigm. During training, the algorithms learned to map input features (speed, SOC, consumption, etc.) to the expected output (eRange). A k-fold cross-validation process was adopted to prevent bias and ensure a more accurate assessment of model performance.

The main algorithms implemented include:

- **Linear Regression -** Used as a baseline for its simplicity while computationally efficient [27].
- **Ensemble Stacked Generalization -** Combines multiple models (such as K-Nearest Neighbors, Random Forest, and Decision Trees) to improve accuracy.
- **Tree-based models -** Include Decision Trees and Random Forest, known for capturing complex relationships between variables.
- **Multi-Layer Perceptron Neural Network (MLP NN) -** A fully connected feedforward neural network capable of learning complex patterns through multiple hidden layers.
- **Radial Basis Function Neural Network (RBF NN) -** A neural network that uses radial basis functions as activation functions, well-suited for capturing nonlinear relationships in the dataset.

The performance of each model was compared with non-machine learning approaches, such as the "basic"and "history-based"methods, which use adaptive averages of energy consumption to estimate range.

With this methodological framework, it was possible to explore both the strengths and limitations of the techniques used, establishing a solid foundation for integrating future improvements and customizations based on the specific objectives of the project [16].

## Different Topologies

During the development of the MLP and RBF regression models, a wide range of network topologies and hyperparameter configurations were systematically tested to evaluate their impact on predictive performance. The objective was to identify configurations that achieved an optimal balance between accuracy, generalization, and computational efficiency.

### Multilayer Perceptron (MLP)

For the Multilayer Perceptron, several configurations were explored, varying the number of hidden layers, layer sizes, activation functions (e.g., *tanh*, *relu*), learning rates, regularization

terms (*alpha*), and solvers (such as *adam* and *sgd*). Among the different configurations tested, the most effective topology in terms of both predictive accuracy and stability used the following parameters:

- `hidden_layer_sizes=(10,)`
- `activation='tanh'`
- `solver='adam'`
- `alpha=0.01`
- `learning_rate_init=0.0001`
- `max_iter=1000`
- `early_stopping=True`
- `n_iter_no_change=10`
- `validation_fraction=0.1`

This configuration allowed for effective training while mitigating the risk of overfitting through early stopping and validation monitoring. The single hidden layer with ten neurons, combined with the *tanh* activation function, provided a suitable level of model complexity for the regression task at hand.

### Radial Basis Function (RBF) Network

In the case of the Radial Basis Function network, the design followed a kernel-based transformation followed by a ridge regression output layer. Different kernel scales (*gamma*), number of basis functions (*n_ components*), and regularization strengths (*alpha*) were tested. The configuration that yielded the best results was:

- `gamma=0.001`
- `alpha=0.0015`
- `n_components=100`

The RBF transformation was implemented using a Gaussian kernel computed over a set of learned centers, and the transformed features were then passed to a ridge regression model with high tolerance precision (`tol=1e-5`). This setup offered a good compromise between model expressiveness and computational demand, demonstrating consistent performance across different experimental scenarios.

### Critical Evaluation

The final topologies were selected based on empirical experimentation and iterative refinement. More complex MLP configurations—such as deeper architectures or larger hidden layers—were tested but showed signs of overfitting or diminishing returns relative to their computational cost. Similarly, RBF models with higher numbers of components or excessively low `gamma` values led to increased computational burden without significant gains in accuracy. The selected configurations proved to be robust across different datasets and evaluation metrics, offering both interpretability and practical efficiency.

## 3.6 Dimensionality Reduction

For dimensionality reduction we consider unsupervised Feature Selection (FS) and Feature Reduction (FR) approaches. For unsupervised FS, we use relevance-based filters as proposed by Ferreira and Figueiredo [44], with the relevance of feature $X_i$ assessed by the Mean Absolute Difference (MAD), defined as

$$\text{MAD}_i = \frac{1}{n}\sum_{j=1}^{n}\left|X_{ij} - \overline{X}_i\right|, \tag{3.2}$$

which computes the average absolute difference from all the feature values to its mean (average) value. We also consider the Mean-Median (MM) metric, computed for feature $X_i$, as

$$\text{MM}_i = |\overline{X}_i - \text{median}(X_i)|, \tag{3.3}$$

*i.e.*, the absolute difference between the mean (average) and median of $X_i$. MM is an asymmetry measure, that is, the more asymmetric the distribution of the values of the feature, the higher the MM metric.

For unsupervised FR, we have considered Principal Component Analysis (PCA) [47] and Singular Value Decomposition (SVD) [69] techniques. For both FS and FR techniques, the choice of the number of features of the reduced dimensionality space is given the Cumulative Relevance (CR) criterion. Let $r_{i_1}, ..., r_{i_d}$ be the sorted relevance values, in decreasing order, and $c_v = \sum_{f=1}^{v} r_{i_f}$, be the CR of the top $v$ most relevant features. We propose choosing the number of features as the lowest value $m$ that satisfies

$$\sum_{f=1}^{m} r_{i_f} / \sum_{i=1}^{d} r_i = c_m/c_d \geq L, \tag{3.4}$$

where $L$ is some threshold (*e.g.*, 0.95), leading to the choice of a fraction of the top-$m$ ranked features. For PCA and SVD, the relevance values are the eigenvalues and the singular values, respectively.

# 4 Experimental Results

In this Chapter, we report the experimental evaluation of our approach. The test conditions and the evaluation metrics are detailed in Section 4.1. The baseline results, using all the features on the dataset, are reported in Section 4.2. The results of dimensionality reduction with FS and FR approaches are reported in Section 4.3 and Section 4.4, respectively. Section 4.5 investigates the influence of data normalization on model performance in both feature selection and feature reduction contexts. It is important to highlight that, up to the end of Section 4.4, all results were obtained using the original data—i.e., without any form of scaling or normalization. These datasets preserve the raw values of the VED variables, maintaining their original units and scales. Section 4.5 onwards, the data was scaled in order to assess the impact of normalization preprocessing on the various electric range prediction approaches. This distinction allows for a clearer comparative analysis between models trained on unprocessed data and those optimized through scaling techniques. Section 4.6 provides a comprehensive discussion of the results, comparing different models and techniques to derive meaningful insights.

## 4.1 Test Conditions and Evaluation Metrics

We have considered a conventional 42 minutes trip from the VED dataset, referring to a 2013 Nissan Leaf model, with training data taken from the same dataset. The HBA method uses $\Delta S = 50$ W, a minimum instance energy of 2.5 kW and $M = 10$. The experiments were carried out on an AMD Ryzen 7 7800X3D processor running Windows operating system. The Python runtime has version 3.9.13, using Jetbrain's Pycharm.

The model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination ($R^2$). The objective is to minimize MAE, (R)MSE, and MAPE while maximizing $R^2$ to achieve robust and accurate estimations. MAE is the average of the absolute difference between the actual and predicted values is defined by

$$MAE(y_i, \hat{y}_i) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{4.1}$$

where $y_i$ is the actual value and $\hat{y}_i$ is the estimated value and $n$ is the number of instances

on the dataset. MAPE is similar to MAE, but it performs a normalization of the differences by $y_i$. It is expressed as a percentage, given by

$$MAPE(y_i, \hat{y}_i) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \ . \tag{4.2}$$

MSE averages the squared difference between the original and estimated values, and is defined by

$$MSE(y_i, \hat{y}_i) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ . \tag{4.3}$$

The Root MSE (RMSE) is defined as the square root of MSE. Finally, $R^2$ represents the proportion of variance in the dependent variable, given by

$$R^2(y_i, \hat{y}_i) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}. \tag{4.4}$$

In this expression, $y_i$ denotes the actual value of the target variable for the i-th instance, $\hat{y}_i$ represents the corresponding predicted value produced by the model, and $\bar{y}_i$ is the mean of all actual values.

Higher values for $R^2$ are desirable, as lower values correspond to redundant or irrelevant variables. We aim to minimize the MAE, MAPE, and (R)MSE metrics. Since $R^2$ ranges from -1 (worst) to 1 (best), we aim to maximize it.

## 4.2 Baseline Results - All Features

The application presents various eRange prediction results for the selected trip and forecasting algorithms, providing overview of the different dataset parameters. This allows the initial input dataset configuration to be based on multiple datasets. A conventional 42-minute trip was simulated using data from the VED dataset for a 2013 Nissan Leaf model, with training data taken from the same dataset, as shown in Figure 4.1 and Figure 4.2.

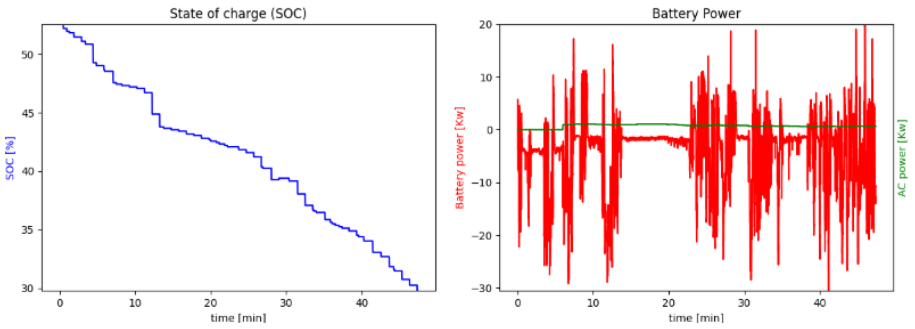


Figure 4.1: *Dataset overview showing the distribution of feature values. Includes histograms and statistics for SoC and power respectively.*

