CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Estimation of Key Parameters Influencing Battery Sizing

by

Zoha Akbar

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Engineering
Department of Electrical and Computer Engineering

2025

I dedicate this thesis to my parents, whose unwavering love, support, and encouragement have been the driving force behind my academic journey. Their belief in my abilities and sacrifices have made this achievement possible. This work is a tribute to their unwavering commitment and serves as a reflection of the values they instilled in me.

CERTIFICATE OF APPROVAL

Estimation of Key Parameters Influencing Battery Sizing

by

Zoha Akbar

(MEE 231001)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Athar Waseem	IIU,Islamabad
(b)	Internal Examiner	Dr. Umer Farooq Ahmed	CUST,Islamabad
(c)	Supervisor	Dr. Noor Muhammad Khan	CUST,Islamabad

Dr. Noor Muhammad Khan
Thesis Supervisor
September, 2025

Dr. Noor Muhammad Khan
Head
Dept. of Electrical Engineering
September, 2025

Dr. Imtiaz Ahmad Taj
Dean
Faculty of Engineering
September, 2025

Author's Declaration

I, **Zoha Akbar** hereby state that my MS thesis titled "**Estimation of Key Parameters Influencing Battery Sizing**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Zoha Akbar)

Registration No: MEE231001

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Estimation of Key Parameters Influencing Battery Sizing**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Zoha Akbar)

Registration No: MEE231001

Acknowledgement

I am profoundly grateful to Allah (S.W.A), the Almighty, for His boundless blessings and guidance throughout my research journey. His unwavering support, divine interven tion, and grace have been the driving force behind my accomplishments. I am indebted to my parents, whose love, unwavering belief in my abilities, and continu ous prayers have provided me with the strength and motivation to overcome challenges and achieve success. Their unwavering support has been an invaluable gift. I extend my heartfelt appreciation to my thesis supervisor, Dr. Noor Muhammad Khan, for his invaluable guidance, mentorship, and expertise. His unwavering dedication and insight ful feedback have played a pivotal role in shaping my research and academic growth. Additionally, I express my gratitude to the faculty members and staff of Department of Electrical Engineering, CUST for fostering a nurturing academic environment and providing resources that have enriched my learning experience. Their commitment to excellence and intellectual growth has been truly inspiring.

(Zoha Akbar)

Abstract

In the era of electric mobility and renewable energy integration, accurate assessment of Battery Autonomy, Battery Life, and Battery Throughput is critical for ensuring the reliability, efficiency, and longevity of modern energy storage systems. This thesis proposes a data-driven methodology for optimal battery sizing and performance prediction by focusing on the prediction of the values of important factors that affect battery sizing. Leveraging advanced machine learning techniques, the study enhances forecast accuracy across three pivotal metrics: Battery Autonomy, Battery Life, and Battery Throughput. To optimize model input, three distinct feature selection algorithms Harmony Search (HS), Linear Forward Search (LFS), and Ranker Search (RS) are systematically applied. Predictive modeling was performed using Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel, and model efficacy is rigorously assessed using four comprehensive evaluation metrics: Root Mean Squared Error (RMSE), Spearman's Rank Correlation Coefficient (SROCC), Kendall's Tau (KCC), and Pearson's Linear Correlation Coefficient (LCC). Empirical results indicate that LFS yields superior performance for Battery Autonomy and Battery Life, achieving minimal RMSE and high correlation fidelity, whereas RS demonstrates optimal predictive accuracy for Battery Throughput. Conversely, in multi-output scenarios targeting simultaneous prediction of Battery Autonomy and Battery Throughput, HS delivers the most consistent and balanced performance across all evaluation dimensions. For Battery Autonomy, LFS achieves the lowest RMSE (0.0005) with strong SROCC (0.9851). For Battery Life, LFS again shows the best results with RMSE 105.4754 and SROCC 0.9933. For Battery Throughput, RS performs best, yielding a remarkably low RMSE (0.000262) and the highest SROCC (0.9852). These findings highlight the necessity of aligning feature selection strategies with specific modeling objectives. LFS and RS are preferable for single target prediction, while HS emerges as the optimal approach for integrated, multi-target forecasting. The proposed framework offers significant utility for applications in electric vehicles (EVs), renewable energy systems, and intelligent energy management, advancing the state of the art in battery analytics and decision support.

Contents

Author's Declaration iv

Plagiarism Undertaking v

Acknowledgement vi

Abstract vii

List of Figures xii

List of Tables xiii

Abbreviations xv

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Micro Grids . . . 1
  - 1.2.1 Challenges in Micro Grids . . . 2
  - 1.2.2 Types of Microgrids . . . 3
    - 1.2.2.1 Grid-Connected . . . 3
    - 1.2.2.2 Remote . . . 3
    - 1.2.2.3 Networked . . . 3
- 1.3 Importance of Batteries in Microgrids . . . 3
- 1.4 Feature Selection in Machine Learning . . . 4
- 1.5 Meta-Heuristic Algorithms . . . 5
  - 1.5.1 Filter-Based Algorithms . . . 5
  - 1.5.2 Embedded Methods . . . 6
  - 1.5.3 Wrapper-Based Algorithms . . . 6
- 1.6 Selected Dataset . . . 6
- 1.7 Statistical Analysis . . . 10
- 1.8 Thesis Objective . . . 10
  - 1.8.1 Applications of Research . . . 11
- 1.9 Thesis Outline . . . 11

2 Literature Survey and Problem Formulation 13
- 2.1 Literature Review . . . 13

2.1.1 Battery Sizing Literature Review . . . 14
2.1.2 Models Used for Battery Sizing . . . 16
2.1.3 Metaheuristic Approaches . . . 19
2.2 Gap Analysis . . . 22
2.3 Problem Statement . . . 23
2.4 Thesis Contribution . . . 23

3 Proposed Methodology and System Model 24
3.1 Proposed Methodology . . . 24
3.2 Dataset . . . 25
3.3 Pre-processing . . . 26
3.4 Feature Selection . . . 26
3.4.1 Harmony Search (HS) . . . 27
3.4.2 Linear Forward Search . . . 27
3.4.3 Ranker Search . . . 28
3.5 Support Vector Regression . . . 29
3.6 Evaluation Parameters . . . 30
3.6.1 Spearman Ranked Correlation Coefficient (SROCC) . . . 30
3.6.2 Kendal Correlation Constant(KCC) . . . 31
3.6.3 Linear Correlation Constant(LCC) . . . 32
3.6.4 Root Mean Square Error(RMSE) . . . 33
3.7 Statistical Analysis . . . 34

4 Results and Discussion 35
4.1 Data Preprocessing . . . 35
4.2 Results . . . 36
4.2.1 Performance Analysis Using Harmony Search For Battery Autonomy With Cleaned and Uncleaned Data by Varying Epsilon . . . 36
4.2.2 Performance Analysis Using Linear Forward Search for Battery Autonomy With Cleaned and Uncleaned Data by Varying Epsilon . . . 37
4.2.3 Performance Analysis Using Ranker Search for Battery Autonomy With Uncleaned Data by Varying Epsilon . . . 38
4.2.4 Performance Analysis Using Harmony Search for Battery Throughput With Uncleaned Data by by Varying Epsilon . . . 39
4.2.5 Performance Analysis Using Linear Forward Search for Battery Throughput with Uncleaned and Cleaned Data by Varying Epsilon . . . 40
4.2.6 Performance Analysis Using RS for BT with Uncleaned and Cleaned Data by Varying Epsilon . . . 42
4.2.7 Performance Analysis Using Harmony Search for Battery Life with Uncleaned and Cleaned Data by Varying Epsilon . . . 43
4.2.8 Performance Analysis Using LFS for Battery Life with Uncleaned and Cleaned Data by Varying Epsilon . . . 44

4.2.9 Performance Analysis Using RS for Battery Life with Uncleaned and Cleaned Data by Varying Epsilon 45
4.2.10 Effect of Epsilon on Evaluation Parameters 46
4.2.11 Comparative Performance of Feature Selection Methods (HS, LFS, and RS) for BA, BT and BL 46
4.3 Statistical Analysis of Battery Metrics 47
4.3.1 Boxplot of Spearman Rank Order Correlation Coefficient (SROCC) by Ranker Serach 47
4.3.2 Boxplot of Spearman Rank Order Correlation Coefficient (SROCC) by LFS Serach 48
4.3.3 Boxplot of Spearman Rank Order Correlation Coefficient (SROCC) by HS 49
4.3.4 Outliers for Battery Autonomy 50
4.4 Statistical Analysis of Battery Throughput 51
4.4.1 Boxplot of SROCC by HS 51
4.4.2 Boxplot of SROCC by LFS 52
4.4.3 Boxplot of SROCC by RS 53
4.4.4 Outliers for Battery Throughput 54
4.5 Statistical Analysis of Battery Life 54
4.5.1 Boxplot of SROCC HS 54
4.5.2 Boxplot of SROCC by LFS 55
4.5.3 Boxplot of Spearman Rank Order Correlation Coefficient (SROCC) by RS 56
4.5.4 Outliers for Battery Life 57
4.6 Impact of Feature Selection and ML on BT and BA 57
4.6.1 Best Performing Algorithm: Harmony Search (HS) 58
4.7 Features Selected 59
4.7.1 Features Selected By HS 59
4.7.2 Features Selected By LFS 60
4.7.3 Features Selected By Ranker Search 60
4.8 Statistical Analysis of Data 61
4.8.1 Statistical Analysis With Linear Regression 62
4.8.2 Scatter Plots for Linear Forward Search 62
4.8.3 Outliers Detection 63
4.8.4 Scatter Plots for Ranker Search 64
4.8.5 Outliers Detection 65
4.8.6 Scatter Plots for Harmony Search 66
4.8.7 Outliers Detection 67
4.9 Statistical Analysis With Polynomial Regression 67
4.9.1 Scatter Plot with Linear Forward Search 68
4.9.2 Outliers with Polynomial Regression 68
4.9.3 Scatter Plot with Harmony search 69
4.9.4 Outliers of HS with Polynomial Regression 70
4.9.5 Scatter Plot with RS 71
4.9.6 Outliers of RS with Polynomial Regression 71

5 Conclusion and Future Work **74**
5.1 Future Work 75

Bibliography **77**

List of Figures

1.1 Microgrid System 2
1.2 Process Diagram for Feature Selection Process 5

3.1 Proposed Methodology for Optimum Performance 25
3.2 SVR Illustration [57] 29
3.3 Monotonic Variable Illustration [58] 31
3.4 Linear Correlation Types [59] 32
3.5 Illustration of RMSE [60] 33

4.1 Box Plot for SROCC of Battery Autonomy Using RS 47
4.2 Box Plot for SROCC of Battery Autonomy Using LFS 48
4.3 Box Plot for SROCC of Battery Autonomy Using HS 49
4.4 Outliers for Battery Autonomy 50
4.5 Box Plot for SROCC Battery Throughput Using HS 51
4.6 Box Plot for SROCC Battery Throughput Using LFS 52
4.7 Box Plot for SROCC Battery Throughput Using RS 53
4.8 Outliers for Battery Throughput 54
4.9 Box Plot for SROCC of Battery Life Using HS 55
4.10 Box Plot for SROCC of Battery Life Using LFS 56
4.11 Box Plot for SROCC of Battery Life Using RS 56
4.12 Outliers for Battery Life 57
4.13 LFS Scatter Plot for BA/BT 62
4.14 Showing Outliers for BA/BT using LFS 63
4.15 RS Scatter Plot for BA/ BT 64
4.16 Showing Outliers for BA/BT Using RS 65
4.17 HS Scatter Plot for BA/ BT 66
4.18 Showing Outliers for BA/BT Using HS 67
4.19 Scatter Plot for LFS with Polynomial Regression 68
4.20 Scatter Plot for LFS with Polynomial Regression 68
4.21 Scatter Plot for HS with Polynomial Regression 69
4.22 Scatter Plot for HS with Polynomial Regression 70
4.23 Scatter Plot for RS with Polynomial Regression 71
4.24 Scatter Plot for RS with Polynomial Regression 72

List of Tables

1.1 Power Generation Factors and External Elements 7

4.1 Battery Autonomy With Harmony Search: Performance Analysis with Uncleaned Data 36
4.2 Battery Autonomy With Harmony Search: Performance Analysis with Cleaned Data 37
4.3 Battery Autonomy With Linear Forward Search: Performance Analysis with Uncleaned Data 37
4.4 Battery Autonomy With Linear Forward Search: Performance Analysis with Cleaned Data 37
4.5 Battery Autonomy With Ranker Search: Performance Analysis with Uncleaned Data 38
4.6 Battery Autonomy With Ranker Search: Performance Analysis with Cleaned Data 39
4.7 Battery Throughput With Harmony Search: Performance Analysis with Uncleaned Data 39
4.8 Battery Throughput With Harmony Search: Performance Analysis with Cleaned Data 40
4.9 Battery Throughput With Linear Forward Search: Performance Analysis with Uncleaned Data 41
4.10 Battery Throughput With Linear Forward Search: Performance Analysis with Cleaned Data 41
4.11 Battery Throughput With Ranker Search: Performance Analysis with Uncleaned Data 42
4.12 Battery Throughput With Ranker Search: Performance Analysis with Cleaned Data 42
4.13 Battery Life With Harmony Search: Performance Analysis with Uncleaned Data 43
4.14 Battery Life With Harmony Search: Performance Analysis with Cleaned Data 43
4.15 Battery Life With Linear Forward Search: Performance Analysis with Uncleaned Data 44
4.16 Battery Life With Linear Forward Search: Performance Analysis with Cleaned Data 44
4.17 Battery Life With Ranker Search: Performance Analysis with Uncleaned Data 45
4.18 Battery Life With Ranker Search: Performance Analysis with Cleaned Data 45

4.19 Comparative Performance of Feature Selection Methods for Battery Metrics 46
4.20 Predicted Values for Battery Autonomy: Comparison Before and After Outlier Removal 57
4.21 Predicted Values for Battery Throughput: Comparison Before and After Outlier Removal 58
4.22 Comparison of Feature Selection Methods for Battery Autonomy and Throughput (After Outlier Removal) 58
4.23 Top features selected by Harmony Search for Battery Life, Battery Throughput, and Battery Autonomy. 59
4.24 Top features selected by Linear Forward Search for Battery Life, Battery Throughput, and Battery Autonomy. 60
4.25 Top features selected by Ranker Search (F-score) for Battery Life, Battery Throughput, and Battery Autonomy. 61

Abbreviations

COE	Cost of Energy
DERs	Distribution Energy Resources
dGen	Distributed Generation
Ev	Electric Vehicle
FS	Feature Selection
HS	Harmony search
KCC	Kendal Correlation Constant
KW	Kilo Watt
LCC	Linear Corelation Constant
LFS	Linear Forward Search
LSTM	Long Short TerM Memory
MHAs	Meta Heuristic Algorithms
NPC	Net Present Cost
PME	Particular Matter Emissions
RMSE	Root Mean Square Error
RS	Ranker Search
SROCC	Spearman Ranked Correlation Constant
SVR	Support Vector Regression

Chapter 1

Introduction

1.1 Background

In today's interconnected world, micro grids and optimal battery sizing are vital for energy resilience and sustainability. Micro grids enable localized power generation, ensuring supply during failures and integrating renewable like solar and wind. Optimized batteries manage supply–demand fluctuations, reduce fossil fuel dependence, and lower carbon emissions. Economically, proper battery sizing minimizes costs while maintaining reliability. Micro grids also support electrification, advanced energy management, and community empowerment by promoting independence and local growth. Together, they drive innovation for a more resilient and sustainable energy future.