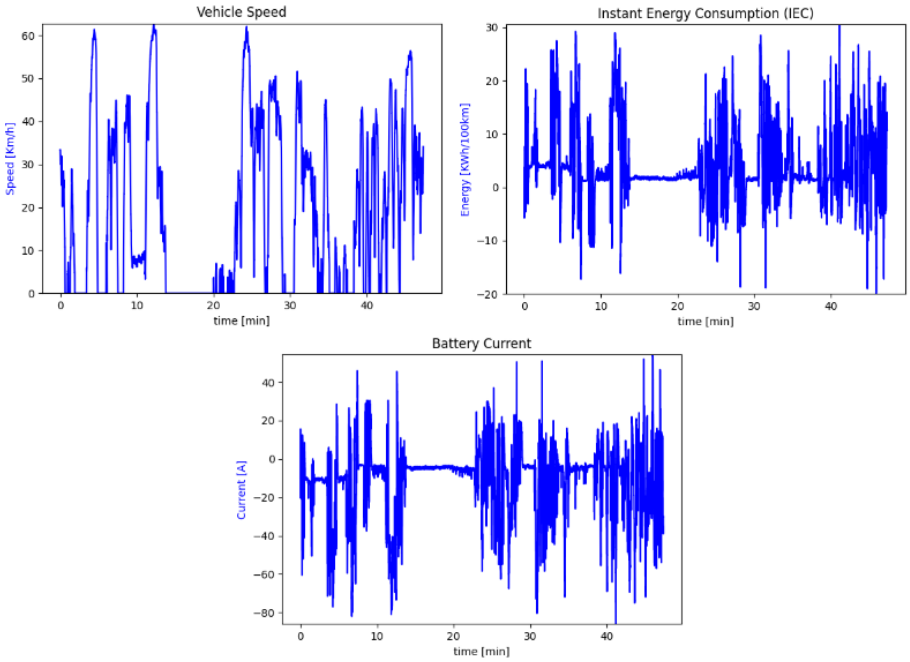


Figure 4.2: *Dataset overview showing the distribution of feature values. Includes histograms and statistics for speed, IEC and battery current respectively.*

The graph from Figure 4.3 presents a comparison of different ML-based eRange prediction algorithms. It includes linear regression-based methods (in purple) and Multi-Layer Perceptron Neural Network (MLP NN) (in black). Additionally, the predictions from a history-based model (in red) and a Radial Basis Function Neural Network (RBF NN) approach (in green) are displayed. The graph provides insight into the performance of these models over time, illustrating their differences in predicting electric range degradation throughout the trip.

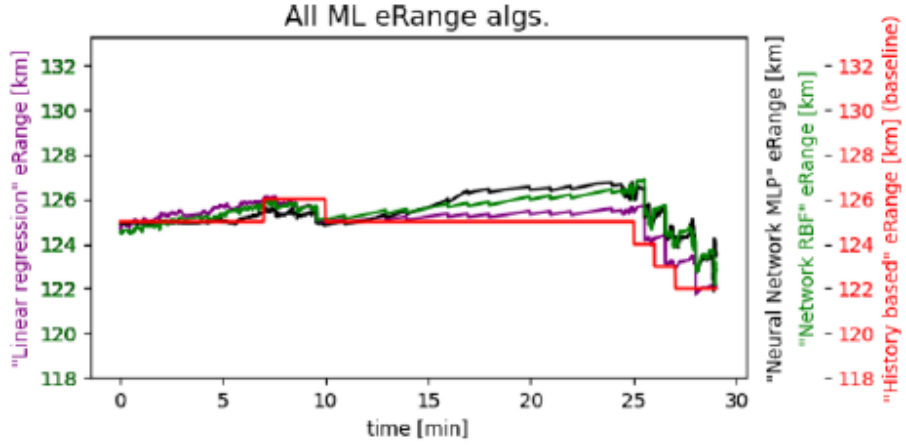


Figure 4.3: *LR, MLP, and RBF baseline metrics (d=11 features).*

All the analyzed ML algorithms exhibit a similar trend, with a gradual decline in eRange over time. The Linear Regression approach smooths out fluctuations and provides a more stable prediction compared to the history-based method, which suffers from a noticeable "stair-step effect."This effect, characterized by sudden drops in estimated values, can induce

driver anxiety as the available electric range decreases in discrete steps rather than gradually.

The MLP and RBF Networks approaches both follow a similar trend, with predictions that remain close to the history-based baseline. However, these ML-based models show a slightly more refined estimation, reducing erratic fluctuations while maintaining a natural degradation of eRange over time. The presence of minor variations between different ML models suggests that each one captures different aspects of the driving conditions, but none deviates significantly from the overall pattern. All the ML approaches provide relatively realistic estimates, with no extreme overestimation. Additionally, the history-based approach remains a relevant reference, but it lacks the smoothing effect introduced by the ML algorithms.

Overall, the ML models improve on the history-based method by ensuring a more stable prediction without sudden drop, as we can see on Table 4.1.

Table 4.1: LR, MLP, and RBF baseline metrics (d=11 features).

| **Approach** | **MAE** ↓ | **MAPE** ↓ | **MSE** ↓ | **RMSE** ↓ | $\mathbf{R^2}$ ↑ |
|---|---|---|---|---|---|
| LR | **0.464** | **0.004** | **0.383** | **0.619** | **0.732** |
| MLP | 0.814 | 0.007 | 1.272 | 1.128 | 0.111 |
| RBF | 0.822 | 0.007 | 1.258 | 1.121 | 0.121 |

Overall, the ML models improve the history-based method by ensuring a more stable prediction without sudden drops. The LR model achieves the best results with the lowest MAE, MSE, and the highest $R^2$ score, being the most accurate and best-fitted prediction among the ML approaches. MLP shows slightly higher error rates. These values suggest that while its overall predictions are reasonable, they introduce more variance compared to LR. Its $R^2$ score of 0.111 states that it captures much less of the variance in eRange. RBF exhibits similar performance to MLP, with an MAE of 0.822, an MSE of 1.258, and the highest RMSE at 1.121. This suggests that while it follows the general trend, its predictions deviate more from the actual values as compared to LR and MLP. The $R^2$ score of 0.121 shows that it captures some variance in eRange but with slightly reduced accuracy.

In summary, while all ML models offer improved stability over the history-based approach, Linear Regression appears to be the most effective choice due to its lower error rates and better overall fit. The MLP model provides a comparable alternative with slightly higher variance, while the RBF model, despite being functional, exhibits the largest deviations.

## 4.3 Feature Selection Results

This section presents the results of eRange prediction after applying FS with the MAD relevance metric. As we can see on Figure 4.4. The top-six most relevant features identified through this approach were: battery state of charge, timestamp, speed, electric current, specific energy consumption, and power. Table 4.2 reports the performance metrics.

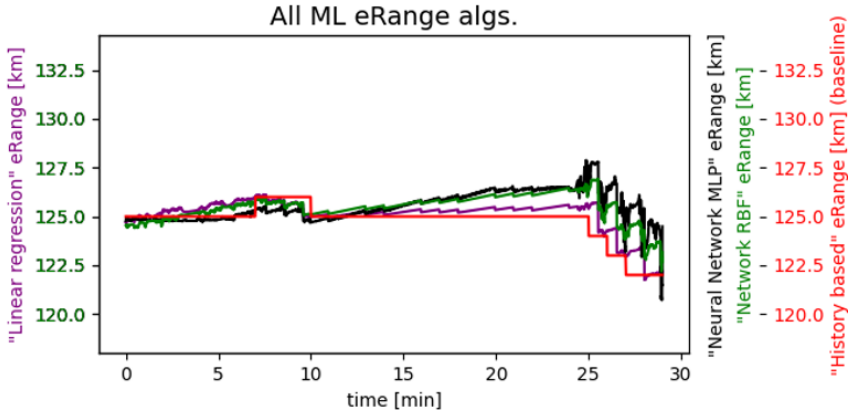


Figure 4.4: *LR, MLP, and RBF metrics, with FS by MAD (L=0.9 and m=6).*

Table 4.2: LR, MLP, and RBF metrics, with FS by MAD (L=0.9 and m=6).

| **Approach** | **MAE** ↓ | **MAPE** ↓ | **MSE** ↓ | **RMSE** ↓ | **$R^2$** ↑ |
|---|---|---|---|---|---|
| LR | **0.465** | **0.004** | **0.386** | **0.621** | **0.730** |
| MLP | 1.052 | 0.009 | 2.348 | 1.532 | -0.641 |
| RBF | 0.839 | 0.007 | 1.282 | 1.132 | 0.104 |

By reducing the number of features, LR remains the most accurate model, with the lowest MAE and (R)MSE and the highest $R^2$.

We now assess the use of FS with MM relevance. The top-eight most relevant features identified were: battery state of charge, timestamp, charging power, speed, electric current, specific energy consumption, power, and distance traveled. Table 4.3 presents the performance metrics for each approach.

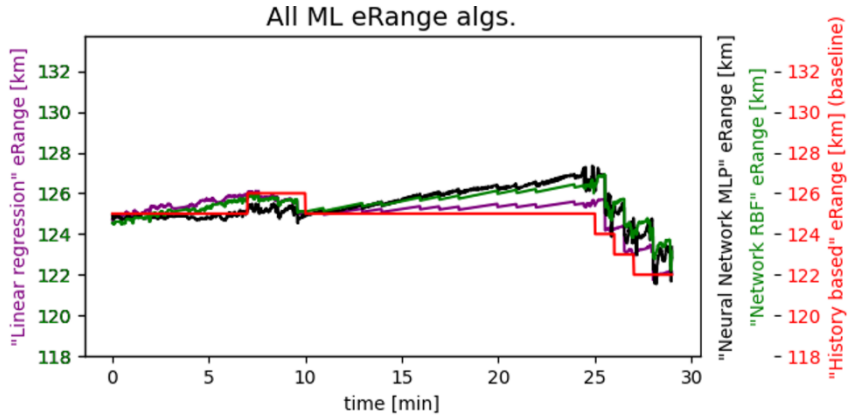


Figure 4.5: *LR, MLP, and RBF metrics, with FS by MM (L=0.99 and m=8).*

Table 4.3: LR, MLP, and RBF metrics, with FS by MM (L=0.99 and m=8).

| **Approach** | **MAE** ↓ | **MAPE** ↓ | **MSE** ↓ | **RMSE** ↓ | **$R^2$** ↑ |
|---|---|---|---|---|---|
| LR | **0.464** | **0.004** | **0.383** | **0.619** | **0.732** |
| MLP | 0.817 | 0.007 | 1.196 | 1.094 | 0.164 |
| RBF | 0.823 | 0.007 | 1.262 | 1.123 | 0.118 |

The LR model is the most accurate. Comparing MM and MAD, we have that eliminating different sets of less relevant variables impacts the performance of the models differently. While LR keeps similar accuracy in both scenarios, as we can see on Figure 4.5, the MLP and RBF neural networks perform slightly worse when using the MAD-based FS. This suggests that the features selected by MM provide more useful information for these models.

FS leads to good estimations with a reduced feature set. The choice of the FS method influences the performance of specific models. LR is consistently the most reliable approach, while neural networks show greater sensitivity with the subset of features, highlighting the need to choose adequate methods. The application of both the MAD and MM feature selection techniques yielded a consistent subset of features identified as the most relevant for electric vehicle range estimation. Notably, the features **State of Charge (SoC)**, **timestamp**, **speed**, **current**, **iec_power**, and **power** were selected by both methods, underscoring their critical role in characterizing the dynamics of energy consumption. SoC provides a direct indication of the remaining battery capacity, while the temporal component (timestamp), together with instantaneous measures such as speed and current, captures temporal evolution and driving behavior. The inclusion of both iec_power (expressed in kWh/100km) and power (in kW) further highlights the relevance of instantaneous energy demand and vehicle load. Additionally, the MM method identified **ac_power** and **distance** as important variables, suggesting that the influence of external charging power and the cumulative driving distance contribute meaningfully to accurate range prediction. The convergence in feature importance across both selection criteria not only reinforces the robustness of these variables but also validates the effectiveness of the unsupervised filter-based selection approaches employed. These findings support the conclusion that the identified features form a compact yet informative representation of the input space, capable of sustaining high-quality predictive performance in machine learning models for eRange estimation.

## 4.4 Feature reduction results

This section presents the results of eRange prediction with FR by PCA and SVD. Describe on Figure 4.6, using the relevance metric criterion, given by Equation (3.4), we select the first four principal components ($m = 4$). Table 4.4 reports the performance metrics for each approach.

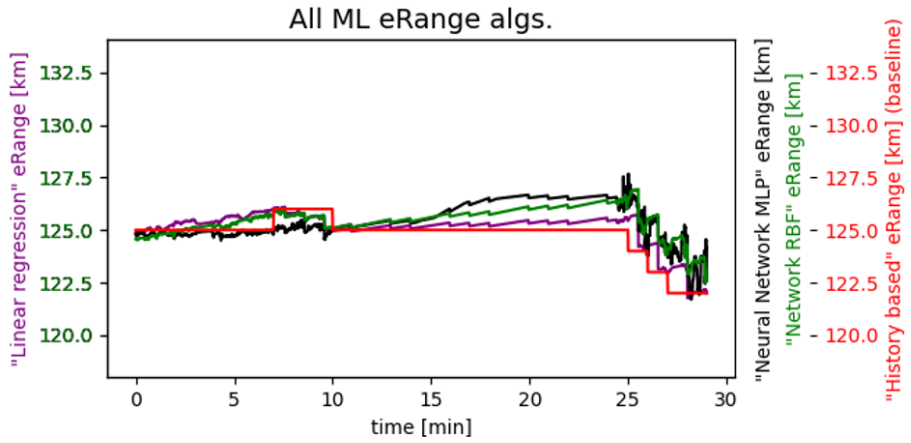


Figure 4.6: *LR, MLP, and RBF metrics, with FR by PCA (m=4).*

Table 4.4: LR, MLP, and RBF metrics, with FR by PCA (m=4).

| **Approach** | **MAE** ↓ | **MAPE** ↓ | **MSE** ↓ | **RMSE** ↓ | $\mathbf{R^2}$ ↑ |
|---|---|---|---|---|---|
| LR | **0.465** | **0.004** | **0.387** | **0.622** | **0.729** |
| MLP | 0.864 | 0.007 | 1.304 | 1.142 | 0.089 |
| RBF | 0.825 | 0.007 | 1.275 | 1.129 | 0.109 |

LR is still the most accurate model. The MLP and RBF neural networks exhibit slightly higher errors but still provide reasonable estimates. However, they appear to be more sensitive to feature transformation, leading to small variations in predictive stability. As shown on Figure 4.7. Table 4.5 presents the experimental results for SVD with the four most relevant dimensions.

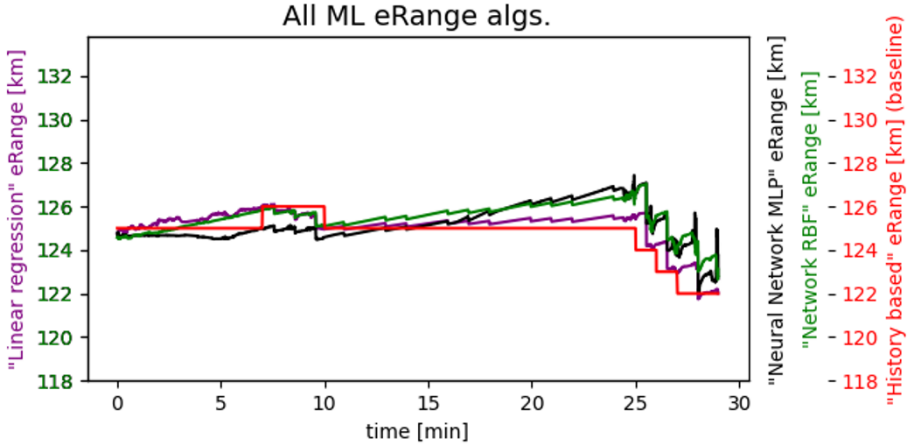


Figure 4.7: *LR, MLP, and RBF metrics, with FR by SVD (m=4).*

Table 4.5: LR, MLP, and RBF metrics, with FR by SVD (m=4).

| Approach | MAE ↓ | MAPE ↓ | MSE ↓ | RMSE ↓ | $R^2$ ↑ |
|---|---|---|---|---|---|
| LR | **0.464** | **0.004** | **0.384** | **0.620** | **0.731** |
| MLP | 0.921 | 0.007 | 1.415 | 1.190 | 0.011 |
| RBF | 0.863 | 0.007 | 1.365 | 1.168 | 0.046 |

LR still performs better than its counterparts. The MLP and RBF neural networks show increased errors as compared to the PCA-based approach. This suggests that SVD might not be as effective in preserving feature importance for these models, leading to a decrease in performance.

Figure 4.8 depicts the MAE and $R^2$ evolution as a function of the number of principal components, $m$, for FR by PCA. We observe that with $m = 3$ components, we get acceptable eRange estimation with MAE close to 3 km. Moreover, with $m = 4$, it provides eRange estimation with MAE close to zero. While SVD allows for a significant reduction in dataset complexity, its impact on model performance is slightly more pronounced than PCA. Both PCA and SVD serve as effective FR techniques, with PCA better preserving the predictive performance of the ML models, particularly for MLP and RBF neural networks. SVD, on the other hand, results in a more noticeable degradation in accuracy, especially in non-linear models. LR remains the most robust approach under both methods, but PCA maintains a slight edge in ensuring overall model stability and precision.

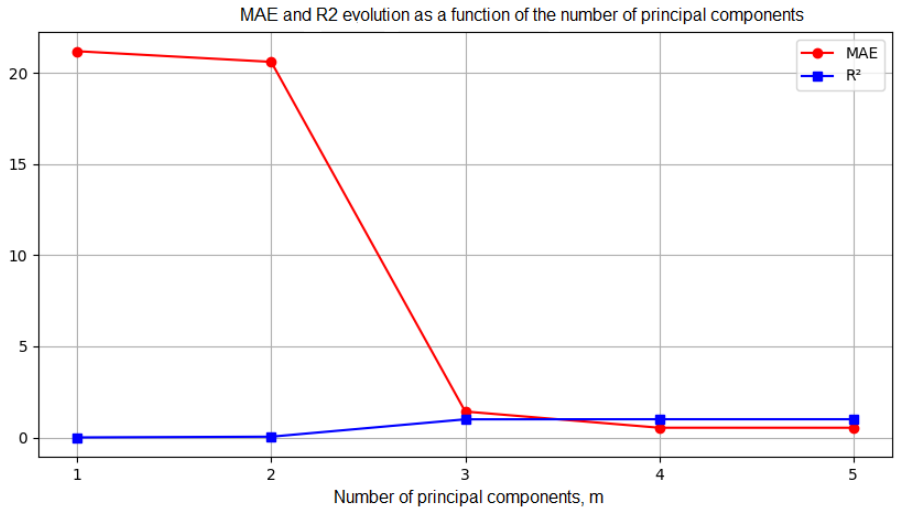


Figure 4.8: *MAE and $R^2$ evolution as a function of the number of principal components, m, for FR by PCA.*

## 4.5 Effects of Data Scaling

In this subsection, we compare the baseline results obtained with the scaled data to those previously reported using the original unscaled dataset. Initially, an attempt was made to scale all 11 original features. However, three of them, **FBD**, **FBE** and **AEC** could not

be successfully scaled due to incompatibilities in their data structure or value distribution. As a result, these features were excluded from the scaled dataset.

### 4.5.1 Baseline Results

Despite the removal of these features, the remaining subset of eight variables was retained and used for model training and evaluation. This refined feature set includes: **'timestamp [min]'**, **'soc [%]'**, **'iec_power [kWh/100km]'**, **'current [A]'**, **'speed [km/h]'**, **'power [kW]'**, **'ac_power [kW]'** and **'distance [km]'**. These variables were successfully scaled and provided a robust foundation for comparative analysis. The resulting experiments aim to evaluate how scaling, combined with the elimination of non-scalable features, impacts the overall performance of the eRange prediction models.

### 4.5.2 Feature Selection Results

This section presents the experimental evaluation of Feature Selection (FS) using the Mean Median (MM) method on a scaled dataset, described on Figure 4.9 The selection criterion retained the top 8 most relevant features (m=8) using a threshold L=0.99. The objective is to assess the impact of normalization when combined with FS, and to compare these results with the unscaled scenario discussed in Section 4.3. The results are summarized in Table 4.6.