1.2 Micro Grids

A microgrid is a localized energy system that delivers power to remote areas and reduces dependence on fossil fuels, which cause pollution. With rising electricity demand, green energy sources like solar and battery storage are increasingly important, making optimal battery sizing a key issue [1]. In Pakistan, energy shortages have persisted for decades.

Since 2013, efforts have expanded fossil, hydro, and solar generation, with many households adopting rooftop solar. Microgrids, a new technology in the country, can help reduce shortages and conserve fossil resources. They integrate distributed generators (PV, wind, mini turbines, diesel), batteries for storage, and smart loads [2].

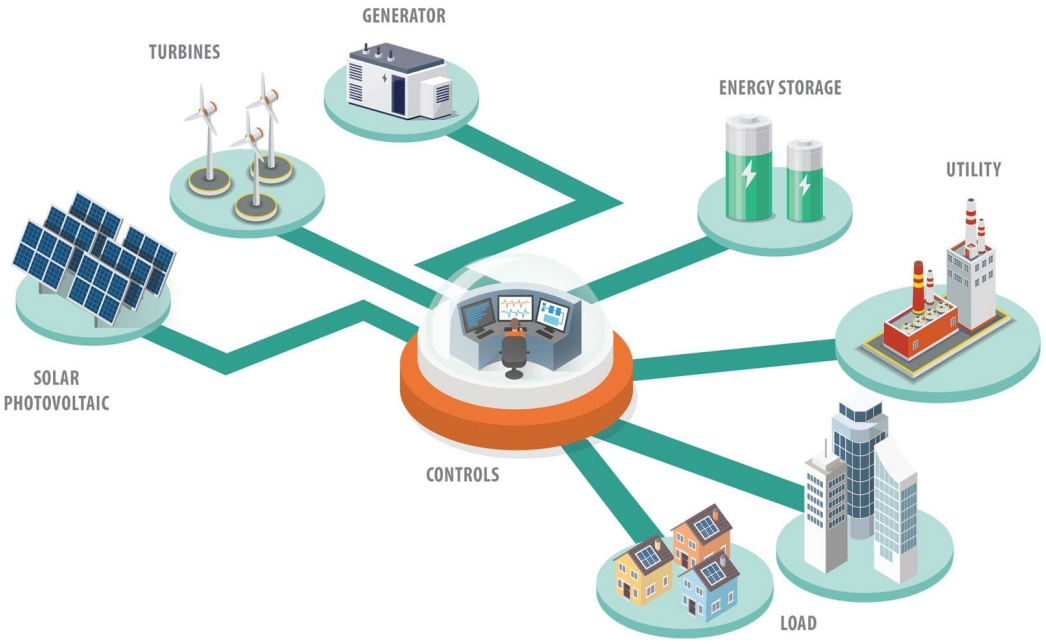

FIGURE 1.1: Microgrid System

1.2.1 Challenges in Micro Grids

Microgrids can disconnect from the main grid during failures, ensuring power for essential loads [3, 4]. Rising electricity demand, driven by population growth and appliance use, has traditionally been met with fossil fuels, causing pollution, resource depletion, and worsening environmental issues [5].

Continued reliance risks irreversible damage. To address this, renewable energy (RE) sources such as solar, wind, and hydropower provide clean, efficient alternatives [6, 7], offering solutions to shortages, price fluctuations, pollution, and climate change. However, integrating multiple renewable sources introduces operational challenges such as intermittency, storage requirements, and balancing supply with demand. Advanced control systems, demand-side management, and robust storage technologies are critical for enhancing stability and reliability. Furthermore,

economic feasibility and policy support play a significant role in determining the pace of microgrid adoption worldwide.

1.2.2 Types of Microgrids

There are three main types of microgrids: grid-connected, remote, and networked.

1.2.2.1 Grid-Connected

These microgrids stay linked to the main grid but can operate independently in island mode. They provide grid support services like demand response, reliability, and voltage regulation, while also lowering energy costs using real-time price monitoring. Common in hospitals, fire stations, and universities.

1.2.2.2 Remote

Also called off-grid microgrids, these operate only in island mode as no utility grid connection is available. They are ideal for remote or hard-to-reach areas, ensuring 100% reliability and independence.

1.2.2.3 Networked

These consist of multiple Distributed Energy Resources (DERs) on the same grid segment, coordinated by a control system. Widely used in community microgrids and smart city projects, they enhance resilience, reduce costs, and secure critical infrastructure like hospitals and police stations [8, 9].

1.3 Importance of Batteries in Microgrids

Battery sizing plays a critical role in the cost and performance of microgrids. The goal is to minimize battery size while meeting voltage, reliability, and frequency

constraints, as batteries are costly and significantly affect system economics [1]. Optimal sizing ensures stable and efficient operation while making storage more accessible to consumers.

Microgrids rely on batteries to supply backup power, improve quality, and balance fluctuating loads. However, challenges such as battery degradation and high costs affect reliability [10–12]. Thus, storage systems must be designed to be both efficient and affordable, supported by real-time, non-invasive measurements and data analysis [13].

Traditional optimization methods like Particle Swarm Optimization and tools such as HOMER have been used for battery sizing [14], but data-driven approaches are emerging as more effective. By using household load data, real-time battery measurements, and regression-based machine learning models, more accurate and cost-efficient estimations can be achieved.

Studies show that complete datasets incorporating power sources, battery parameters, and environmental factors can improve predictions. Techniques like Mixed Integer Linear Programming (MILP) and Support Vector Regression (SVR) are particularly effective for optimizing battery sizing in microgrids [15].

1.4 Feature Selection in Machine Learning

Feature Selection Algorithms (FSAs) are essential in applications such as computer vision, speech recognition, bioinformatics, and machine learning. They enhance model performance by identifying the most informative attributes, reducing dimensionality, and eliminating noise or irrelevant correlations. By focusing on relevant variables, FSAs improve predictive accuracy, reduce overfitting, and lower computational costs, making them particularly valuable for high-dimensional data and real-time or resource-constrained applications [16, 17]. Moreover, FSAs can provide better interpretability of models by highlighting the key features driving predictions. They also support transfer learning by identifying features that generalize well across different datasets or domains. In practice, a combination of filter, wrapper, and embedded methods is often employed to balance efficiency with accuracy.

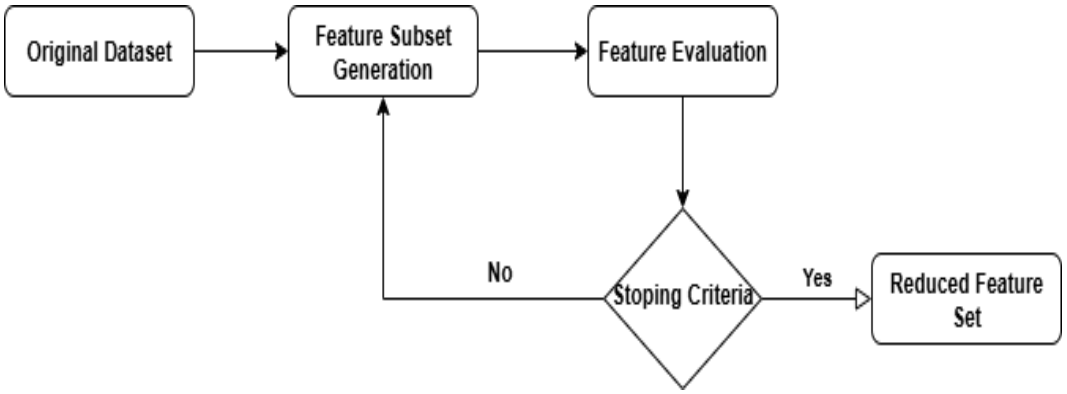

FIGURE 1.2: Process Diagram for Feature Selection Process

1.5 Meta-Heuristic Algorithms

Meta-heuristic algorithms (MHAs') are a form of modern algorithms that have been utilised to tackle a range of optimization problems, including feature selection. These algorithms are very effective for handling computationally intensive problems, as they can identify near optimal solutions by exploring the search space with heuristics and stochastic processes [18] . One of the primary benefits of utilizing MHAs' for feature selection is their ability to handle complex datasets. This is because MHAs' can explore a huge number of possible solutions efficiently, allowing them to discover salient features even when working with high-dimensional data [19]. FSAs for feature selection are mainly classified into three types:

- Filter-based algorithms
- Embedded algorithms
- Wrapper-based algorithms

1.5.1 Filter-Based Algorithms

Filter methods evaluate each feature's relevance independently of the classifier, using statistical criteria for ranking [20]. They are simple and computationally efficient.

1.5.2 Embedded Methods

Embedded methods integrate feature selection into the model training process [21].Examples include LASSO and Random Forests [22, 23].One of the main problems with EM-based FSA is that they necessitate the use of a specific classifier or algorithm, which might result in inefficient feature selection if the classifier or algorithm is not well suited to the dataset. They balance accuracy and efficiency but depend heavily on the chosen classifier and may be biased by outliers [24, 25].

1.5.3 Wrapper-Based Algorithms

Wrapper methods evaluate feature subsets based on classifier performance [26]. Techniques like Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) [27] can improve prediction accuracy by considering feature interactions but require higher computational resources. Hybrid approaches often combine wrapper, filter, and embedded methods for better scalability and performance.

In this research, **wrapper-based feature selection** is employed, as it evaluates feature subsets based on model performance, unlike filter (statistical) or embedded (algorithm-specific) methods. Although computationally intensive, wrappers better capture feature interactions and yield subsets tailored to the predictive task, improving generalization. Techniques such as **Harmony Search, Linear Forward Search, and Ranker Search** are applied to efficiently explore the search space and enhance prediction accuracy while balancing computational effort and robustness.

1.6 Selected Dataset

The input data has 15,000 samples and 40 attributes. Each row in the data set presents yearly consumption of load, battery size, and generation sources. The generation sources are the diesel generator, PV, and grid and important features of the data set are battery autonomy, battery throughput, and battery life. A

dataset representing the domestic load of a microgrid, consisting of 15000 samples with 40 features each, is produced using the MILP approach. This enables the evaluation of optimal battery sizing in microgrids [8].

TABLE 1.1: Power Generation Factors and External Elements

Sr.No.	**Parameter**	**Description**
1.	Photovoltaic power generation	Electrical power produced by photovoltaic cells [8].
2.	Distributed generation market demand (DGEN)	Elements that impact future energy market demand [8].
3.	Hoppecke 6 OPzS 300	A 300 Ah battery with dimensions $147 \times 208 \times 420$ mm and weight ≈ 24.9 kg.
4.	Converter	Standard power rating for the inversion process performed by an operational inverter.
5.	Total Capital Cost (TCC)	Combines overhead and operational expenses to determine the total capital cost.
6.	Unmet Load Fraction	Share of the annual electrical demand unfulfilled due to limited generation.
7.	Total Net Present Cost (TNPC)	Economic indicator used in feasibility analyses of power systems.
8.	Total Emissions	Quantity of pollutants released into the environment.
9.	Total Annual Capital Cost (TACC)	Total operational costs over the project lifetime.
10.	Total Annual Replacement Cost (TARC)	Yearly expense for replacing components in the grid system.
11.	Operations & maintenance cost	Annual operation and maintenance cost.

TABLE 1.1: Power Generation Factors and External Elements (continued)

Sr.No.	**Parameter**	**Description**
12.	Total fuel cost	Annual fossil or gas fuel cost.
13.	Total annual cost (TAC)	Total yearly operating cost.
14.	Operating Cost	Annual expenses related to power generation.
15.	Cost of energy (COE)	Average cost per kWh of useful electrical energy.
16.	Photovoltaic (PV) Production	Projection of PV power production.
17.	Distributed Generation Production (dGENP)	Tool to analyze factors influencing energy resource production.
18.	Grid purchases (GP)	Expense of purchasing grid power from generating companies.
19.	Grid net purchases (GNP)	Grid generation cost excluding operational expenditures.
20.	Total electrical production (TEP)	Maximum power generated by the system convertible to electricity.
21.	AC primary Load Served (AC-PLS)	Total annual energy available to supply AC loads.
22.	Deferrable Load Served (DLS)	Demand that must receive specific energy within a defined time frame.
23.	Renewable Fraction (RF)	Ratio of renewable energy supplied to total energy delivered.
24.	Capacity Shortage (CS)	Total capacity shortfall in annual energy.

TABLE 1.1: Power Generation Factors and External Elements (continued)

Sr.No.	**Parameter**	**Description**
25.	Unmet Load (UL)	Portion of annual demand unserved due to insufficient generation.
26.	Excess Electricity (EE)	Power generated beyond base load requirements.
27.	Diesel	Volumetric consumption of diesel fuel over a given period.
28.	Carbon Dioxide (CO_2) Emissions	CO_2 released into the atmosphere per time period.
29.	Carbon Monoxide (CO) Emissions	CO released into the atmosphere per time period.
30.	UHC Emissions	Unburned hydrocarbon compounds discharged into the air.
31.	Particulate Matter (PM) Emissions	Particulate matter released into the atmosphere per time period.
32.	Sulfur Dioxide (SO_2) Emissions	SO_2 released into the air per time period.
33.	Nitrogen Oxides (NO_x) Emissions	NO_x released into the atmosphere per time period.
34.	DGEN Fuel	Distributed generation fuel consumption in liters/year.
35.	DGEN model Hours	Active hours of distributed generation operation.
36.	DGEN model Starts/yr	Number of starts of the distributed generation system per year.
37.	DGEN Model Life	Operational lifespan of the distributed generation system.
38.	Battery Throughput	Battery lifespan (total capacity divided by usage period).