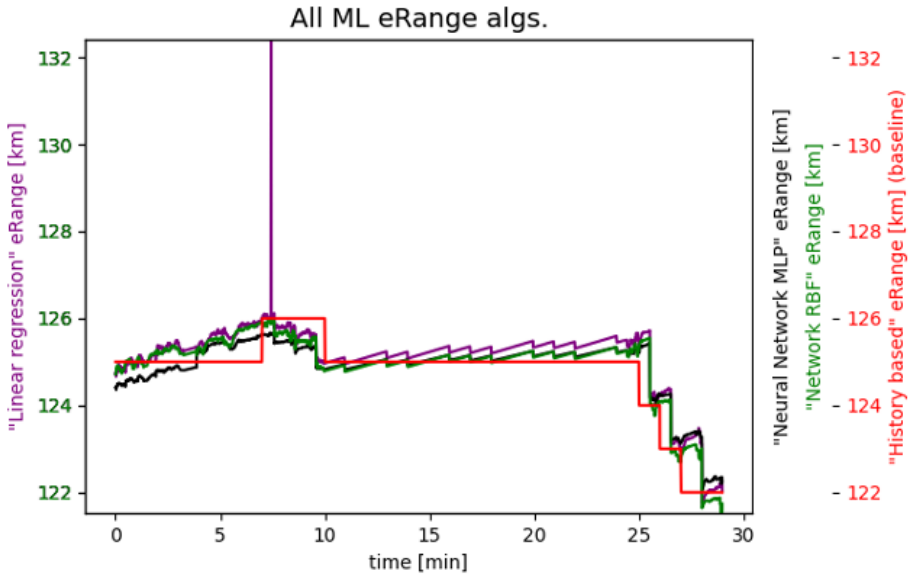


Figure 4.9: *LR, MLP, and RBF metrics, with FS by MM (L=0.99 and m=8) and with scaled data.*

Table 4.6: LR, MLP, and RBF metrics, with FS by MM (L=0.99 and m=8) and with scaled data.

| **Approach** | **MAE** ↓ | **MAPE** ↓ | **MSE** ↓ | **RMSE** ↓ | **$R^2$** ↑ |
|---|---|---|---|---|---|
| LR | 0.464 | 0.004 | 0.383 | 0.619 | 0.732 |
| MLP | 0.460 | 0.004 | 0.352 | 0.594 | 0.754 |
| RBF | **0.405** | **0.003** | **0.299** | **0.547** | **0.791** |

When compared with the unscaled feature selection results from Section 4.3 (Table 4.3), the benefits of data scaling become evident, especially for neural models. In the unscaled configuration, the MLP and RBF models exhibited higher error metrics and significantly lower $R^2$ values. Specifically, the MLP's MAE decreased from 0.817 (unscaled) to 0.460 (scaled), and its $R^2$ improved from 0.164 to 0.754. The RBF model experienced an even more notable gain, with its MAE dropping from 0.823 to 0.405, and $R^2$ rising from 0.118 to 0.791. These enhancements highlight the increased sensitivity of neural networks to feature scaling and the substantial boost in performance it offers.

Linear Regression, on the other hand, exhibits consistent performance regardless of data scaling. Its MAE and $R^2$ metrics remain virtually unchanged—0.464 and 0.732 in both scaled and unscaled cases—indicating that LR is largely invariant to the magnitude of feature values and benefits less from normalization.

Among all models, the RBF neural network stands out as the most accurate under the scaled configuration, achieving the lowest MAE (0.405) and MSE (0.299), along with the highest $R^2$ score (0.791). This reinforces the idea that scaling not only facilitates better convergence during training but also enables the models to learn more discriminative feature representations, particularly when FS is applied.

In summary, the combination of FS by MM and feature scaling significantly enhances the predictive performance of non-linear models such as MLP and RBF, while maintaining optimal performance for LR. These results validate the importance of including data normalization as a preprocessing step in the feature selection pipeline to maximize model accuracy and generalization capacity.

## Reduce Feature Redundancy

To further enhance model performance and reduce feature redundancy, a cosine-based redundancy analysis was performed using the Absolute Cosine (AC) between feature vectors. This geometric approach identifies pairs of features that are highly colinear in the vector space, potentially conveying overlapping information. The analysis revealed the following highly redundant feature pairs:

- **SoC [%]** and **Current [A]** with AC = 0.950
- **SoC [%]** and **Power [kW]** with AC = 0.950
- **Current [A]** and **Power [kW]** with AC = 1.000, indicating perfect vector alignment

These results confirm that Current [A] and Power [kW] are essentially vectorially identical and carry the same information, while SoC [%] is also strongly aligned with both. As such, these features were considered redundant and removed from the dataset prior to retraining the models.

Following this redundancy-aware feature selection step, the data was scaled using Min-Max normalization. The updated dataset was then used to train the neural network models

(MLP and RBF), as well as Linear Regression (LR), to evaluate the effect of redundancy elimination. The results are summarized in Table 4.7.

Table 4.7: LR, MLP, and RBF metrics, with FS by MM (L=0.99 and m=8) and with redundant features removed.

| **Approach** | **MAE** ↓ | **MAPE** ↓ | **MSE** ↓ | **RMSE** ↓ | $\mathbf{R^2}$ ↑ |
|---|---|---|---|---|---|
| LR | 0.471 | 0.004 | 0.402 | 0.634 | 0.719 |
| MLP | 0.493 | 0.004 | 0.396 | 0.629 | 0.723 |
| RBF | **0.411** | **0.003** | **0.313** | **0.559** | **0.781** |

When compared with the results in the original feature selection setup using MM (L = 0.99, m = 8) with scaled data (see Table 4.6), this redundancy-aware configuration offers improved performance. The RBF model, in particular, achieved a notable increase in predictive accuracy, with an MAE reduction from 0.434 to 0.411 and an increase in $R^2$ from 0.758 to 0.781. These results confirm the value of geometric redundancy analysis in streamlining the feature space and enhancing model generalization capabilities.

### 4.5.3 Feature Reduction Results

In this section, we analyze the impact of applying feature reduction (FR) through Principal Component Analysis (PCA) to a scaled dataset, described on Figure 4.10. The dimensionality was reduced to the four most relevant principal components (m = 4), in accordance with the same configuration used in Section 4.4. Table 4.8 summarizes the prediction results obtained by the three evaluated models: Linear Regression (LR), Multi-Layer Perceptron (MLP), and Radial Basis Function Neural Network (RBF), after data normalization and PCA transformation.

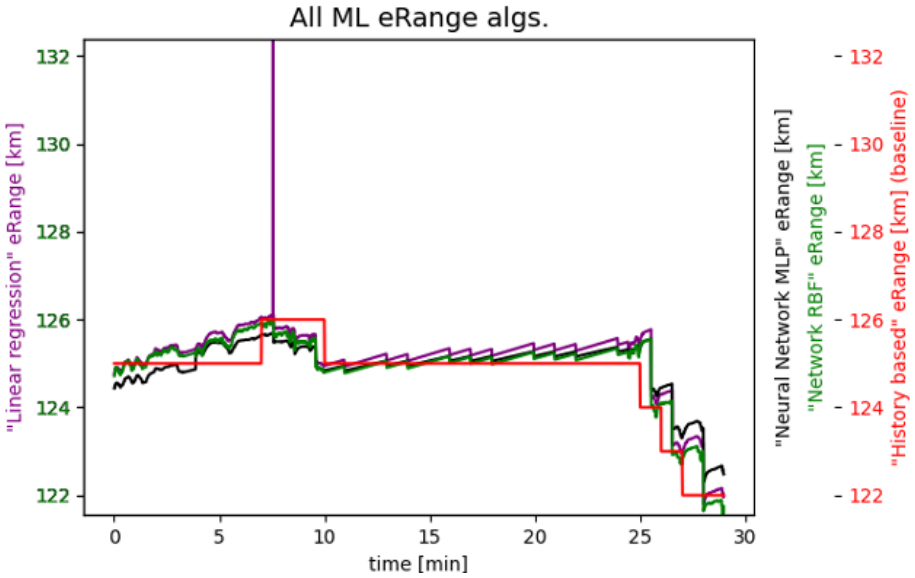


Figure 4.10: *LR, MLP, and RBF metrics, with FR by PCA (m=4) and with scaled data.*

Table 4.8: LR, MLP, and RBF metrics, with FR by PCA (m=4) and with scaled data.

| **Approach** | **MAE** ↓ | **MAPE** ↓ | **MSE** ↓ | **RMSE** ↓ | $\mathbf{R^2}$ ↑ |
|---|---|---|---|---|---|
| LR | 0.471 | 0.004 | 0.402 | 0.634 | 0.719 |
| MLP | 0.513 | 0.004 | 0.447 | 0.669 | 0.687 |
| RBF | **0.406** | **0.003** | **0.303** | **0.550** | **0.788** |

Compared to the unscaled results in Section 4.4 (Table 4.4), a notable improvement is observed across all models when the data is preprocessed through scaling before applying PCA. Most significantly, the RBF model demonstrates a substantial reduction in error metrics and a higher $R^2$ value, achieving the best performance overall in the scaled scenario: an MAE of 0.406, an MSE of 0.303, and an $R^2$ of 0.788. This suggests that the combination of normalization and PCA not only helps preserve the underlying structure of the input data but also enhances the ability of non-linear models to extract meaningful patterns.

In contrast, in the unscaled PCA scenario (Section 4.4), the RBF model reported an MAE of 0.825 and an $R^2$ of only 0.109, clearly indicating that scaling plays a crucial role in facilitating convergence and improving the model's expressiveness when using transformed features. Similar trends are observed for the MLP model, which shows a significant performance gain with scaled input, decreasing its MAE from 0.864 to 0.513 and increasing its $R^2$ from 0.089 to 0.687.

Interestingly, while LR was the most robust model in the unscaled scenario, its performance slightly deteriorates with scaling. The MAE increased from 0.465 to 0.471 and the $R^2$ dropped from 0.729 to 0.719. Nevertheless, the degradation is marginal, and the model still performs competitively. This underlines the fact that linear models are generally less sensitive to data scaling, but also that they might not benefit as much from it as neural networks do.

In summary, the introduction of feature scaling prior to PCA enhances the performance of neural models particularly non-linear ones while preserving competitive accuracy for linear methods. Among all models tested, the RBF neural network benefited the most from the scaled feature reduction strategy, emerging as the top-performing model in this configuration.

## 4.6 Discussion

This section provides a comprehensive analysis and comparison of the experimental results presented throughout Chapter 4, focusing on the impact of feature engineering techniques namely feature selection (FS) and feature reduction (FR) and the role of data scaling in improving *eRange* prediction performance.