TABLE 1.1: Power Generation Factors and External Elements (continued)

Sr.No.	Parameter	Description
39.	Battery Life	Estimated operational lifespan of the battery [8].

1.7 Statistical Analysis

Statistical analysis of battery data involves evaluating various performance metrics to assess the quality, efficiency, and lifespan of batteries. The process begins with data cleaning, where outliers and missing values are addressed to ensure accuracy. This research focus on analyzing battery autonomy and battery throughput using statistical methods to understand and optimize battery performance under various operating conditions. Battery autonomy, defined as the duration a battery can sustain a load before requiring a recharge, and battery throughput, representing the total energy output delivered over time, are critical performance indicators. To evaluate these parameters, we collect and preprocess data, ensuring the removal of outliers and handling of missing values for accurate analysis.Time-series plots and scatter diagrams are used to visualize the relationship between actual and predicted values. Correlation analysis is applied to determine dependencies between autonomy, throughput, and Life.

1.8 Thesis Objective

A significant amount of research has been conducted on batteries, considering different battery features, but much of it remains limited in scope. This thesis aims to address this gap by focusing on the optimal battery sizing of microgrids using machine learning techniques applied to 40 different features, while minimizing human error in the process. The main objectives of this research are:

- Predict key factors for battery sizing: life, autonomy, and throughput.

- Apply ML methods for automated battery size estimation.
- Compare ML models to identify the most effective one.
- Implement and evaluate Linear Regression and Polynomial Regression for battery sizing prediction.

1.8.1 Applications of Research

The proposed research methodology has several practical applications. some of the most important practical work are given below:

- This research can significantly aid microgrid planners and engineers in making data-driven decisions for battery storage systems.
- The developed models can be adapted for both residential and commercial microgrid systems.
- These methods can also be extended to optimize battery sizing in electric vehicle (EV) charging stations and renewable energy-based off-grid systems.
- The approach promotes sustainable energy storage planning by improving the reliability and economics of battery systems.

1.9 Thesis Outline

This thesis is organized into five chapters. Chapter 1 introduces microgrids, outlining their architecture, benefits, and challenges especially with renewable energy integration. It also highlights the role of feature selection in managing high-dimensional data and introduces machine learning for optimizing microgrid performance. Chapter 2 reviews existing literature on microgrid modeling, control strategies, feature selection in energy systems, and machine learning applications, identifying key research gaps. Chapter 3 develops mathematical models for microgrid components such as distributed energy resources, storage, and loads and

formulates the optimization problem, integrating feature selection with machine learning. Chapter 4 describes the simulation setup, dataset, and evaluation metrics, followed by result analysis and a summary of key findings. Chapter 5 compiles all references used, adhering to academic citation standards. Finally, the thesis concludes with insights for future research directions and practical implications.

Chapter 2

Literature Survey and Problem Formulation

2.1 Literature Review

Numerous studies have explored determining the optimal battery size in microgrids. Several of these investigations have employed Mixed-Integer Linear Programming (MILP)-based optimization techniques to rigorously handle constraints and achieve cost-effective designs. In addition, many researchers have introduced heuristic and metaheuristic techniques, such as Genetic Algorithms, Particle Swarm Optimization, and Harmony Search, to address the non-convex, multi-objective nature of the battery sizing problem. Certain studies have also adopted a data-driven approach, leveraging machine learning and statistical models to predict battery performance and inform sizing decisions using historical and simulated data.

The detailed literature review of this study is threefold, aiming to provide a comprehensive foundation for the proposed methodology:

1. Battery Sizing
2. Model Used for Battery Sizing

3. Metaheuristic Approaches for Feature Selection

2.1.1 Battery Sizing Literature Review

Kazemtarghi and Mallik formulated the microgrid (MG) design problem as an integer linear programming (ILP) model to minimize the total investment and operational cost of MGs while ensuring optimal hourly dispatch of MG assets. Their study further analyzed the impact of energy storage system (ESS) parameters, such as depth of discharge (DoD), ESS lifetime, and battery C-rate, along with fuel procurement costs, on the optimal MG design in both grid-connected and islanded modes. Simulation results demonstrated that grid-connected operation achieved about 14% annual cost savings compared with islanded operation, highlighting the techno-economic benefits of grid flexibility in MG planning [28] .

El Shamy, Aduama, and Al-Sumaiti presented a chance-constrained optimal sizing approach for an isolated hybrid microgrid composed of PV, battery storage, fuel cell, electrolyzer, and hydrogen tank, with the objective of minimizing the system's life cycle cost. The problem was formulated as a mixed integer linear programming (MILP) model that incorporates uncertainties in PV generation and load demand, requiring the system to achieve at least 80% success across 100 simulated scenarios. Simulation results identified the optimal capacities of PV, battery, and hydrogen storage components, with the total life cycle cost of the system estimated at US$1.221 million over a 25-year period [29] .

Adaptive control based on ML for decentralized storage in microgrids proposes the method of neural network architecture for the prediction of cooperative behavior locally. The machine control and the convolution technique used in machine learning provide a forecast horizon of the charging and discharging rates of the battery. The best training performance of the batteries can be monitored and controlled with the help of a convolutional layer and trial/error method [30]. Cooperative and Greedy optimization approaches are used to analyze and optimize the storage and modeling of batteries. Although the system is feasible, the degradation cost of the applied optimization method is a concern for the final charge constraints.

Improvements in the ML-based control in the proposed study can help in further exploration in the field of battery optimization.

The Machine Learning Technique used for multi-objective predictive energy management strategy proposes three strategies for optimization, logical level, dual prediction, and multi objective optimization helps in maximization of the battery bank's state of charge and overall reduction of carbon dioxide emissions [31]. The accuracy of the Machine Learning Technique helps in the prediction of storage capacity and health of the battery which includes the battery's state-of-charge limits. Battery Charge, Energy Trade, and Carbon Dioxide Emission are the key objects for optimization in this research paper. The proposed algorithm can be envisioned under the scope of IOT Technology for real-time improvements in the model.

Pilati, proposed the method of Heuristic Algorithm and Mixed Integer Linear Program for the optimization of economic constraints in Hybrid Energy Systems (HES) [32]. The MILP adoption allows the optimization and sizing of batteries in micro-grids and smart buildings with the help of Efficient Mixed-Integer Linear Programming. The study highlighted the deficiencies of the Heuristic Algorithm, and the future work of this study proposes further developments to be done in HA for the handling of more parameters and optimization of the system.

A streamlined mixed-integer linear programming model for battery storage sizing and optimization in microgrids and smart buildings proposes mixed-integer nonlinear programming model and Linearized Efficient Mixed Integer Programming which consists of McCormick Relaxation method and Piecewise Linearization [33]. Accurate results with optimized runtime help in predicting the battery capacity. The future work for this research includes stochastic modeling and stochastic analysis for reasonable optimization.

Optimal Battery Sizing Procedure through the coordinated integration of BESS, LED loads, and PV systems based on the MG Frequency Security Criterion proposes optimization of battery sizing algorithm using sensors to operate Microgrid in safe limits by controlling the excessive loads and adjusting the load as per the safety of the batteries. Primary Frequency Control in this method is carried out

by considering the safe limits of the battery and their operational ranges [34]. The contribution of PV in PFC is considered an important part of this research. The permissible energy limits for discharge and charging of the battery are calculated in this research. The frequency control scheme shows the adequate amount of difference in battery size while considering the full capacity of PVs and LED.

2.1.2 Models Used for Battery Sizing

Artificial Neural Network for the designing of a charge-state estimator is a useful technique for optimization and sizing of battery in the microgrid. Network Configurations for finding the ideal weights to be processed in ANN help achieve the Mean Squared Values in the domain of battery sizing [35]. The SOC values of the battery bank can be found easily by the ANN technique, and it is quite effective for the estimations done in battery banks. Training Function, Adapting Learning Function, and Number of layers are observed and applied for better estimation of the results. Levenberg-Marquardt and gradient descent along with Hyperbolic Tangent Sigmoid helped in network training and design of the artificial neural network. This research can be optimized by including neural networks with the EKF. The nonlinearities of battery cell chemistry are also to be considered to deal with battery banks in a much-detailed perspective.

Battery Sizing focuses on managing the overall cost structure within a microgrid. Its primary goal is to reduce battery size while ensuring constraints like voltage, reliability, and frequency are maintained, allowing the microgrid to perform effectively with a smaller battery bank. Jayashree introduced the use of Mixed Integer Programming (mathematical models) along with professional tools such as MATLAB for optimizing Battery Energy Storage Systems (BESS) [36]. The research made use of the Generic Algebraic Modelling System and CPLEX Optimization Studio for BESS applications, highlighting decision-making and detailed system simulations as key components to achieve results and reduce battery size by maintaining the same constraints in a microgrid. Beyond Jayashree's work, there remains considerable potential in this area to apply more advanced tools

for BESS system optimization and to investigate various applications involving battery banks.

The Techno-Economic Method for the optimization of the annual demand forecast and the use of HOMER Pro allowed the researchers to analyze the advantages of renewable systems as compared to conventional grid application [37]. The drawbacks of this research are that the research done by Ramesh et al does not include the future data and is only valid for one year data of the plant at a rural site. The perfect optimization is still to be considered as an important parameter that is not yet catered in detail in the research done by Ramesh et al. The enhancement of the RE hybrid system using the pattern search technique was performed in MATLAB Simulink Design Optimization (SDO), utilizing algorithms such as Latin Hypercube, Genetic Algorithm (GA), and Nelder-Mead. Analysis with HOMER Pro software indicated that applying the Nelder-Mead algorithm reduced the optimal penetration of the diesel generator (DG). However, detailed analysis of energy usage and demand over time was not conducted in the above research and there is still enough room for research to be done in long-term demand predictions and consumption patterns.

The configuration and placement of the BESS in a microgrid play a crucial role in controlling key microgrid parameters. In his research, Jagdesh Kumar proposed using the PSCAD Grid Modelling Software, effectively applying it to estimate configuration constraints for BESS in isolated renewable energy plants [38]. The battery bank's design parameters were also examined through simulations that were made by MATLAB and the results were shared accordingly in the research done by Jagdesh. The limitation of the research done by Jagdesh is that he does not consider the battery aging phenomenon can be incorporated into the design methods of BESS for future research.

Hannan introduced several methods and algorithms for battery energy storage system (BESS) sizing, including a filter-based battery sizing approach, a DFT-based ESS sizing technique, and a multi-stage decision model for optimal sizing.Optimization was further enhanced through the application of the Grey Wolf Optimization Algorithm and swarm-based optimization techniques. In addition,

Hannan explored the use of a Model Predictive Control algorithm to achieve optimal BESS sizing [39]. The work of Hannan *et al.* has provided valuable guidance for researchers, supporting advancements in battery sizing for efficient and cost-effective microgrid operation. These algorithms form a fundamental foundation for future microgrid optimization and serve as essential tools for implementing advanced sizing strategies in modern energy systems.

Bidari [40] proposed the Grey Wolf Optimizer Approach and the development of Grey Wolf Optimizer for the optimization of sizing parameters of the battery bank and regulating the constraints in a microgrid by reducing the battery size [40]. The Optimizer approach along with GWO Algorithm is used as an efficient tool for battery sizing in the research done by El-Bidari. GWO offered strong robustness and served as a meta-heuristic algorithm for addressing frequency deviation issues. In the study conducted by El-Bidari, DIgSILENT PowerFactory software was used as the primary simulation tool.A higher penetration level of variables is proposed to work in the future and pursue with this research to optimize the system even more.

Yang, in his research on battery capacity and fluctuations in renewable systems, proposed the use of Sodium Sulphur (NAS) batteries to optimize system sizing and minimize fluctuation rates in renewable energy integration [41]. He further highlighted that traditional Markov Decision Processes (MDP) are insufficient to address the complexities of BESS, advocating instead for sensitivity-based optimization theory as a more effective solution. An iterative optimization algorithm was developed to properly address BESS optimization, demonstrating significant progress toward enhancing renewable energy stability. Nevertheless, Yang emphasized that further advancements are required, particularly in developing faster and more adaptable optimization iterations to reduce system processing time. This opens pathways for integrating advanced machine learning and hybrid optimization techniques to achieve greater scalability and real-time adaptability in modern energy systems. Future research may also explore integrating reinforcement learning strategies to dynamically optimize storage operations under uncertain renewable generation conditions.

2.1.3 Metaheuristic Approaches

Gao performed deep research on the optimal sizing of batteries and proposed deep learning and algorithmic approaches to solve the matter of battery sizing and to achieve the optimal size [42]. Auto Encoders Gao introduced the Extreme Learning Machine (SDAE-ELM) approach for battery size optimization. The study also explored the application of Single Layer Feed-Forward Neural Networks (SLFNN) and DNN for similar configuration tasks. However, a limitation of deep learning algorithms is their need for extensive training data. This requirement can reduce training efficiency, particularly for models like CNNs and RNNs.The future work for this research is to continue machine learning with a high level of artificial intelligence and macro-scale numerical approaches must be considered for the future.

In his research, Boonluk proposed using Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) techniques to optimize the size of the battery bank. The algorithms processed Fourier Coefficients, with simulations conducted in MATLAB and MATPOWER 7.0 [43].Boonluk also noted that both GA and PSO achieved the same estimated battery lifetime of 8.8 years. PSO demonstrated greater efficiency than GA in optimizing the objective function therefore the future studies must continue by involving PSO instead of GA for more functional optimization. Talent's research employed MILP and GAMS in conjunction with the CPLEX algorithm for battery sizing [44]. However, a limitation of the study is its omission of the batteries' temperature profiles when calculating panel efficiencies.For future work, it is recommended that temperature constraints must be considered in the computational data to carry on the research.Optimal battery sizing was achieved using GA and PSO, with the IEEE 30 Test System serving as the basis for implementing optimal BESS configurations [45]. OpenDNS integrated with the IEEE 30 Test System via COM interface was employed for the optimization process.Future work of this research recommends the time series analysis and control of BESS. The drawback of the research is that it does not include multi-type BESS in the system.

In his research, Gupta proposed a battery sizing method that uses a MATLAB-based algorithm incorporating all user comfort and requirement constraints [46].

The algorithm computes the sizing parameters and generates outputs accordingly, with Loss of Load Probability (LOLP) included as a key consideration in the study. Higher reliability and economic benefits are important constraints in this study's battery sizing approach. The future work for this research is to adopt more dynamic algorithms to increase the computational speed of the system and to further optimize the size of the battery.

In research done by Nouhaila Lazaar, the The Genetic Algorithm (GA) was employed to determine the optimal battery size [47]. The primary goals of Nouhaila's research were to reduce the NPC of the system and to account for the Equivalent Loss Factor (ELF) as an indicator of reliability.The future research that can be pursued in the same domain can be the use of advanced algorithms that can consider more factors while dealing with the sizing of batteries in a microgrid

Shaobo et al. [48], explained that convex programming is a mathematical optimization approach focused on reducing convex functions over convex sets. This method is applied to the co-optimization of battery sizing, energy control, and battery life. Additionally, the concept of battery modeling was discussed in detail However, one limitation was the inaccuracy of the battery model, as it neglected key factors such as state of charge. Peiman Mirhoseini [49] developed a MILP-based framework to analyze and assess the operating and the transaction costs associated with a battery charging station, thereby enhancing Operational reliability. Nevertheless, since the system focuses on deploying a charging station as a microgrid (MG) to supply clean electricity for its own needs, it excludes dispatchable units such as diesel generators and fuel cells.

Sampietro *et al.* [50] examined the optimal battery and supercapacitor sizing in automotive applications to achieve the lowest overall cost.They utilized adaptive programming to identify the most effective use of energy storage devices and fuel cells, contributing to understanding thelink between battery capacity and cost. T. Terzimehic *et al.* [51] focused on battery degradation modeling using Support Vector Regression. Their work demonstrated the application of data-driven techniques for battery performance forecasting, employing data from batteries operating at different temperaturesto verify the machine learning outcomes.

Jiaming Li *et al.* [52], described that the usage of a grid-connected PV battery system solves the optimal sizing problem of PV and battery to maximize economic advantage. The battery model utilized in current research still incorporates unrealistic assumptions like no leakage and full charge capabilities etc. To provide more realistic results, more realistic battery models could be considered in future investigations. Sufyan *et al.* [53], have applied various stochastic approaches, such as PSO and GA, as tools of optimization algorithms. Several technical, economic, and environmental considerations must be considered while engineering a battery storage system, however, these aspects were not considered. Ji Wo *et al.* [54] investigated the use of a Feedforward Neural Network (FFNN) to model the link between the Remaining Useful Life and the charge curve, highlighting its straightforwardness and efficiency. However, their study did not account for assessing the RUL of the battery under varying charge current rates. Carlos Vidal *et al.* [55], FNNs recurrent neural networks (RNNs), support vector machines (SVM), radial basis functions (RBF), and Hamming networks are all used to estimate battery state. To give readers a wider picture view of Machine Learning, comparisons between approaches are shown in terms of data quality, test settings, battery kinds, and declared accuracy. The networks that are being compared must have a similar number of learnable parameters and be trained and tested with the same data. Otherwise, it's difficult to draw broad conclusions about the accuracy of a particular estimating technique. The knowledge of many sorts of machine learning algorithms and how we may apply them to battery sizing are well explained.

Although the research addresses a parameter that has been the center of interest for most of the authors and electrical engineers for quite a long time there were certain limitations in the work of each effort done by different authors at different times. The current research will focus on the optimization and the sizing parameter of the battery to enhance the system's stability and to make it much more compact. The design parameters of the storage system will be compared in contrast with the total number of emissions that were reduced due to the proposed enhancement in the previous model of Distributed Generation System. The overall capacity of the plant should not be reduced by the proposed model and the new model of the storage system should be capable enough to stabilize the system.

The proposed model replaces parallel generators with batteries during emergencies, aiming to reduce overall pollution and associated emissions compared to the conventional approach. The objective is to develop a system that not only supports grid stability but also serves as a reliable backup power source. Key battery design parameters are considered to minimize size while enhancing charge storage efficiency, thereby creating a more compact and space efficient solution. Li Guo *et al.* [56] developed probability-based planning for a single microgrid system to get benefits while also saving the environment by decreasing two objectives: TNPC, carbon, and nitrogen emissions while designing the microgrids. The system included DGen, turbines, and solar panels and photovoltaic power generation, etc. The operational strategy involved coordinating fuel-powered generators and batteries to ensure smooth performance, accounting for multi-unit DGen operational constraints and maintaining reserve capacity to limit diesel generator usage hours. The plan emphasized maximizing the use of renewable energy resources whenever the generators were active. Additionally, it evaluated the variability of wind speed and the clearness index.

2.2 Gap Analysis

Although substantial research has been conducted on battery performance evaluation, most existing studies are constrained by limited feature sets and fail to account for the complex, multidimensional nature of battery behavior. Many conventional approaches rely on manual analysis or simplistic models, which restrict their ability to capture nonlinear patterns such as battery degradation over time. Furthermore, the application of advanced machine learning techniques remains underexplored, particularly in the context of selecting optimal input features and modeling multiple target variables simultaneously. Previous works also seldom perform a rigorous comparison of feature selection algorithms, which is essential for improving model efficiency and accuracy. Additionally, integrated prediction of key metrics Battery Autonomy, Battery Life, and Battery Throughput has not been sufficiently addressed, especially in high stakes applications such as electric vehicles, renewable energy systems, and intelligent energy management. These

drawbacks underscore the need for a comprehensive, automated, and adaptive framework for battery performance prediction and sizing.

2.3 Problem Statement

There is a critical need for a robust and data-driven framework that accurately predicts key battery performance metrics Battery Autonomy, Battery Life, and Battery Throughput using feature selection and advanced machine learning approaches. Existing methods fall short in handling high-dimensional data, multi output modeling, and delivering consistent, accurate predictions essential for battery sizing in microgrids.

2.4 Thesis Contribution

The thesis makes several key contributions to the field of battery performance analysis.These are enlisted below:

- Predicts key targets: **Battery Autonomy**, **Battery Life**, and **Battery Throughput**.
- Improves accuracy by systematically **removing outliers**.
- Performs **statistical analysis** to reveal trends and correlations.
- Uses **machine learning** to model data, detect patterns, and predict outcomes.
- Validates predictions with **RMSE, LCC, SROCC, and KCC**.

Chapter 3

Proposed Methodology and System Model

3.1 Proposed Methodology

This study employs a machine learning (ML)-based methodology to predict and analyze battery performance in terms of Battery Autonomy, Battery Life, and Battery Throughput. The methodology is divided into the following key stages. It begins with data pre-processing, including cleaning and outlier removal to ensure reliability. Next, statistical analysis is performed to uncover correlations and patterns in battery behavior.

- Dataset Collection
- Pre-Processing
- Feature Selection
- Machine Learning Model Training
- Prediction of Target Variables
- Statistical Analysis

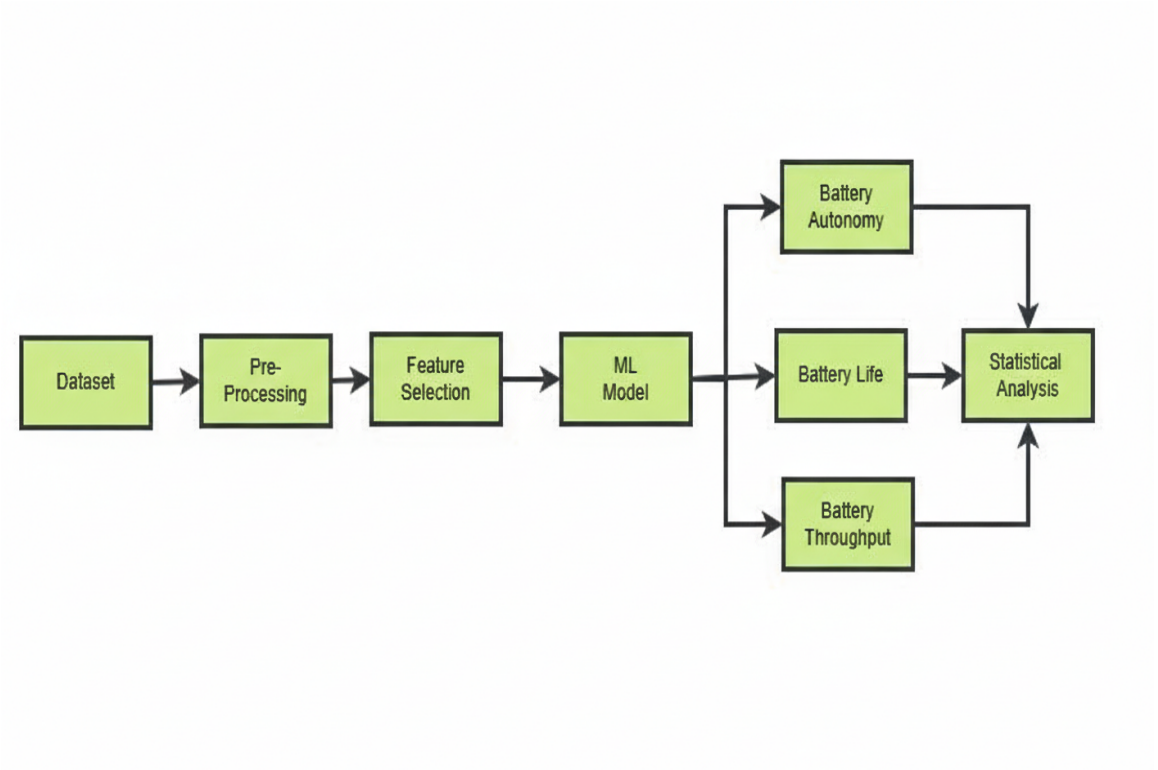

FIGURE 3.1: Proposed Methodology for Optimum Performance

3.2 Dataset

The process begins with the collection of datasets containing key operational parameters of batteries. The considered generation sources include diesel generators, photovoltaic (PV) systems, and the utility grid, while critical features of the dataset are battery autonomy, battery throughput, and battery life. To represent the residential load of a microgrid, a dataset with numerous instances and 40 attributes per instance is generated using the Mixed Integer Linear Programming (MILP) technique. This method enables the evaluation of optimal battery sizing within microgrids [8]. MILP, a widely adopted linear optimization approach, is particularly effective for addressing sizing and selection problems involving Distributed Energy Resources (DERs) and energy storage devices. The optimization framework ensures power balance between DERs, storage systems, and overall load demand. To simulate diverse scenarios, randomized component sizes are entered into matrices to generate multiple generation mixes, each evaluated based on its capacity to meet the total load requirement [8].

3.3 Pre-processing

Data pre processing is a critical initial phase in any machine learning workflow, serving to guarantee the quality and suitability of the data that will ultimately be fed into the models. A fundamental aspect of this stage is data cleaning, which involves a meticulous examination and rectification of imperfections within the dataset. This typically includes the removal of missing values, where various imputation techniques might be employed if outright deletion is not feasible. Furthermore, outliers, which are data points that deviate significantly from the general pattern, are identified and addressed, as they can disproportionately influence model training.Following these pre-processing steps, our refined dataset now comprises a substantial 15000 samples. This cleaned and prepared dataset, consisting of this significant number of instances, is now ready to be utilized for training and evaluating machine learning models. The pre-processing stage has ensured that these samples are of high quality, free from major issues like missing values, outliers, and inconsistencies, thereby setting a strong foundation for building robust and reliable predictive models.

3.4 Feature Selection

The input features undergo a selection process, during which multiple feature selection methods were evaluated.Feature selection methods identify the most important variables for determining the optimal battery size.In this research, wrapper-based feature selection techniques such as Harmony Search, Linear Forward Search, and Ranker Search are employed. Wrapper methods are chosen because they evaluate feature subsets directly on the predictive performance of the model, making them more accurate and better suited for capturing interactions among variables compared to filter or embedded methods. By using multiple wrapper techniques, we ensure that the selected features are robust, reliable, and tailored to the specific prediction task. The final selection of features is made on the basis of their correlation with the target variables and the R^2 score, ensuring that only the most

influential features are retained for accurate prediction of battery autonomy, battery life, and battery throughput.Descriptions of each feature selection method are provided below.

3.4.1 Harmony Search (HS)

Harmony Search (HS) is a metaheuristic optimization technique known for its simplicity, fast convergence, and low computational cost. It has been applied to diverse engineering problems, including nonlinear and non-convex functions with strict constraints, where conventional methods struggle. HS mimics the process of musical improvisation to iteratively search for the optimal solution, making it well-suited for complex optimization tasks.

$$x_{\text{new}} = x + bw \tag{3.2}$$

where x_{new} is the new harmony, $\mathbf{x}_{\text{old}}$ is the old harmony, b is a constant (which controls the magnitude of the change), and ω is a random variable that introduces randomness. The adjustment using a random walk derived from pitch can be illustrated as follows:

$$x_{\text{new}} = x_{\text{old}} + b(2\epsilon - 1) \tag{3.3}$$

where x_{old} is the predefined variable for pitch control b is the constant for the pitch displacement and ϵ is a random number between 0 and 1 [8].

3.4.2 Linear Forward Search

The methodology would employ a sequential approach, which is fundamental for locating a specific element within a collection of items. Upon successful identification of the target item, its corresponding index is typically returned. The traversal during the search process proceeds in a forward direction. Linear search finds its primary application in datasets containing discrete values and a significant number of elements. In the context of n models, the function representing the standard

linear regression model can be formulated as:

$$y = Q\theta + \epsilon \tag{3.4}$$

where Q represents the regression constant (matrix of input features), and θ denotes the regression variable. The error term accounts for the second-order differential equation. A key assumption is that the variance is additive, which enables estimating the parameter θ using the least squares approach [8].

3.4.3 Ranker Search

A ranking search algorithm retrieves information from diverse data sources using evaluation metrics. A well-known example is Google's PageRank, which ranks URLs based on webpage importance and term relevance. Such algorithms generally follow three stages: crawling (using bots to gather updates), indexing (categorizing content by text, images, and tags), and serving (ranking results by query relevance). Similar frameworks, with variations in attributes like price or traffic, are used by other search engines. The basic form of a ranking search algorithm can be expressed as:

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} \tag{3.7}$$

where A, B, C, and D are interconnected web pages, $L(\cdot)$ represents the outbound links, and PR denotes the respective probability functions of the pages. The overall PR function can be defined as:

$$PR(u) = \sum_{\forall v \in B_u} \frac{PR(v)}{L(v)} \tag{3.5}$$

The notation B_u represents the set encompassing all hyperlinks directed towards the Uniform Resource Locator (URL) page u, while $L(v)$ denotes the total count of outbound hyperlinks originating from the URL v. Furthermore, the algorithm incorporates a damping factor to account for the probability of a user randomly

navigating to any page, thereby preventing rank sink issues and ensuring a more realistic representation of web navigation [8].