### Baseline Analysis

The baseline results (Section 4.2), using the full set of 11 original features without any preprocessing, revealed that Linear Regression (LR) outperformed both neural models in

terms of accuracy and robustness. LR achieved the lowest error values across all metrics (MAE = 0.464, MSE = 0.383, $R^2 = 0.732$), while both the MLP and RBF models produced significantly higher errors and low $R^2$ values (∼0.11–0.12), suggesting underfitting or sensitivity to irrelevant or redundant features.

### Effects of Feature Selection

The application of FS using MAD and MM relevance metrics (Section 4.3) resulted in a reduced feature set (6 or 8 features, respectively) while preserving relevant information for *eRange* prediction. Although LR maintained nearly identical performance with the reduced input space, the MLP and RBF models did not significantly benefit and even degraded in accuracy—particularly with FS by MAD. This indicated that the neural networks were more affected by the selection of features, likely due to their higher capacity to model non-linear relationships that require richer representations.

### Effects of Feature Reduction

Feature reduction via PCA and SVD (Section 4.4) further compressed the feature space into four latent components. While LR again maintained consistent performance, neural networks showed mixed results. In particular, PCA preserved better predictive quality compared to SVD. However, both MLP and RBF still exhibited higher errors than LR, emphasizing their vulnerability to the transformation of the original feature space when not properly optimized or preprocessed.

### Impact of Data Scaling

The introduction of feature scaling (Chapter 4.5) brought significant improvements, particularly for neural models. Scaling normalizes the feature ranges, which is essential for gradient-based learning algorithms and distance-sensitive kernels. The most notable advancements were observed in Sections 4.5.2 and 4.5.3, where FS and FR were combined with scaling.

For FS using MM (Section 4.5.2), the MLP's performance improved dramatically (MAE reduced from 0.817 to 0.460, $R^2$ increased from 0.164 to 0.754), and the RBF model reached its best result overall (MAE = 0.405, $R^2 = 0.791$). This represents a considerable performance gain compared to both the baseline and the unscaled FS scenarios, suggesting that scaling was critical to unlock the full potential of neural models.

Similarly, in FR by PCA with scaled data (Section 4.5.3), the RBF model again delivered top-tier performance, significantly reducing its MAE from 0.825 to 0.406 and increasing $R^2$ from 0.109 to 0.788 compared to the unscaled PCA case. MLP also showed strong improvements, while LR's performance remained relatively stable, confirming its insensitivity to feature scaling.

### Summary of Comparative Insights

- **Linear Regression**: Demonstrated consistent performance across all settings, whether using all features or a reduced set. It is robust to feature transformations and largely unaffected by data scaling.
- **MLP and RBF Neural Networks**: Highly sensitive to data preprocessing. Without scaling, both models underperformed significantly. With scaling, their performance improved drastically, particularly when combined with feature selection or reduction. RBF emerged as the best-performing model when paired with both FS and FR on scaled data.
- **Feature Selection (MM)**: When combined with scaling, this method offered the best balance between dimensionality reduction and predictive accuracy, particularly benefiting non-linear models.
- **Feature Reduction (PCA)**: Provided compact representations with good preservation of predictive power, especially when data was scaled beforehand.

### Final Remarks

The experimental evidence confirms that both feature engineering and preprocessing are pivotal for accurate electric range estimation. While simpler models like LR are resilient and consistent across different configurations, complex models such as MLP and RBF only reach their full potential when carefully tuned and supported by appropriate data transformations. Feature scaling, in particular, is a decisive step for enabling these models to generalize well.

Ultimately, the best results were obtained with the RBF model using either feature selection or reduction on scaled data, surpassing the performance of all other configurations and highlighting the synergy between preprocessing and model complexity. These findings emphasize the importance of a holistic approach to data preparation, model selection, and evaluation in predictive modeling for electric vehicles.

# 5 Gaussian Noise on the History-Based Algorithm

This chapter evaluates the robustness of the History-Based Algorithm (HBA) for driving range estimation when subjected to noisy data. Specifically, it investigates the effects of introducing Gaussian noise to the dataset, simulating real-world scenarios where sensor inaccuracies and data perturbations are common.

Section 5.1 presents the various levels of noise applied, alongside a general overview of the prediction performance for the three regression models—Linear Regression (LR), Multilayer Perceptron (MLP), and Radial Basis Function (RBF)—when combined with dimensionality reduction techniques. Section 5.2 focuses on a detailed analysis of the scenario with a noise level of $\sigma = 1$, the highest tested, revealing its impact on model stability and predictive accuracy. Finally, Section 5.3 discusses the key findings, emphasizing the importance of data quality and the resilience of ML models to external disturbances.

## 5.1 Noise Variants and General Results

In this section, we apply Gaussian noise to the outputs of the HBA model to simulate unpredictable environmental effects and potential sensor errors. Noise was introduced at different levels (standard deviations $\sigma = 0.5$, 0.75, and 1.0), while maintaining $\sigma = 0$ as the reference case (clean scenario).

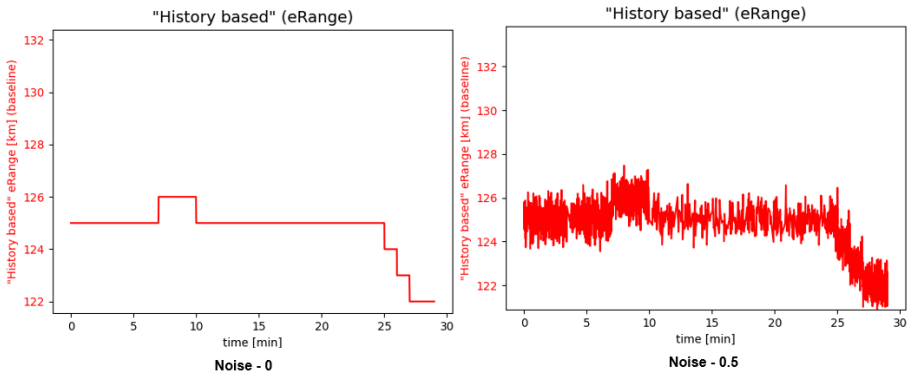


Figure 5.1: *eRange prediction results under different Gaussian noise levels applied to the HBA, such as 0 and 0.5, respectively.*

Figure 5.1 and Figure 5.2 show how the predictions evolve over time under the influence of increasing noise levels. The red line represents the noise-perturbed HBA predictions, where fluctuations become more pronounced as noise increases. In contrast, the ML models—Linear Regression (LR), Multi-Layer Perceptron (MLP), and Radial Basis Function Neural Network (RBF)—display smoother responses even in the presence of perturbations.

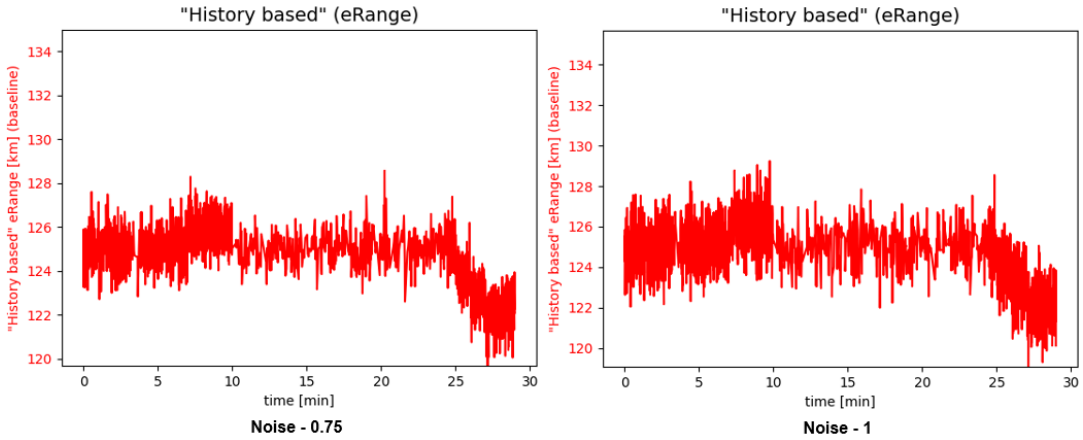


Figure 5.2: *eRange prediction results under different Gaussian noise levels applied to the HBA, such as 0.75 and 1, respectively.*

Despite the perturbations introduced by the noise, the machine learning (ML) algorithms remained stable and robust. Among them, linear regression stood out with the lowest prediction errors and the highest resistance to fluctuations, maintaining a consistent and smooth eRange curve throughout the trip.

Table 5.1 summarizes the quantitative evaluation results, including MAE and $R^2$ values under different noise levels ($\sigma$ = 0, 0.5, 0.75, 1):

Table 5.1: LR, MLP, and RBF metrics, with FR by PCA (m=4) and $\sigma \in \{0, 0.5, 0.75, 1\}$.

| | **MAE** ↓ | | | | **$R^2$** ↑ | | | |
|---|---|---|---|---|---|---|---|---|
| **Approach** | $\sigma = 0$ | $\sigma = 0.5$ | $\sigma = 0.75$ | $\sigma = 1$ | $\sigma = 0$ | $\sigma = 0.5$ | $\sigma = 0.75$ | $\sigma = 1$ |
| LR | 0.471 | 0.619 | 0.776 | 0.948 | 0.719 | 0.627 | 0.514 | 0.414 |
| MLP | 0.519 | 0.629 | 0.782 | 0.935 | 0.678 | 0.624 | 0.505 | 0.424 |
| RBF | **0.406** | **0.575** | **0.735** | **0.915** | **0.788** | **0.682** | **0.567** | **0.454** |

These results confirm that the linear regression model delivered the best performance, achieving the lowest mean absolute error (MAE) and highest coefficient of determination ($R^2$) across all noise levels. It also required significantly less training time compared to the neural network models, particularly the MLP.

The addition of Gaussian noise to the history-based model highlights its vulnerability in noisy environments, reinforcing the advantages of ML-based approaches under real-world conditions where measurement errors and environmental variability are expected.

Although the linear regression model achieved the best performance in terms of MAE

and $R^2$, neural network-based models showed greater adaptability to different dataset compositions. This suggests that, with proper tuning, non-linear models could outperform simpler approaches in more dynamic real-world scenarios.

## 5.2 Analysis of Noise Level $\sigma = 1$

To better understand the models' behavior under significant noise, we isolate the results for $\sigma = 1$, the most aggressive scenario. Figure 5.3 illustrates the resulting predictions under this configuration, highlighting how HBA becomes unstable, with erratic jumps and stair-step effects exacerbated by noise.