3.5 Support Vector Regression

Support Vector Regression (SVR) predicts target values by fitting a hyperplane within a defined margin. Kernel functions such as sigmoid, polynomial, and Gaussian RBF map data into higher dimensions to enable linear separation. SVR provides sparse solutions, with the margin and hyperplane orientation (via the normal vector) determining model performance and data fitting.

$$\min_{w} \quad \frac{1}{2}\|w\|^2 \tag{3.6}$$

where w represents the weights, and the error is addressed within the constraints.

$$|y_i - w_i x_i| \leq \epsilon \tag{3.7}$$

where y_i is the initial y constraint for the variable, and x_i is the initial x constraint for the variable [8].

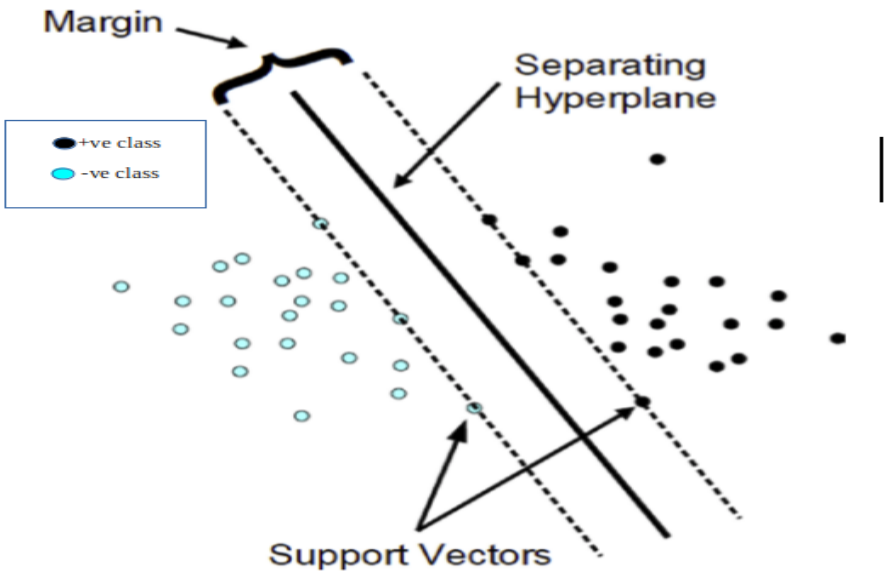

FIGURE 3.2: SVR Illustration [57]

Figure 3.2 illustrates the concept of Support Vector Regression (SVR), where the central black line represents the regression hyperplane predicting continuous values. SVR aims to fit this line within a defined margin of tolerance (shown by the two red lines marking the maximum margin), forming an ϵ-insensitive tube around the hyperplane. Data points lying on or outside these margins are called support vectors (highlighted in blue), and they directly influence the model by defining the position and shape of the regression line. Points within the margin do not affect the model, allowing SVR to balance fitting the data while maintaining model simplicity and avoiding overfitting.This ability to ignore minor deviations makes SVR highly robust to noise in the dataset. Moreover, by tuning kernel functions, SVR can effectively capture both linear and nonlinear relationships. As a result, it is widely used for prediction tasks in energy systems, including battery performance forecasting.

3.6 Evaluation Parameters

The performance of the proposed models is assessed using standard statistical evaluation parameters, including Spearman's Rank Order Correlation Coefficient (SROCC), Kendall's Correlation Coefficient (KCC), Linear Correlation Coefficient (LCC), and Root Mean Square Error (RMSE).

3.6.1 Spearman Ranked Correlation Coefficient (SROCC)

The Spearman Ranked Correlation Coefficient (SROCC) is a non-parametric measure used to assess the strength and direction of a monotonic relationship between two variables. It is particularly useful when data do not meet the assumptions required for parametric correlation tests. SROCC is applicable to both ordinal variables and continuous data measured on ordinal, interval, or ratio scales. Its application relies on three key assumptions First, the variables must be measured on an appropriate scale, Second,the dataset should consist of paired observations, and (iii) the relationship between the variables should be monotonic [8].

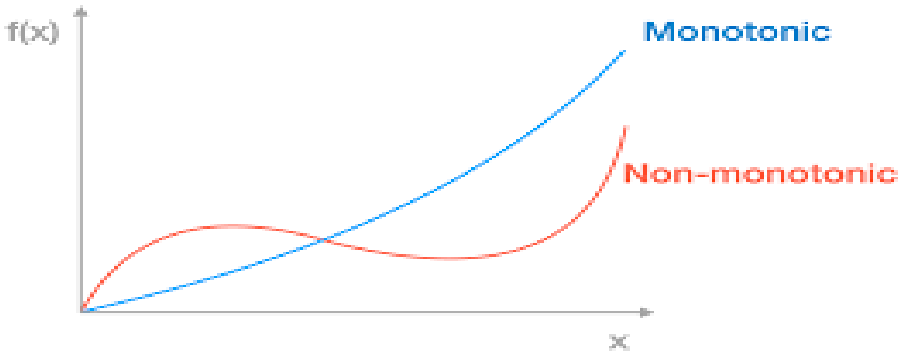

FIGURE 3.3: Monotonic Variable Illustration [58]

The formulation of the SROCC can be seen below

$$\text{SROCC} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{3.8}$$

In this context, let x_i denote the i-th value of the variable x, and let $\bar{x}$ represent the mean of the variable x. Similarly, y_i denotes the i-th value of the variable y, while $\bar{y}$ represents the mean of the variable y.

3.6.2 Kendal Correlation Constant(KCC)

The KCC evaluates the ordinal-level relationship between two measured variables. Kendall's rank correlation serves as a non-parametric alternative to Pearson's correlation, particularly when one or more parametric assumptions are violated. It assesses the similarity in the ranking of data values. A coefficient value of 1 ($\tau = 1$) indicates perfect agreement in ordering between the two sets, while a value of -1 ($\tau = -1$) signifies complete inverse ordering. When $\tau = 0$, it implies there is no association between the rankings of the two sets [14].The rank correlation can be expressed by

$$\tau = \frac{n_c - n_d}{n(n-1)} \tag{3.9}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and n is the total number of pairs. A KCC value close to 1 indicates best performance, suggesting that the predicted battery autonomy from the proposed system

closely matches the actual battery autonomy. Conversely, a KCC value near 0 indicates no relation, showing that the predicted battery autonomy does not align well with the original battery autonomy. [8].

3.6.3 Linear Correlation Constant(LCC)

This metric quantifies the strength of the linear association between two variables, such as x and y. The LCC value, r, indicates the strength of this relationship. When r is close to 1 or -1, it reflects a strong linear association. Conversely, when r is near 0, it indicates a weak linear relationship.

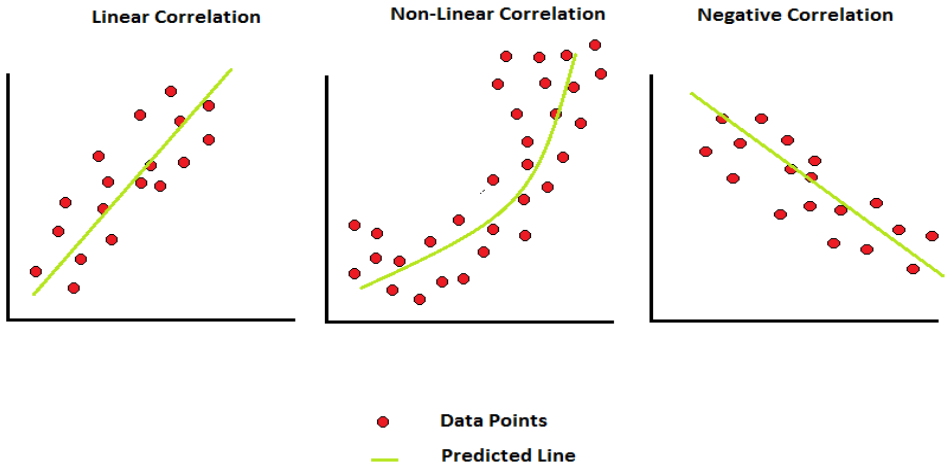

FIGURE 3.4: Linear Correlation Types [59]

The assumption used is that the correlation coefficient would require the underlying relationship between the two variables to be linear. The conclusions are mainly drawn from observable variables in most cases for the tests. In addition, LCC assumes that the variables are measured on an interval or ratio scale and that the dataset contains paired observations with minimal measurement error. It further relies on the assumptions of normality and homoscedasticity, ensuring that variability in one variable remains consistent across the values of the other. Although widely used due to its simplicity and interpretability, LCC is sensitive to outliers and may not accurately represent relationships that are monotonic but non-linear. Therefore, in cases where linearity cannot be guaranteed, alternative correlation measures such as SROCC or Kendall's coefficient are preferred.

The formulation of the linear coefficient can be seen below

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{3.10}$$

where x_i is the i-th value of the variable x, and y_i is the i-th value of the variable y [8].

3.6.4 Root Mean Square Error(RMSE)

The RMSE is the statistical tool that is used for the prediction of standard deviation error (Residual). The residuals are the measure of how far the regression data points are and therefore the measure of the spread. This would mean that the RMSE would sow the concentration of the data around the line of best fit for a statistical finding. The scatter plot below shows the spread of the data around the line of best fit. The concentration along the line of best fit is sparce hence a slightly higher RMSE.

FIGURE 3.5: Illustration of RMSE [60]

RMSE is a crucial metric in statistics for showing data relationships and measuring how much data points deviate from the studied set. The RMSE can be formulated as follows

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \tag{3.11}$$

The value of P represents the predicted values, while i denotes the observed values for the observations in a sample of size n [8].

3.7 Statistical Analysis

After the evaluation of parameters for target variables battery autonomy,battery Life and battery throughput this research presents a statistical analysis of the data, including the use of scatter plots and the computation of correlation coefficients to assess the relationship between the predicted and actual values. These plots serve as a crucial diagnostic tool, highlighting trends, deviations, and potential outliers in the model's predictions.In addition, the study computes various correlation coefficients, such as the Pearson Linear Correlation Coefficient (LCC), Spearman Rank-Order Correlation Coefficient (SROCC), and Kendall's Tau Coefficient (KCC). These metrics quantitatively assess the strength and direction of the linear and monotonic relationships between the predicted and actual values for both battery autonomy and throughput. High values of these coefficients indicate strong agreement between predictions and ground truth, validating the model's effectiveness. This statistical evaluation not only confirms the predictive accuracy of the model but also supports the identification of any systematic errors or inconsistencies in performance. By combining visual and numerical assessments, the analysis ensures a robust interpretation of model behavior and reliability, guiding future improvements in battery performance prediction frameworks.

Chapter 4

Results and Discussion

This section presents the experimental evaluation of the proposed technique, which uses feature selection and a Support Vector Regression (SVR) model to predict battery autonomy and throughput. These parameters are vital for accurate battery sizing in microgrid systems. By selecting relevant features, we improved predictive accuracy and reduced complexity. Results show the SVR-based model effectively captures nonlinear relationships, with strong agreement between predicted and actual values, confirming its robustness for data-driven battery management and microgrid design.

4.1 Data Preprocessing

In the data preprocessing phase, missing values were carefully handled using appropriate imputation to ensure a complete, unbiased dataset. A feature selection technique was then applied to 15,000 samples with 41 battery related attributes, including operational conditions and load profiles. By selecting the most relevant features, dimensionality was reduced while preserving predictive power, improving SVR model accuracy and delivering more precise insights for microgrid battery design. Additionally, normalization was performed to standardize the data, ensuring consistent scale and stability during model training.

4.2 Results

Despite their critical role in the design and optimization of energy storage systems, battery autonomy, battery throughput, and battery life have not been the primary focus in most existing studies. These parameters are essential for accurate battery sizing in microgrids. Battery autonomy indicates how long the battery can supply energy under specific conditions. Without precise estimation of these two metrics, it becomes challenging to design battery systems that are both efficient and reliable. Therefore, greater emphasis should be placed on modeling and analyzing autonomy and throughput in future research to ensure the development of robust microgrid solutions. Moreover, integrating these metrics into predictive models can enhance system-level optimization and long-term performance evaluation.

4.2.1 Performance Analysis Using Harmony Search For Battery Autonomy With Cleaned and Uncleaned Data by Varying Epsilon

TABLE 4.1: Battery Autonomy With Harmony Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.0017	0.9985	0.9857	0.9123
0.1	0.0024	0.9977	0.9857	0.9123
0.5	0.0079	0.9847	0.9855	0.9119
1	0.0189	0.9827	0.9843	0.9093

Tables 4.1 and 4.2 compare SVR performance across epsilon (ϵ) values using uncleaned and cleaned data. On uncleaned data, higher ϵ increases RMSE and weakens LCC, though SROCC and KCC stay stable. After outlier removal, RMSE decreases and correlations improve, confirming enhanced accuracy and robustness. Moreover, performance degradation with larger ϵ is less severe, showing greater resilience. These results highlight the importance of preprocessing for reliable battery performance prediction.

TABLE 4.2: Battery Autonomy With Harmony Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.001614	0.998483	0.985170	0.911297
0.1	0.002250	0.997621	0.985170	0.911297
0.5	0.007884	0.985187	0.984895	0.910663
1	0.017436	0.984999	0.984812	0.910523

4.2.2 Performance Analysis Using Linear Forward Search for Battery Autonomy With Cleaned and Uncleaned Data by Varying Epsilon

TABLE 4.3: Battery Autonomy With Linear Forward Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.0006	0.9900	0.9800	0.9120
0.1	0.0021	0.9980	0.9800	0.9120
0.5	0.0081	0.9770	0.9800	0.9090
1	0.0192	0.9540	0.9600	0.8600

TABLE 4.4: Battery Autonomy With Linear Forward Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.0005	0.9999	0.9851	0.9115
0.1	0.0018	0.9977	0.9851	0.9115
0.5	0.0079	0.9784	0.9850	0.9113
1	0.0182	0.9586	0.9705	0.8806

After applying Linear Feature Selection (LFS), the SVR model's performance was evaluated for varying epsilon (ϵ) values on both uncleaned and cleaned datasets, as

shown in Table 4.3 and 4.4. The goal was to assess whether dimensionality reduction improves robustness to outliers and how ϵ affects accuracy and correlation. On the uncleaned data, smaller ϵ (e.g., 0.01) led to lower RMSE and higher LCC, SROCC, and KCC, indicating better predictions. As ϵ increased, RMSE rose and LCC declined, while SROCC and KCC stayed stable. On the cleaned data, outlier removal improved all metrics, with the best results at $\epsilon = 0.01$. As ϵ increased, RMSE and LCC worsened slightly, SROCC stayed stable, but KCC dropped at $\epsilon = 1$, showing reduced ability to maintain ranking relationships. These results highlight the value of LFS in improving SVR prediction for battery sizing. This approach supports more accurate, data-driven design of microgrid storage systems, ultimately contributing to greater efficiency and reliability in renewable energy integration.

4.2.3 Performance Analysis Using Ranker Search for Battery Autonomy With Uncleaned Data by Varying Epsilon

TABLE 4.5: Battery Autonomy With Ranker Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.0017	0.9985	0.9857	0.9123
0.1	0.0024	0.9977	0.9857	0.9123
0.5	0.0079	0.9847	0.9855	0.9119
1	0.0189	0.9827	0.9843	0.9093

Table 4.5 and 4.6 presents the performance of a Support Vector Regression (SVR) model tuned using the Ranker Search algorithm on uncleaned data—i.e., data with outliers retained. At $\epsilon = 0.01$, the model achieves the lowest RMSE and the highest values for all correlation metrics (LCC, SROCC, KCC), demonstrating that a smaller epsilon margin results in more precise and consistent predictions despite the presence of outliers. As the value of ϵ increases, RMSE rises and LCC

TABLE 4.6: Battery Autonomy With Ranker Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.001614	0.998483	0.985170	0.911297
0.1	0.002250	0.997621	0.985170	0.911297
0.5	0.007884	0.985187	0.984895	0.910663
1	0.017436	0.984999	0.984812	0.910523

(linear correlation) gradually declines, indicating a drop in predictive accuracy due to the model's increased tolerance for error. Meanwhile, SROCC and KCC remain relatively steady up to $\epsilon = 0.5$, reflecting stable rank-based correlations. However, at $\epsilon = 1$, both metrics decline—especially KCC, which highlights a reduction in ordinal consistency and ranking reliability of the SVR model's outputs. These results emphasize the sensitivity of the Ranker Search approach to epsilon tuning. Proper parameter selection is therefore essential for balancing accuracy and stability in battery performance prediction.

4.2.4 Performance Analysis Using Harmony Search for Battery Throughput With Uncleaned Data by by Varying Epsilon

TABLE 4.7: Battery Throughput With Harmony Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.8334	0.9911	0.8710	0.7630
0.1	0.8354	0.9909	0.8704	0.7593
0.5	2.0450	0.9445	0.8680	0.7473
1	4.7724	0.9288	0.8601	0.7249

Table 4.7 and 4.8 presents the performance of the Support Vector Regression (SVR) model optimized using the Harmony Search algorithm for predicting Battery Throughput, evaluated on both uncleaned (with outliers) and cleaned (outliers

removed) datasets.

TABLE 4.8: Battery Throughput With Harmony Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.6350	0.9937	0.9048	0.8128
0.1	0.6427	0.9935	0.9048	0.8081
0.5	1.5177	0.9653	0.9012	0.7885
1	4.3443	0.9521	0.8952	0.7709

The model's behavior is analyzed by varying the ϵ parameter, which defines the SVR's margin of tolerance for prediction error. For the uncleaned data, the best performance is observed at $\epsilon = 0.01$, with the lowest RMSE (0.8334) and highest correlation metrics (LCC = 0.9911, SROCC = 0.8710, KCC = 0.7630). As ϵ increases, the model becomes more tolerant to deviations, resulting in a sharp increase in RMSE and a decline in all correlation metrics, especially at $\epsilon = 1$, where performance significantly deteriorates. In contrast, the cleaned data consistently yields better results across all epsilon values. At $\epsilon = 0.01$, the model achieves its optimal performance with RMSE = 0.6350, LCC = 0.9937, SROCC = 0.9048, and KCC = 0.8128. Even at $\epsilon = 1$, the degradation is much less severe compared to the uncleaned case. These results demonstrate that: Smaller epsilon values lead to more accurate and consistent predictions. Outlier removal significantly enhances the model's robustness and correlation reliability. While increasing ϵ reduces model sensitivity, it must be carefully tuned to avoid compromising performance, particularly in the presence of noise or outliers.

4.2.5 Performance Analysis Using Linear Forward Search for Battery Throughput with Uncleaned and Cleaned Data by Varying Epsilon

Table 4.9 and 4.10 presents the results of applying Linear Forward Search (LFS) feature selection to a Support Vector Regression (SVR) model for predicting Battery Throughput, evaluated on both uncleaned data (with outliers) and cleaned

data (after outlier removal). The performance is analyzed by varying the epsilon (ϵ) parameter, which controls the SVR model's tolerance for error margins during training. For the uncleaned data, smaller ϵ values (e.g., $\epsilon = 0.01$ and 0.1) yield lower RMSE and higher correlation metrics (LCC, SROCC, KCC), indicating that the model performs well even in the presence of outliers when tight error tolerance is applied. As ϵ increases, particularly at $\epsilon = 1$, the RMSE rises significantly and correlation metrics drop, showing a loss in model accuracy and ranking reliability.

TABLE 4.9: Battery Throughput With Linear Forward Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.9874	0.9878	0.8683	0.7503
0.1	0.9998	0.9891	0.8683	0.7485
0.5	1.9243	0.9891	0.8580	0.7268
1	4.5135	0.9203	0.8584	0.7245

TABLE 4.10: Battery Throughput With Linear Forward Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.4853	0.9964	0.9061	0.8188
0.1	0.5133	0.9959	0.9058	0.8133
0.5	1.6088	0.9668	0.9015	0.8133
1	4.8099	0.9158	0.9035	0.8005

For the cleaned data, performance improves significantly across all epsilon values. At $\epsilon = 0.01$, the RMSE drops by almost 50% compared to the uncleaned case, and correlation metrics increase, reflecting better generalization after outlier removal. Even with larger ϵ values like 0.5 and 1, the cleaned data retains higher consistency and correlation than the uncleaned counterpart, though some performance degradation is still observed at $\epsilon = 1$. This demonstrates that data preprocessing not only enhances prediction accuracy but also improves the stability of correlation measures across a wider range of parameter values. Moreover, the results highlight that while small epsilon values are ideal for precision, cleaned datasets

provide resilience even under less optimal parameter choices, ensuring more robust and reliable modeling of battery performance.

4.2.6 Performance Analysis Using RS for BT with Uncleaned and Cleaned Data by Varying Epsilon

Table 4.11 and 4.12 presents the performance of a Support Vector Regression (SVR) model for predicting Battery Throughput using features selected through the Ranker Search (RS) method. The analysis compares results on both uncleaned data (containing outliers) and cleaned data (with outliers removed), across different values of the SVR epsilon parameter (ϵ). It is observed that the removal of outliers significantly improves model accuracy and stability. Furthermore, the cleaned dataset consistently achieves lower RMSE and higher R^2, confirming the importance of preprocessing in enhancing prediction reliability.

TABLE 4.11: Battery Throughput With Ranker Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.8554	0.9908	0.8696	0.7547
0.1	0.7967	0.9917	0.8696	0.7530
0.5	1.9061	0.9917	0.8625	0.7328
1	4.6066	0.9271	0.8542	0.7220

TABLE 4.12: Battery Throughput With Ranker Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	0.000262	0.999959	0.985170	0.911297
0.1	0.002134	0.997297	0.985170	0.911297
0.5	0.007513	0.992683	0.985077	0.911056
1	0.017592	0.992540	0.985077	0.910969

For the uncleaned dataset, the best performance occurs at $\epsilon = 0.1$, with the lowest RMSE and highest correlation metrics, while larger ϵ values increase error and

reduce consistency, especially KCC. In contrast, the cleaned dataset shows substantial improvement across all settings, with RMSE near zero and correlations close to 1 at $\epsilon = 0.01$, and consistently strong results even at higher ϵ. These findings highlight the importance of outlier removal and feature selection in enhancing SVR accuracy, stability, and reliability for battery sizing in microgrids.

4.2.7 Performance Analysis Using Harmony Search for Battery Life with Uncleaned and Cleaned Data by Varying Epsilon

TABLE 4.13: Battery Life With Harmony Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	707.8630	0.9918	0.9363	0.8342
0.1	718.5447	0.9915	0.9377	0.8350
0.5	1587.7944	0.9643	0.9031	0.7682
1	3234.0393	0.9179	0.5911	0.4422

TABLE 4.14: Battery Life With Harmony Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	430.2162	0.9940	0.9752	0.9056
0.1	457.5584	0.9932	0.9749	0.8951
0.5	1097.1627	0.9745	0.8955	0.7472
1	2665.8849	0.9481	0.8737	0.7102

Table 4.13 and 4.14 presents a comparative performance analysis using Harmony Search for battery life prediction with uncleaned and cleaned data as the epsilon (ϵ) parameter varies. The results show that for uncleaned data, increasing ϵ leads to a significant degradation in model performance. Specifically, RMSE rises sharply from approximately 707.86 at $\epsilon = 0.01$ to over 3200 at $\epsilon = 1$, while correlation metrics (LCC, SROCC, and KCC) decline noticeably, indicating poorer predictive accuracy and weaker rank correlations. In contrast, with cleaned data, the model

consistently achieves lower RMSE and higher correlation values across all ϵ levels. Even as ϵ increases, the drop in performance is less severe compared to uncleaned data. For example, at $\epsilon = 0.01$, RMSE is reduced to around 430.21, with LCC, SROCC, and KCC all exceeding 0.90, demonstrating strong linear and rank-order relationships. This comparison highlights the importance of data cleaning in improving model robustness and predictive accuracy when using Harmony Search for battery life estimation.

4.2.8 Performance Analysis Using LFS for Battery Life with Uncleaned and Cleaned Data by Varying Epsilon

TABLE 4.15: Battery Life With Linear Forward Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	121.9373	0.9998	0.9731	0.9192
0.1	313.6227	0.9984	0.9528	0.8718
0.5	1560.2174	0.9692	0.8750	0.7514
1	3492.5126	0.9554	0.8971	0.7725

TABLE 4.16: Battery Life With Linear Forward Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	105.4754	0.9996	0.9933	0.9612
0.1	219.0220	0.9985	0.9836	0.9195
0.5	1013.6790	0.9758	0.9069	0.7616
1	2832.2980	0.9614	0.9521	0.8240

Table 4.15 and 4.16 shows the SVR model's performance using Linear Feature Selection (LFS) under different ϵ values, both before and after data cleaning.At $\epsilon = 0.01$, the model yields the best overall performance in both uncleaned and cleaned datasets. The cleaned data shows slightly higher rank-based correlation metrics (SROCC = 0.9907, KCC = 0.9410) but a slightly higher RMSE. At $\epsilon = 0.1$,

RMSE is minimized (242.6478) in the cleaned dataset while maintaining strong correlation scores, suggesting this setting offers a good trade-off between accuracy and consistency. As ϵ increases to 0.5 and 1, performance (especially RMSE and SROCC) drops in both datasets. However, cleaned data consistently results in lower RMSE and higher rank-based metrics, highlighting its effectiveness in improving robustness. Overall, data cleaning enhances ranking correlations, particularly at higher ϵ, and leads to more stable and interpretable SVR models when using LFS.

4.2.9 Performance Analysis Using RS for Battery Life with Uncleaned and Cleaned Data by Varying Epsilon

TABLE 4.17: Battery Life With Ranker Search: Performance Analysis with Uncleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	161.1253	0.9996	0.9717	0.9141
0.1	312.6126	0.9984	0.9480	0.8653
0.5	1499.2330	0.9695	0.8460	0.7169
1	3604.7290	0.9601	0.8693	0.7380

TABLE 4.18: Battery Life With Ranker Search: Performance Analysis with Cleaned Data

Epsilon (ϵ)	**RMSE**	**LCC (Pearson)**	**SROCC**	**KCC (Kendall)**
0.01	120.9112	0.9966	0.9813	0.9458
0.1	241.7497	0.9980	0.9851	0.9259
0.5	1085.3274	0.9729	0.9706	0.8728
1	2717.6412	0.9522	0.9562	0.8289

Table 4.17 and 4.18 shows that increasing epsilon (ϵ) degrades RS performance on uncleaned data, with higher RMSE and weaker correlations. Cleaned data

consistently yields lower RMSE and stronger correlations, confirming the benefit of data cleaning for accurate battery life prediction.

4.2.10 Effect of Epsilon on Evaluation Parameters

The epsilon ϵ parameter in Support Vector Regression (SVR) plays a crucial role in controlling the margin of tolerance around the actual target value, within which no penalty is given during training. To understand its impact, epsilon values were varied while predicting Battery Autonomy,Batery Throughput and Battery Life and the corresponding evaluation metrics Root Mean Squared Error (RMSE), Linear Correlation Coefficient (LCC), Spearman's Rank-Order Correlation Coefficient (SROCC), and Kendall's Rank Correlation Coefficient (KCC) were analyzed.

4.2.11 Comparative Performance of Feature Selection Methods (HS, LFS, and RS) for BA, BT and BL

The comparative performance of feature selection methods for Battery Autonomy, Battery Life, and Battery Throughput is shown below in table 4.19 .

TABLE 4.19: Comparative Performance of Feature Selection Methods for Battery Metrics

Metric	**Method**	**RMSE**	**SROCC**
Battery Autonomy	HS	0.0016	0.9852
	LFS	**0.0005**	**0.9851**
	RS	0.0016	0.9852
Battery Throughput	HS	0.6350	0.9048
	LFS	0.4853	0.9061
	RS	**0.0003**	**0.9852**
Battery Life	HS	430.22	0.9752
	LFS	**105.48**	**0.9933**
	RS	120.91	0.9813

At $\epsilon = 0.01$, Linear Forward Search (LFS) yields the best results for Battery Autonomy (BA) and Battery Life (BL), while Ranker Search (RS) excels in Battery Throughput (BT). Since no method dominates all metrics, feature selection must align with the specific battery characteristic. This underscores the need for flexible, data-driven approaches and adaptive pipelines that can select or combine methods based on the target outcome.