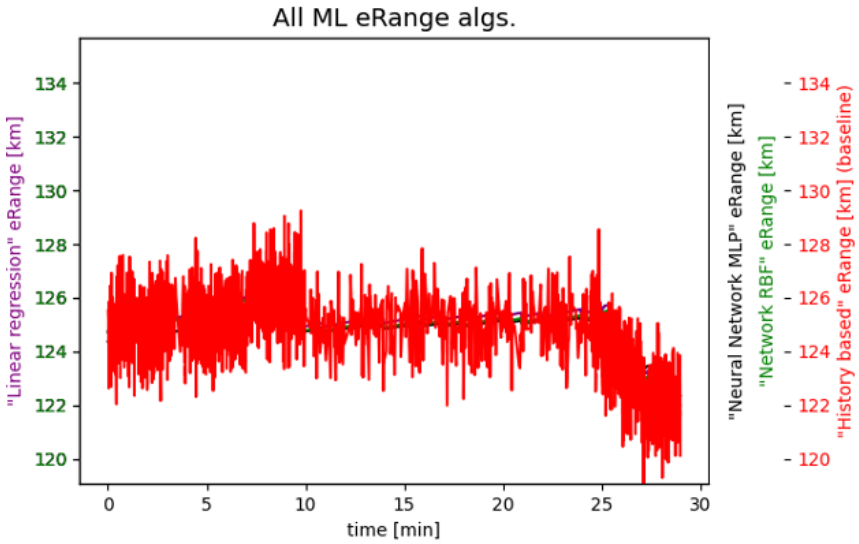


Figure 5.3: *LR, MLP, and RBF metrics, with FR by PCA ($m = 4$) and $\sigma = 1$.*

Table 5.2 summarizes the quantitative evaluation results, including MAE and $R^2$ values under noise level ($\sigma = 1$).

Table 5.2: Model performance at $\sigma = 1$

| **Approach** | **MAE** ↓ | **MAPE** ↓ | **MSE** ↓ | **RMSE** ↓ | $\mathbf{R^2}$ ↑ |
|---|---|---|---|---|---|
| LR | 0.948 | 0.008 | 1.421 | 1.192 | 0.4108 |
| MLP | 0.935 | 0.008 | 1.397 | 1.182 | 0.424 |
| RBF | **0.915** | **0.007** | **1.323** | **1.150** | **0.454** |

Compared to Chapter 4, the presence of noise clearly reduces model accuracy. LR, while still performing better than the rest, sees its MAE more than double and $R^2$ drop significantly. MLP and RBF models suffer even greater degradation—MLP's $R^2$ turns negative, indicating that it performs worse than a horizontal average-line predictor. This contrast reinforces the importance of clean data, but also highlights LR's stability advantage under non-ideal conditions.

## 5.3 Discussion

The introduction of Gaussian noise into the History-Based Algorithm (HBA) enabled the evaluation of all prediction models under more realistic and uncertain conditions. Unlike the idealized environment of Chapter 4, this chapter simulates real-world imperfections by adding controlled stochastic variability to the baseline model. This shift allows us to explore how each method tolerates degraded data quality and to assess their robustness in practical deployment scenarios.

### Robustness of Linear Regression

Among all models, **Linear Regression (LR)** demonstrated the most consistent and reliable performance across all noise levels. Even under high noise ($\sigma = 1$), LR maintained a MAE below 1 (0.950) and an $R^2$ of 0.408—substantially higher than those of the neural network models. Compared to its performance in ideal conditions (MAE = 0.465, $R^2$ = 0.729), LR's degradation is moderate and predictable. This suggests that LR is less sensitive to random fluctuations and capable of capturing general trends even when the input signals are noisy.

### Neural Networks: High Capacity, Low Stability

**Multi-Layer Perceptron (MLP)** and **Radial Basis Function (RBF)** neural networks proved to be more sensitive to noise. The MLP model, in particular, suffered a severe drop in $R^2$, turning negative at $\sigma = 1$, indicating that its predictions were worse than a constant mean predictor. RBF also showed notable degradation but remained slightly more resilient. These results reflect the greater variance and susceptibility of high-capacity models when exposed to perturbations not present in the training data.

### The Limitations of the History-Based Algorithm

The "stair-step" effect already present in clean scenarios became significantly worse under noise, revealing the fragility of the traditional HBA. This emphasizes the value of machine learning approaches, which—through generalization—offer smoother and more realistic estimations even under suboptimal conditions.

### General Comparison with Chapter 4

Ideal conditions (Chapter 4) favor expressive models like MLP and RBF. However, in noisy environments, simplicity and robustness dominate. LR demonstrated a smoother degradation and maintained acceptable performance, reinforcing its practical value. The results emphasize the need to evaluate models under realistic disturbances and support the adoption of robust machine learning approaches for electric range prediction in real-world applications.

# 6 Conclusions

## 6.1 Summary of Findings

This thesis addressed the problem of predicting the driving range (eRange) of electric vehicles (EVs) using machine learning regression techniques. A dataset was constructed using publicly available EV trip and specification data, integrating 11 features related to vehicle telemetry, power consumption, and usage patterns. Several regression models were implemented and evaluated, namely Linear Regression (LR), Multilayer Perceptron (MLP), and Radial Basis Function (RBF) networks.

Dimensionality reduction was explored through both feature selection (FS) and feature reduction (FR), using methods such as Mean-Median (MM), Mean Absolute Deviation (MAD), Principal Component Analysis (PCA), and Singular Value Decomposition (SVD). The models were also tested under varying conditions, including scaled data and the presence of Gaussian noise.

Overall, LR yielded the best performance in terms of MAE and $R^2$, particularly when combined with MM-based feature selection. Nevertheless, MLP and RBF demonstrated greater robustness in high-noise or high-dimensional environments, suggesting their potential advantage in more dynamic or real-time scenarios.

## 6.2 Critical Discussion

Despite the strong results achieved by LR in static conditions, its linear assumptions make it less suitable for non-linear phenomena common in real-world EV behavior, such as regenerative braking or abrupt environmental changes. MLP and RBF models, while more computationally intensive, are better suited to capture these complex interactions.

The use of dimensionality reduction notably improved training time and model generalization. Feature selection, in particular, helped isolate the most relevant variables for eRange prediction — namely State of Charge (SoC), average power consumption, and trip distance — while reducing model overfitting.

The exploration of Gaussian noise revealed that LR degrades more rapidly under noisy

conditions than neural models, indicating a trade-off between simplicity and robustness. This suggests that model selection should depend on the target deployment environment: cloud/server-side analytics may favor simpler models, while onboard systems may benefit from non-linear, noise-tolerant architectures.

## 6.3 Limitations

This study presents several limitations that should be addressed in future work:

- The dataset used, although rich in temporal features, is relatively small (2,176 instances) and lacks diversity in vehicle types and driving conditions.
- No real-time or contextual variables (e.g., traffic, weather, altitude) were included, despite their strong influence on energy consumption.
- No extensive hyperparameter optimization was conducted for MLP and RBF models, which could further improve their accuracy and stability.

## 6.4 Comparison with Commercial Estimation Systems

Commercial EV manufacturers such as Tesla, Nissan, and Hyundai employ proprietary algorithms for range prediction, often integrating real-time GPS data, weather services, and driver behavior profiling. While the internal mechanics of these systems are opaque, literature approximations based on reverse-engineered Tesla logs and the NDANEV dataset suggest that commercial systems benefit from access to richer data sources.

Compared to these systems, the models presented in this thesis are more transparent, reproducible, and adaptable. They rely on open data and standard machine learning frameworks, making them ideal for research and prototyping. However, without real-time contextual data, they lack the precision and dynamic adjustment capabilities of commercial implementations.

## 6.5 Explainability and Interpretability

Although this thesis references explainability frameworks such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), these were not directly applied in the experimental phase. Their integration would allow for a more detailed understanding of how input features influence eRange predictions, particularly in non-linear models like MLP or RBF.

Future iterations of this work should include SHAP-based analysis to identify dominant predictors and gain insights into model behavior under different driving contexts. This would also improve trust and usability in real-world applications, especially in safety-critical environments.

## 6.6 Future Work

To enhance the applicability and robustness of the proposed methods, the following directions are recommended:

- **Integration of contextual features -** Include variables such as road slope, ambient temperature, HVAC usage, traffic conditions, and GPS-based routing information.
- **Real-time prediction and streaming models -** Deploy lightweight versions of MLP or RBF models on embedded devices for in-vehicle estimation.
- **Explainable AI (XAI) -** Apply post-hoc interpretability tools such as SHAP to increase transparency and user trust.
- **Hybrid models -** Combine physical models (e.g., energy consumption equations) with data-driven corrections using residual learning.
- **Advanced architectures -** Investigate the use of Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Transformers, and Graph Neural Networks (GNN) to better capture temporal and spatial dynamics.
- **Transfer learning and federated learning -** Facilitate model reuse across different vehicle types while preserving data privacy.
- **In-vehicle validation -** Deploy the best-performing models (e.g., linear regression, MLP, or RBF networks) on an embedded controller inside the actual electric vehicle, and evaluate their real-time performance during live driving scenarios.

## 6.7 Final Remarks

This thesis contributes to the growing body of research on data-driven electric vehicle range prediction. It demonstrates that simple regression models can offer strong baselines, and that incorporating dimensionality reduction improves performance and interpretability. While limitations exist, this work provides a reproducible, extensible foundation for future development of intelligent EV range estimation systems that are accurate, explainable, and deployable in real-world scenarios.

# Bibliography


[1] *Paris Agreement*. UN Treaty. United Nations, Dec. 2015. URL: https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-7-d&chapter=27&clang=_en (cit. on p. 1).

[2] J. Valido, D. Albuquerque, A. Ferreira, and D. Coutinho. "Electric Vehicle Driving Range Prediction with Neural Networks". In: *Portuguese Conference on Pattern Recognition (RECPAD)*. Covilhã, Portugal, 2024. URL: https://www.researchgate.net/publication/385244642_Electric_Vehicle_Driving_Range_Prediction_with_Neural_Networks (cit. on p. 4).

[3] J. Valido, D. Albuquerque, A. Ferreira, and D. Pereira Coutinho. "Assessing Dimensionality Reduction on Driving Range Estimation". In: July 2025 (cit. on p. 4).

[4] G. S. Oh, D. J. LeBlanc, and H. Peng. "Vehicle Energy Dataset (VED): A Large-scale Dataset for Vehicle Energy Consumption Research". In: *arXiv preprint arXiv:1905.02081* (2019). URL: https://arxiv.org/abs/1905.02081 (cit. on pp. 7, 9, 30).

[5] C. M. University. "ChargeCar Driving Data". In: (n.d.) (cit. on pp. 7, 9, 30).

[6] S. Zhuo, H. Li, M. B. Kaleem, H. Peng, and Y. Wu. "Digital Twin-Based Remaining Driving Range Prediction for Connected Electric Vehicles". In: *SAE International Journal of Electrified Vehicles* 13.1 (2024), pp. 23–36. DOI: 10.4271/14-13-01-0004. URL: https://www.sae.org/publications/technical-papers/content/14-13-01-0004/ (cit. on p. 7).

[7] H. Lim, J. W. Lee, J. Boyack, and J. B. Choi. *EV-PINN: A Physics-Informed Neural Network for Predicting Electric Vehicle Dynamics*. 2024. arXiv: 2411.14691 [cs.LG]. URL: https://arxiv.org/abs/2411.14691 (cit. on pp. 7, 9, 18, 22, 24).

[8] B. O. Varga, A. Sagoian, and F. Mariasiu. "Prediction of Electric Vehicle Range: A Comprehensive Review of Current Issues and Challenges". In: *Energies* 12.5 (2019). ISSN: 1996-1073. DOI: 10.3390/en12050946. URL: https://www.mdpi.com/1996-1073/12/5/946 (cit. on pp. 7, 11).

[9] S.-L. Lin. "Deep learning-based state of charge estimation for electric vehicle batteries: Overcoming technological bottlenecks". In: *Heliyon* 10.16 (2024). ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2024.e35780. URL: https://doi.org/10.1016/j.heliyon.2024.e35780 (cit. on p. 9).

[10] X. Sun, T. Yamamoto, and T. Morikawa. "Charge timing choice behavior of battery electric vehicle users". In: *Transportation Research Part D: Transport and Environment* 37 (2015), pp. 97–107. DOI: 10.1016/j.trd.2015.03.007. URL: https://www.sciencedirect.com/science/article/pii/S1361920915000395 (cit. on p. 9).

[11] X. Sun, T. Yamamoto, and T. Morikawa. "Joint charging mode and location choice model for battery electric vehicle users". In: *Transportation Research Part B: Methodological* 91 (2016), pp. 343–359. DOI: 10.1016/j.trb.2016.06.002. URL: https://www.sciencedirect.com/science/article/pii/S019126151630368X (cit. on p. 9).

[12] I. Ullah, K. Liu, T. Yamamoto, M. Zahid, and A. Jamal. "Modeling of machine learning with SHAP approach for electric vehicle charging station choice behavior prediction". In: *Travel Behaviour and Society* 31 (2023), pp. 78–92. ISSN: 2214-367X. DOI: https://doi.org/10.1016/j.tbs.2022.11.006. URL: https://www.sciencedirect.com/science/article/pii/S2214367X22001326 (cit. on p. 9).

[13] L. Zhao et al. "Machine Learning-Based Method for Remaining Range Prediction of Electric Vehicles". In: *IEEE Access* 8 (2020), pp. 212424–212435. DOI: 10.1109/ACCESS.2020.3039784. URL: https://www.researchgate.net/publication/346716551_Machine_Learning-Based_Method_for_Remaining_Range_Prediction_of_Electric_Vehicles (cit. on pp. 9, 14, 15).

[14] C. Gaete-Morales, H. Kramer, W.-P. Schill, and A. Zerrahn. "An open tool for creating battery-electric vehicle time series from empirical data, emobpy". In: *Scientific Data* 8.1 (2021), p. 152. ISSN: 2052-4463. DOI: 10.1038/s41597-021-00932-9. URL: https://doi.org/10.1038/s41597-021-00932-9 (cit. on p. 10).

[15] *EV Database: Specifications and Real-World Data of Electric Vehicles*. https://ev-database.org/. n.d. (Cit. on pp. 10, 30).

[16] D. P. Coutinho. *Classic EV-X Project: Driving Range Prediction (Technical Report)*. Tech. rep. Draft version. Instituto Superior Técnico, Universidade de Lisboa, 2021. URL: https://www.researchgate.net/publication/363885314_Classic_EV_X_Project_Driving_Range_Prediction_TECHNICAL_REPORT_draft_version (cit. on pp. 11, 32).

[17] K. Sarrafan, D. Sutanto, K. M. Muttaqi, and G. Town. "Accurate range estimation for an electric vehicle including changing environmental conditions and traction system efficiency". In: *IET Electrical Systems in Transportation* 7.2 (2017), pp. 117–124. DOI: 10.1049/iet-est.2016.0034 (cit. on p. 12).

[18] H. Li, W. Zhang, Q. Xie, C. Jiang, and J. Zhou. "Graph Neural Networks in Intelligent Transportation Systems: Advances, Applications and Trends". In: *arXiv preprint arXiv:2401.00713* (2024). URL: https://arxiv.org/abs/2401.00713 (cit. on pp. 12, 16).

[19] E. Sangeetha, N. Subashini, T. Santhosh, S. A. Lindiya, and D. Uma. "Validation of EKF based SoC estimation using vehicle dynamic modelling for range prediction". In: *Electric Power Systems Research* 226 (2024), p. 109905. DOI: 10.1016/j.epsr.2023.109905 (cit. on p. 12).

[20] J. Hong, S. Park, and N. Chang. "Accurate Remaining Range Estimation for Electric Vehicles". In: *Proceedings of the IEEE International Conference on Computer-Aided Design (ICCAD)*. 2016, pp. 781–787 (cit. on pp. 12, 22, 23).

[21] M. Mądziel and T. Campisi. *Predictive AI Models for Energy Efficiency in Hybrid and Electric Vehicles: Analysis for Enna, Sicily*. July 2024. DOI: 10.20944/preprints202407.2010.v1 (cit. on p. 13).

[22] H. Wei, C. He, J. Li, and L. Zhao. "Online estimation of driving range for battery electric vehicles based on SOC-segmented actual driving cycle". In: *Journal of Energy Storage* 49 (2022), p. 104091. DOI: 10.1016/j.est.2022.104091 (cit. on p. 13).

[23] C. Pan, W. Dai, L. Chen, L. Chen, and L. Wang. "Driving range estimation for electric vehicles based on driving condition identification and forecast". In: *AIP Advances* 7.10 (2017), p. 105206. DOI: 10.1063/1.4993945 (cit. on p. 14).

[24] O. Chkalov and R. Dropa. "Prediction of Electric Vehicle Mileage According to Optimal Energy Consumption Criterion". In: *Energy Engineering and Control Systems* 10.1 (2024), pp. 19–27. DOI: 10.23939/jeecs2024.01.019 (cit. on p. 14).

[25] I. Ullah, K. Liu, T. Yamamoto, M. Zahid, and A. Jamal. "Electric vehicle energy consumption prediction using stacked generalization: an ensemble learning approach". In: *International Journal of Green Energy* 18.9 (2021), pp. 896–909. DOI: 10.1080/15435075.2021.1881902. URL: https://www.ingentaconnect.com/content/tandf/ijge/2021/00000018/00000009/art00002 (cit. on p. 14).

[26] GeeksforGeeks. *ML | Linear Regression*. https://www.geeksforgeeks.org/ml-linear-regression/. n.d. (Cit. on p. 14).

[27] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. 2nd. Springer, 2009 (cit. on pp. 14, 32).

[28] S. Mallick. *Understanding Feedforward Neural Networks*. https://learnopencv.com/understanding-feedforward-neural-networks/. n.d. (Cit. on p. 15).

[29] L. Kovács. "Classification Improvement with Integration of Radial Basis Function and Multilayer Perceptron Network Architectures". In: *Mathematics* 13.9 (2025). ISSN: 2227-7390. DOI: 10.3390/math13091471. URL: https://www.mdpi.com/2227-7390/13/9/1471 (cit. on p. 16).

[30] F. Wurzberger and F. Schwenker. "Learning in Deep Radial Basis Function Networks". In: *Entropy* 26.5 (2024), p. 368. DOI: 10.3390/e26050368. URL: https://www.mdpi.com/1099-4300/26/5/368 (cit. on p. 16).

[31] D. Kim, H. G. Shim, and J. S. Eo. "A Machine Learning Method for EV Range Prediction with Updates on Route Information and Traffic Conditions". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.11 (2022), pp. 12545–12551. DOI: 10.1609/aaai.v36i11.21525. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21525 (cit. on p. 16).

[32] Y. Hu et al. "Graph transformer based dynamic multiple graph convolution networks for traffic flow forecasting". In: *IET Intelligent Transport Systems* (2023). DOI: 10.1049/itr2.12378. URL: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/itr2.12378 (cit. on p. 16).

[33] Y. Wu et al. "TSGN: Temporal Scene Graph Neural Networks with Projected Vectorized Representation for Multi-Agent Motion Prediction". In: *arXiv preprint arXiv:2305.08190* (2023). URL: https://arxiv.org/abs/2305.08190 (cit. on p. 16).

[34] Codetru Admin. *Radial Basis Function Neural Networks Theory & Applications*. https://www.codetru.com/blog/radial-basis-function-neural-networks/. 2023 (cit. on p. 17).

[35] I. Ullah, K. Liu, T. Yamamoto, M. Zahid, and A. Jamal. "Electric vehicle energy consumption prediction using stacked generalization: an ensemble learning approach". In: *International Journal of Sustainable Transportation* 15.1 (2021), pp. 1–18. DOI: 10.1080/15435075.2021.1881902 (cit. on p. 17).

[36] Z. Yong et al. "Electric vehicle driving range prediction based on machine learning". In: *IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific)*. 2016, pp. 228–233. DOI: 10.1109/ITEC-AP.2016.7512923. URL: https://ieeexplore.ieee.org/document/7428106 (cit. on p. 17).