4.3 Statistical Analysis of Battery Metrics

This section analyzes Battery Autonomy, Battery Throughput, and Battery Life, using box plots of SROCC values to compare predictive performance across Harmony Search, Linear Forward Search, and Ranker Search. A focused analysis of Battery Autonomy highlights model performance and shows how box plots reveal data distribution, outliers, and the consistency of SROCC values across methods, illustrating each technique's ability to handle outliers and improve prediction reliability.

4.3.1 Boxplot of Spearman Rank Order Correlation Coefficient (SROCC) by Ranker Serach

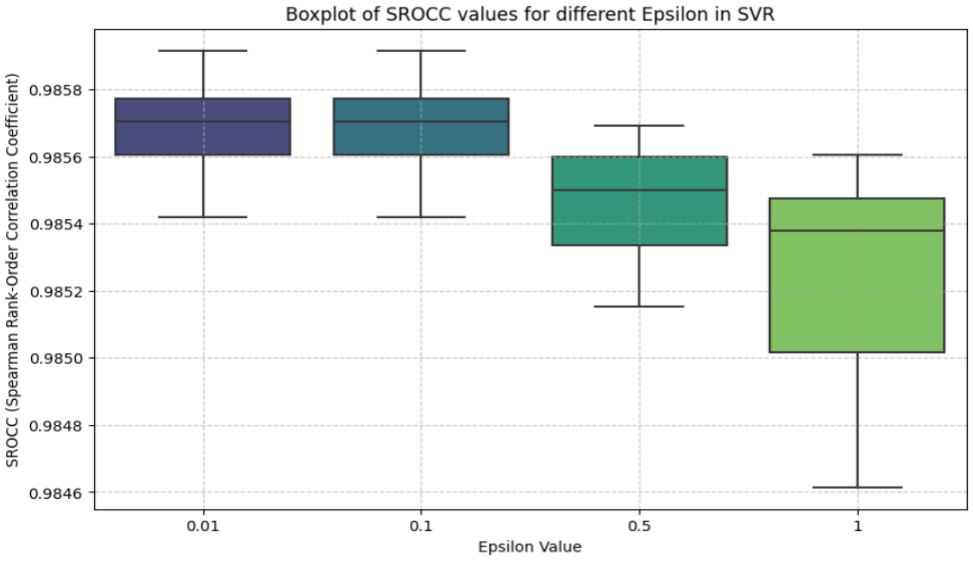

FIGURE 4.1: Box Plot for SROCC of Battery Autonomy Using RS

Figure 4.1 illustrates that, for the given SVR model, smaller epsilon values (0.01 and 0.1) generally result in better and more consistent Spearman Rank-Order Correlation Coefficient (SROCC) values. X-axis (Epsilon Value): This axis shows the four different epsilon values tested: 0.01, 0.1, 0.5, and 1. Epsilon (ϵ) is a hyperparameter in SVR that defines the margin of tolerance within which no penalty is given to errors.

The Y-axis (SROCC) ranges from approximately 0.9846 to 0.9859, representing the rank-order agreement between predicted and actual values, where higher values indicate better performance. Each boxplot corresponds to a specific ϵ, showing the distribution of SROCC values (interquartile range with median), with whiskers extending to 1.5×IQR. No visible outliers are observed in this plot.

4.3.2 Boxplot of Spearman Rank Order Correlation Coefficient (SROCC) by LFS Serach

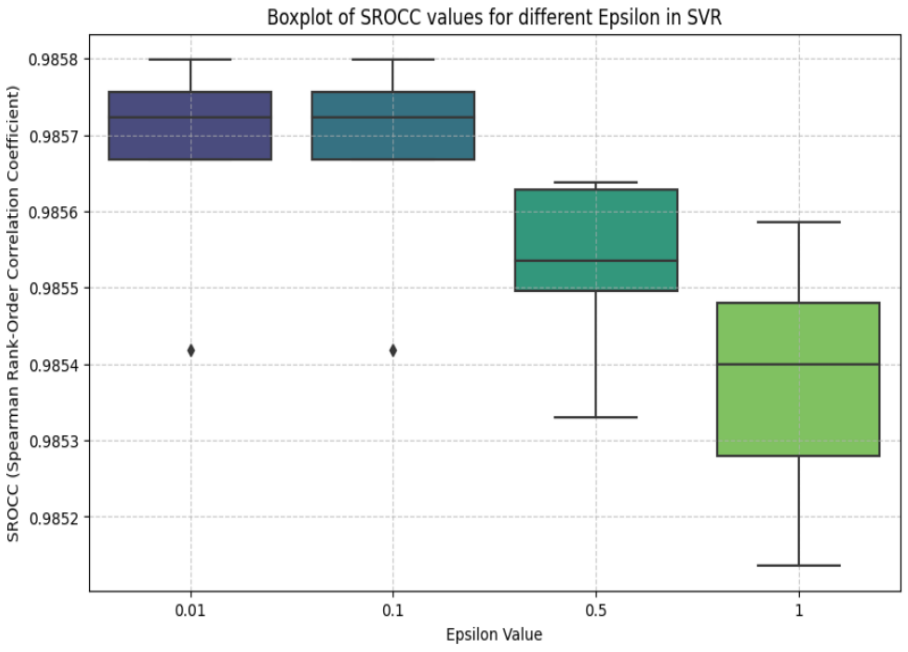

FIGURE 4.2: Box Plot for SROCC of Battery Autonomy Using LFS

Figure 4.2 presents a boxplot that depicts the distribution of Spearman Rank-Order Correlation Coefficient (SROCC) values corresponding to various values of

the Epsilon parameter in a Support Vector Regression (SVR) model. The Epsilon values examined include 0.01, 0.1, 0.5, and 1.0. Each box in the plot represents the interquartile range (IQR) of the SROCC values, with the horizontal line inside the box indicating the median. The "whiskers" extend to the smallest and largest values within 1.5 times the IQR from the lower and upper quartiles, respectively. Data points lying outside this range are marked as outliers.

This visualization provides insight into how the choice of Epsilon impacts the rank-order correlation between predicted and actual values in the SVR model. It helps identify the range and consistency of SROCC scores for different Epsilon values, showing how tighter margins often lead to higher and more stable correlations. By highlighting the presence of outliers, the boxplot also emphasizes variability in model performance and guides the selection of Epsilon values for improved predictive reliability.

4.3.3 Boxplot of Spearman Rank Order Correlation Coefficient (SROCC) by HS

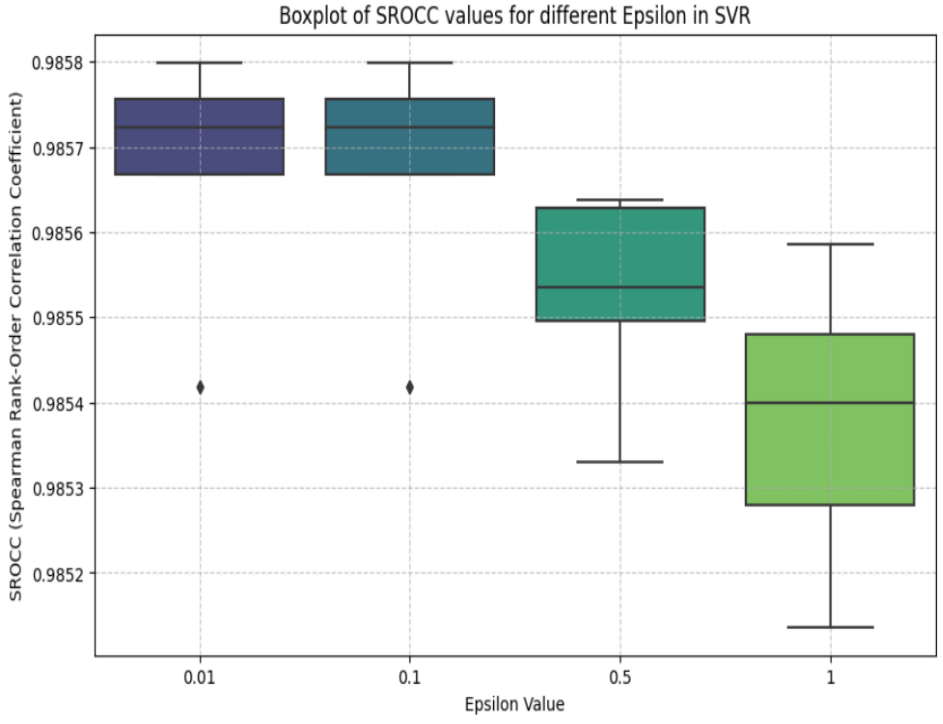

FIGURE 4.3: Box Plot for SROCC of Battery Autonomy Using HS

Figure 4.3 presents a boxplot that illustrates the distribution of Spearman Rank-Order Correlation Coefficient (SROCC) values for different Epsilon settings in a Support Vector Regression (SVR) model. The plot compares four Epsilon values: 0.01, 0.1, 0.5, and 1.0, with each boxplot summarizing the spread of SROCC scores for its respective Epsilon. The central box represents the interquartile range (IQR), the line inside the box indicates the median SROCC value, and the whiskers extend to 1.5 times the IQR to show the typical range of the data. Any points outside this range are marked as outliers.

The purpose of the figure is to highlight how model performance varies with changes in the Epsilon parameter, where higher SROCC values indicate stronger agreement between predicted and actual outcomes. From the figure, it is observed that smaller Epsilon values (0.01 and 0.1) lead to higher and more consistent SROCC scores, while larger Epsilon values (0.5 and 1.0) result in lower and more variable performance.

4.3.4 Outliers for Battery Autonomy

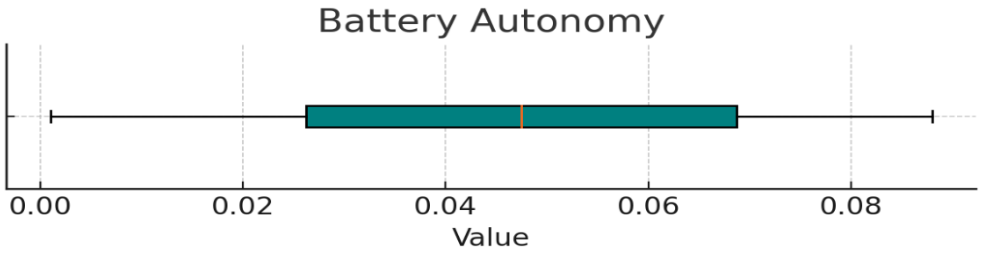

FIGURE 4.4: Outliers for Battery Autonomy

Figure 4.4 explains the distribution of "Battery Autonomy" values. Box: The central box represents the interquartile range (IQR). The bottom edge of the box is the first quartile (Q1), which is approximately 0.02.This means that 25% of the data points for battery autonomy are at or below this value.

The line inside the box is the median (Q2), which is approximately 0.04. This means 50% of the data points are at or below this value. The top edge of the box is the third quartile (Q3), which is approximately 0.07. This means 75% of the

data points are at or below this value. Whiskers: The vertical lines extending from the top and bottom of the box are called whiskers. The bottom whisker extends down to approximately 0.00. This represents the minimum value in the dataset (excluding any outliers). The top whisker extends up to approximately 0.09. This represents the maximum value in the dataset (excluding any outliers).

No Outliers: In this particular boxplot, there are no individual points plotted beyond the whiskers, which indicates that there are no detected outliers in this dataset based on the standard 1.5 * IQR rule for defining outliers.

Y-axis: The vertical axis represents the numerical values of "Battery Autonomy," ranging from 0.00 to 0.09. The specific units or meaning of these values would depend on the context of the data.

4.4 Statistical Analysis of Battery Throughput

4.4.1 Boxplot of SROCC by HS

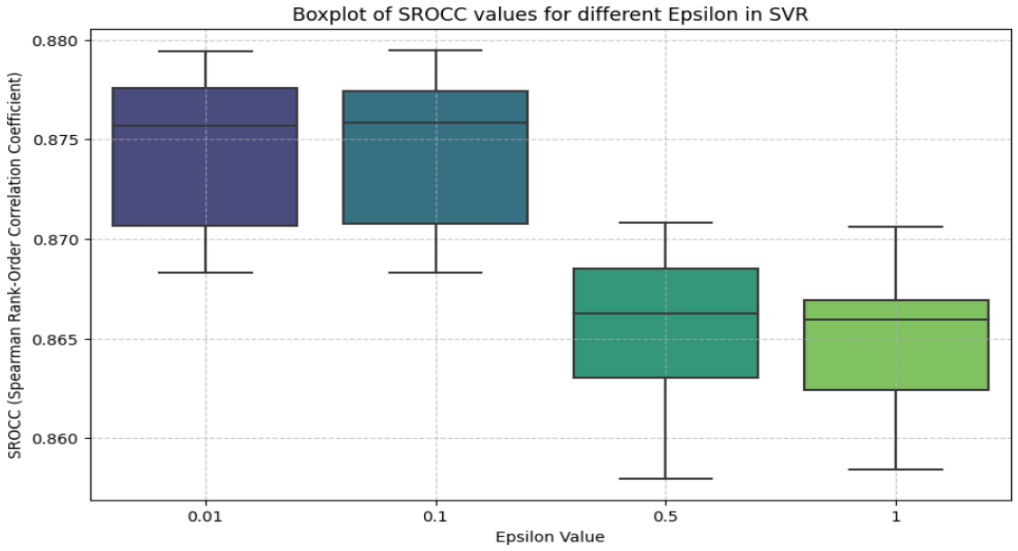

FIGURE 4.5: Box Plot for SROCC Battery Throughput Using HS

Figure 4.5 shows how the SROCC scores (a measure of prediction accuracy) change for an SVR model when different "Epsilon" values (0.01, 0.1, 0.5, 1) are used. The

figure indicates that smaller Epsilon values (0.01 and 0.1) generally result in higher and more consistent SROCC scores, suggesting better model performance.Overall, these results underscore the importance of carefully tuning the Epsilon parameter in SVR models. A smaller Epsilon is generally more suitable when the objective is to achieve high correlation with target values, particularly in applications where model precision and ranking consistency are critical.

These findings reinforce the need for parameter optimization as a key step in developing robust predictive models.The presence or absence of outliers provides additional context regarding system anomalies or rare events that significantly deviate from normal operating conditions. By capturing both central and extreme behavior, the boxplot serves as an effective tool for assessing the overall reliability and robustness of battery throughput in various applications.