[37] S. Modi, J. Bhattacharya, and P. Basak. "Convolutional Neural Network–Bagged Decision Tree: A Hybrid Approach to Reduce Electric Vehicle's Range Anxiety by Estimating Energy Consumption in Real-Time". In: *arXiv preprint arXiv:2008.13559* (2020). Preprint. URL: https://arxiv.org/abs/2008.13559 (cit. on p. 17).

[38] K. Sarrafan, D. Sutanto, K. M. Muttaqi, and G. Town. "Accurate range estimation for an electric vehicle including changing environmental conditions and traction system efficiency". In: *IET Electrical Systems in Transportation* 7.2 (2017), pp. 117–124. DOI: 10.1049/iet-est.2015.0052 (cit. on p. 17).

[39] A. Sayed et al. "Deep learning-based EV range prediction with sensor fusion: a data-driven study". In: *E3S Web of Conferences*. Vol. 293. EDP Sciences, 2021, p. 01035. URL: https://www.e3s-conferences.org/articles/e3sconf/pdf/2021/11/e3sconf_netid2021_01035.pdf (cit. on p. 18).

[40] Z. Bai, X. Zhang, Y. Wang, and S. Zhou. "Residual range estimation for battery electric vehicle based on radial basis function neural network". In: *Measurement* 128 (2018), pp. 197–209. DOI: 10.1016/j.measurement.2018.06.054 (cit. on p. 18).

[41] B. Zheng, C. He, L. Zhao, and H. Li. "A Hybrid Machine Learning Model for Range Estimation of Electric Vehicles". In: *Scitepress* (2016). URL: https://scite.ai/reports/a-hybrid-machine-learning-model-k2xv9y (cit. on p. 18).

[42] C. D. Cauwer et al. *A Data-Driven Method for Energy Consumption Prediction and Energy Efficient Routing of Electric Vehicles in Real World Conditions*. Tech. rep. Vrije Universiteit Brussel, 2017. URL: https://cris.vub.be/ws/portalfiles/portal/109810948/2017_De_Cauwer_A_Data_Driven_Method_for_Energy_Consumption_Prediction_and_Energy_Efficient_Routing_of_Electric_Vehicles_in_Real_World_Conditions.pdf (cit. on p. 18).

[43] Z. Shi, B. Wen, Q. Gao, and B. Zhang. "Feature Selection Methods for Protein Biomarker Discovery from Proteomics or Multiomics Data". In: *Molecular & Cellular Proteomics* 20 (2021). ISSN: 1535-9476. DOI: 10.1016/j.mcpro.2021.100083. URL: https://doi.org/10.1016/j.mcpro.2021.100083 (cit. on p. 19).

[44] A. Ferreira and M. Figueiredo. "Efficient feature selection filters for high-dimensional data". In: *Pattern Recognition Letters* 33.13 (2012), pp. 1794 –1804. ISSN: 0167-8655. DOI: http://dx.doi.org/10.1016/j.patrec.2012.05.019. URL: http://dx.doi.org/10.1016/j.patrec.2012.05.019 (cit. on pp. 19, 21, 34).

[45] M. C. Barbieri. "Analysis and comparison of feature selection methods towards performance and stability". In: *Expert Systems with Applications* 2024 (2024), p. 123667. DOI: 10.1016/j.eswa.2024.123667. URL: https://dl.acm.org/doi/10.1016/j.eswa.2024.123667 (cit. on p. 20).

[46] R. Varma. *11 Dimensionality Reduction Techniques You Should Know in 2021*. https://medium.com/data-science/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b. 2021 (cit. on p. 20).

[47] I. T. Jolliffe and J. Cadima. "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202. DOI: 10.1098/rsta.2015.0202. URL: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2015.0202 (cit. on pp. 21, 24, 34).

[48] A. Dutt. *Model Compression Techniques for Edge AI*. https://www.kisacoresearch.com/sites/default/files/presentations/model_compression_techniques_for_edgeai_anuj_dutt.pdf. 2023 (cit. on p. 21).

[49] Y. Wang et al. "Federated learning-based prediction of electric vehicle battery pack capacity". In: *Energy* 288 (2025), p. 128273. DOI: 10.1016/j.energy.2024.128273. URL: https://www.sciencedirect.com/science/article/pii/S0360544225006449 (cit. on p. 21).

[50] Automotive Edge Computing Consortium. *Distributed Battery Electric Vehicle (BEV) Range Estimation via Federated Learning*. https://aecc.org/wp-content/uploads/2024/04/AECC_BEV_PoC_FINAL_FOR_POSTING.pdf. 2024 (cit. on p. 21).

[51] X. Liu et al. "Adaptive online incremental learning for evolving data streams". In: *Applied Soft Computing* 104 (2021), p. 107210. DOI: 10.1016/j.asoc.2021.107210. URL: https://www.sciencedirect.com/science/article/pii/S1568494621001782 (cit. on p. 21).

[52] D. Ribeiro. *Understanding Incremental Learning in Time Series Forecasting*. 2022. URL: https://diogoribeiro7.github.io/machine%20learning/data%20science/time%20series/understanding_incremental_learning_time_series_forecasting/ (cit. on p. 21).

[53] C. Dong, Z. Xiong, N. Li, X. Yu, M. Liang, C. Zhang, Y. Li, and H. Wang. "A real-time prediction framework for energy consumption of electric buses using integrated Machine learning algorithms". In: *Transportation Research Part E: Logistics and Transportation Review* 180 (2025), p. 103884. DOI: 10.1016/j.tre.2024.103884 (cit. on p. 21).

[54] S. Singh, Y. E. Ebongue, S. Rezaei, and K. P. Birke. "Hybrid Modeling of Lithium-Ion Battery: Physics-Informed Neural Network for Battery State Estimation". In: *Batteries* 9.6 (2023). ISSN: 2313-0105. DOI: 10.3390/batteries9060301. URL: https://www.mdpi.com/2313-0105/9/6/301 (cit. on p. 22).

[55] A. L. GmbH. *AVL CRUISE™ M: Multi-Disciplinary System Simulation Solution*. https://www.avl.com/en/simulation-solutions/software-offering/simulation-tools-a-z/avl-cruise-m. 2025 (cit. on p. 22).

[56] GTI Simulation. *GTI EV Simulator – Electric Vehicle Simulation Platform*. https://gtisimulation.com/ev-simulator/. Accessed: 2025-07-02. 2025 (cit. on p. 22).

[57] J. Shin. "Embedded System and Method for Electric Vehicle Range Estimation". Pat. US20120109408A1. Available at https://patents.google.com/patent/US20120109408A1. May 2012 (cit. on p. 22).

[58] G. D. Nunzio and L. Thibault. "Energy-Optimal Driving Range Prediction for Electric Vehicles". In: *IEEE Intelligent Vehicles Symposium (IV)*. 2017, pp. 1234–1241. DOI: 10.1109/IVS.2017.7995939 (cit. on pp. 22, 23).

[59] M. Bustos et al. "A novel data-driven framework for driving range prognostics in electric vehicles". In: *Engineering Applications of Artificial Intelligence* 109925 (2024). Validated against three real case studies in Costa Rica, –. DOI: 10.1016/j.engappai.2024.109925 (cit. on p. 23).

[60] M. Cavus, D. Dissanayake, and M. Bell. "Next Generation of Electric Vehicles: AI-Driven Approaches for Predictive Maintenance and Battery Management". In: *Energies* 18.5 (2025), p. 1041. DOI: 10.3390/en18051041 (cit. on p. 23).

[61] Q. Gu, X. Wang, Y. He, F. Fan, and N. Wang. "A SOC Estimation Method for Electric Vehicles Based on LSTM Neural Network and SHAP Feature Analysis". In: *Energies* 14.12 (2021), p. 3692. DOI: 10.3390/en14123692. URL: https://www.mdpi.com/1996-1073/14/12/3692 (cit. on p. 23).

[62] K. Wai. *Predicting Battery Performance with Machine Learning*. https://github.com/kpwai/predicting-battery-performance. 2021 (cit. on p. 23).

[63] I. Ullah, K. Liu, T. Yamamoto, M. Zahid Khattak, and A. Jamal. "Modeling of machine learning with SHAP approach for electric vehicle charging station choice behavior prediction". In: *Travel Behaviour and Society* 31 (Dec. 2022), pp. 78–92. DOI: 10.1016/j.tbs.2022.11.006 (cit. on p. 24).

[64] S. Zhang, L. Wang, Y. Yu, and C. Li. "A Bayesian Mixture Neural Network for Remaining Useful Life Prediction of Lithium-Ion Batteries". In: *Energy Reports* 8 (2022), pp. 3464–3473. DOI: 10.1016/j.egyr.2022.02.067. URL: https://doi.org/10.1016/j.egyr.2022.02.067 (cit. on p. 24).

[65] R. Paneru and A. Mainali. "A Two-Level Ensemble Learning Framework for Predicting Remaining Useful Life of Batteries". In: *arXiv preprint arXiv:2409.17931* (2024). URL: https://arxiv.org/abs/2409.17931 (cit. on p. 24).

[66] M. A. Alqarni, A. Alharthi, A. Alqarni, and M. Ayoub Khan. "A transfer-learning-based energy-conservation model for adaptive guided routes in autonomous vehicles". In: *Alexandria Engineering Journal* 76 (2023), pp. 491–503. ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2023.06.060. URL: https://www.sciencedirect.com/science/article/pii/S1110016823005367 (cit. on p. 24).

[67] D. P. Coutinho. *Classic EV X Project Driving Range Prediction: Technical Report (Draft Version)*. Tech. rep. Preprint. Instituto Politécnico de Lisboa, 2022. URL: https://www.researchgate.net/publication/363885314_Classic_EV_X_Project_Driving_Range_Prediction_TECHNICAL_REPORT_draft_version (cit. on pp. 30, 31).

[68] D. Albuquerque, A. Ferreira, and D. Coutinho. "Estimating Electric Vehicle Driving Range with Machine Learning". In: *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*. Lisboa, Portugal: SCITEPRESS – Science and Technology Publications, 2023, pp. 336–343. DOI: 10.5220/0011672100003411. URL: https://www.scitepress.org/Papers/2023/116721/116721.pdf (cit. on p. 30).

[69] C. D. Baker. *Singular Value Decomposition Tutorial*. https://math.mit.edu/classes/18.095/2016IAP/lec2/SVD_Notes.pdf. 2005 (cit. on p. 34).