4.4.2 Boxplot of SROCC by LFS

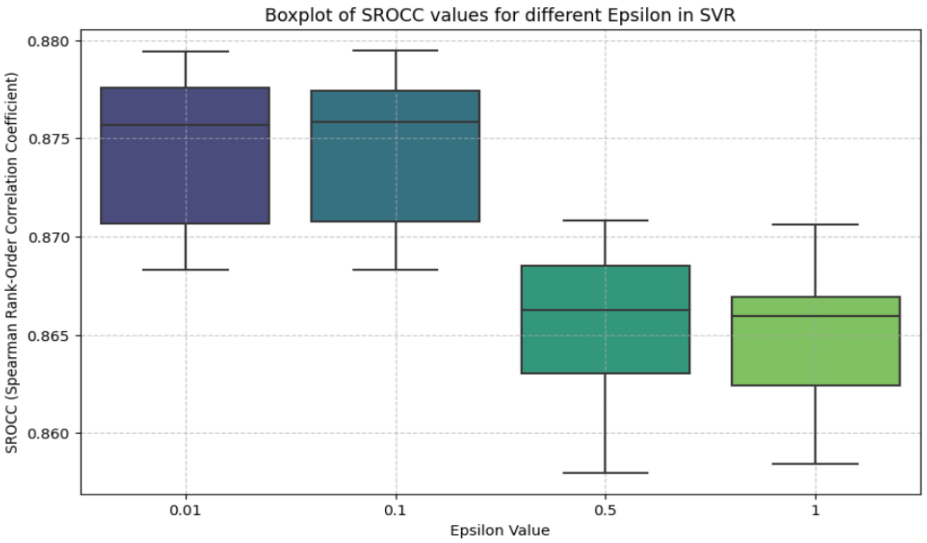

FIGURE 4.6: Box Plot for SROCC Battery Throughput Using LFS

Figure 4.6 uses a boxplot to illustrate the distribution of SROCC values for an SVR model across various Epsilon settings (0.01, 0.1, 0.5, 1). It highlights that Epsilon values of 0.01 and 0.1 tend to produce superior and more stable SROCC scores, implying improved predictive accuracy. The narrower interquartile ranges

at these lower Epsilon values indicate greater consistency in predictions, while higher Epsilon settings show wider variability and more outliers. This suggests that tighter error margins lead to stronger rank-order correlation between predicted and actual values, making them preferable for robust and reliable model performance.

Overall, the boxplot visualization offers clear evidence of the impact of Epsilon selection on model stability, helping to identify optimal parameter settings for minimizing prediction error and maximizing correlation strength. By showing the spread, median, and presence of outliers, it enables a more nuanced understanding of how SVR configurations influence the quality of battery throughput predictions.These insights are valuable for practitioners aiming to fine-tune SVR models for battery throughput prediction in microgrid applications.It also highlights the role of data visualization in revealing patterns and guiding informed model optimization.

4.4.3 Boxplot of SROCC by RS

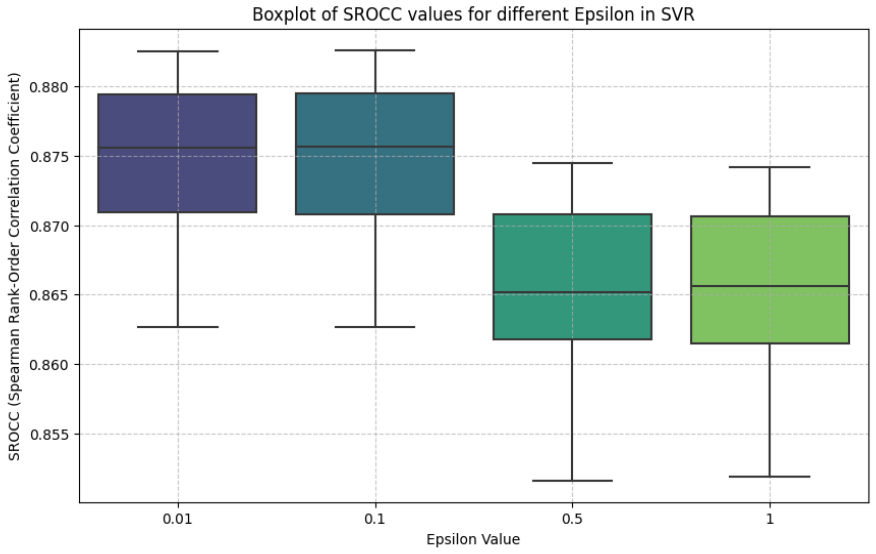

FIGURE 4.7: Box Plot for SROCC Battery Throughput Using RS

Figure 4.7 displays the SROCC scores obtained from an SVR model for different Epsilon values (0.01, 0.1, 0.5, 1). The visual evidence suggests that the model achieves better and more reliable SROCC performance when Epsilon is set to 0.01 or 0.1.

4.4.4 Outliers for Battery Throughput

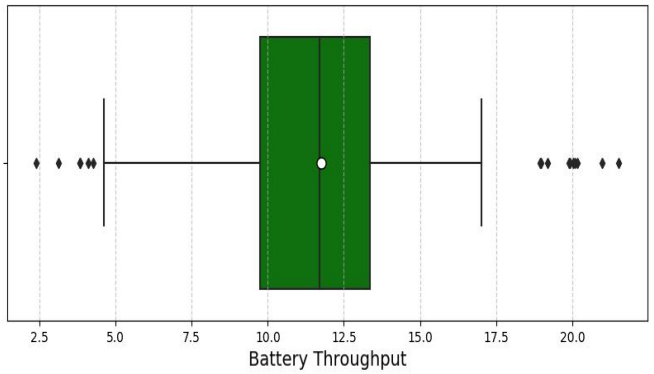

FIGURE 4.8: Outliers for Battery Throughput

Figure 4.8 illustrates the distribution of Battery Throughput values using a box plot. The dark green box represents the interquartile range (IQR), covering the middle 50% of the data between the first (Q1) and third quartiles (Q3). The box length reflects variability within this central range, with a longer box indicating greater spread. The median (Q2), though not clearly visible due to the fill, lies within the box and divides the data into two halves. Its position provides insights into skewness, and numerical analysis confirms a slight lean toward lower values.

4.5 Statistical Analysis of Battery Life

The statistical analysis of Battery Life examines how different feature selection methods influence prediction stability and accuracy. By comparing distributions and correlation values, the analysis highlights consistency across models, identifies potential outliers, and evaluates how effectively each method captures long-term battery degradation patterns.

4.5.1 Boxplot of SROCC HS

Figure 4.9 illustrates how SROCC scores (a measure of how accurate a model's predictions are) vary when an SVR model uses different "Epsilon" values (0.01, 0.1, 0.5, 1, and 2).

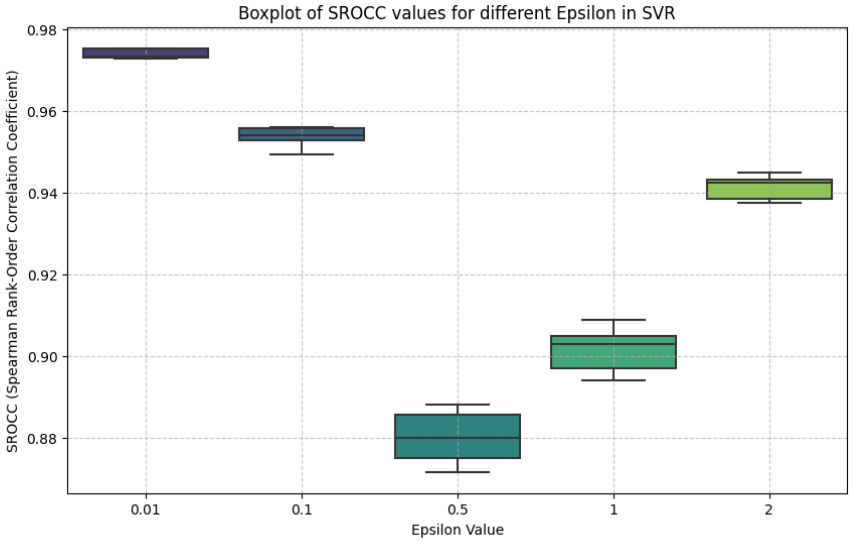

FIGURE 4.9: Box Plot for SROCC of Battery Life Using HS

It shows that smaller Epsilon values (0.01 and 0.1) tend to yield higher and more consistent SROCC scores, indicating better predictive performance, while larger values lead to lower and more varied scores. This pattern suggests that tighter error margins in SVR (lower Epsilon) help capture subtle variations in the data. In contrast, higher Epsilon values allow a wider error margin, reducing sensitivity and degrading correlation with actual outcomes. The figure highlights the importance of careful Epsilon tuning to achieve optimal model accuracy. Overall, it emphasizes that selecting a small Epsilon value is generally more effective for ensuring robust and reliable predictions.

4.5.2 Boxplot of SROCC by LFS

Figure 4.10 is a boxplot that displays the distribution of SROCC values for an SVR model across different Epsilon settings (0.01, 0.1, 0.5, 1, and 2). The graph suggests that the SVR model generally achieves better and more stable SROCC scores with smaller Epsilon values (0.01 and 0.1). As Epsilon increases, the variability of SROCC widens, indicating reduced consistency in prediction accuracy. This trend highlights the importance of careful tuning of Epsilon to maintain robust correlation performance. Overall, smaller Epsilon values provide a more reliable balance between accuracy and stability in battery life prediction.

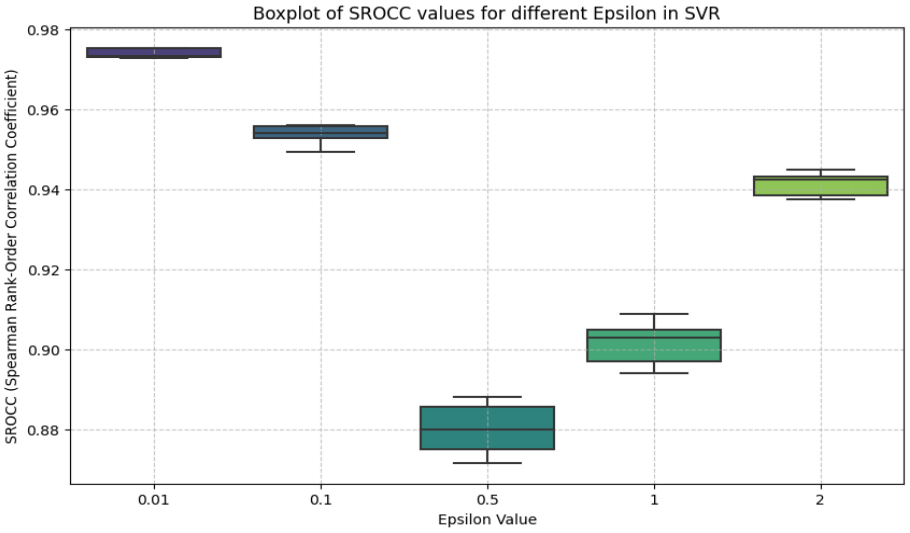

FIGURE 4.10: Box Plot for SROCC of Battery Life Using LFS

4.5.3 Boxplot of Spearman Rank Order Correlation Coefficient (SROCC) by RS

Figure 4.11 shows the SROCC scores of an SVR model tested with different Epsilon values (0.01, 0.1, 0.5, 1, and 2). Results indicate that smaller Epsilon values yield higher and more consistent SROCC with fewer outliers, while larger values increase variability and reduce reliability. This highlights the need to fine-tune Epsilon to balance stability, accuracy, and generalization for robust battery life prediction.

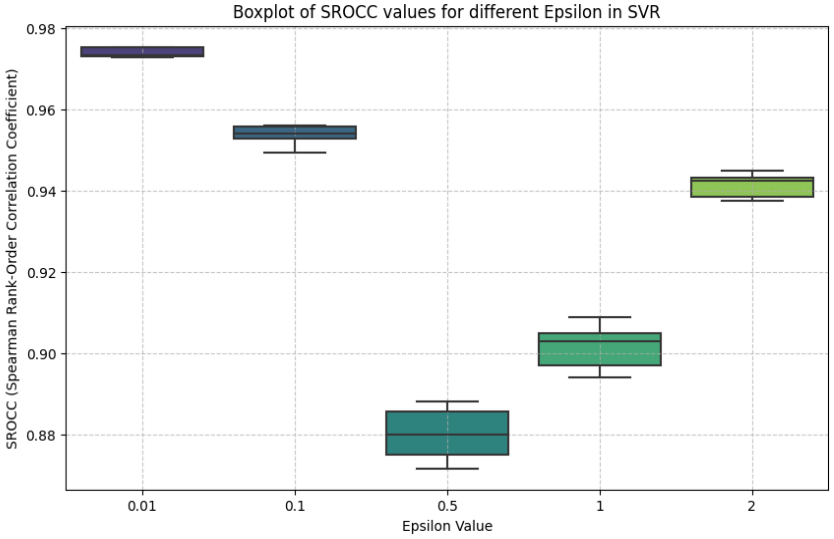

FIGURE 4.11: Box Plot for SROCC of Battery Life Using RS

4.5.4 Outliers for Battery Life

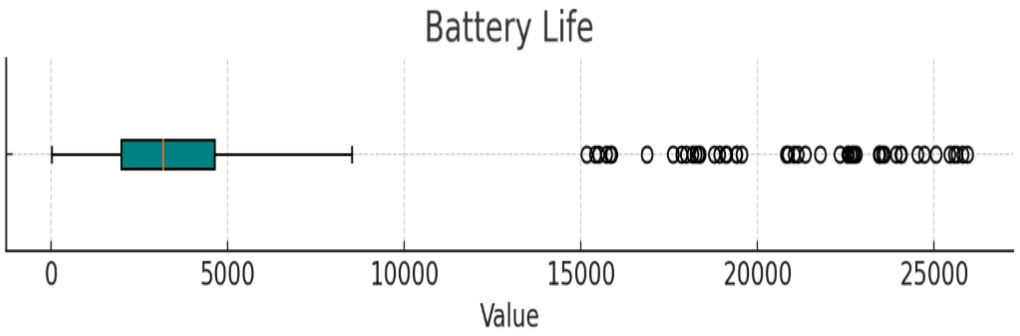

FIGURE 4.12: Outliers for Battery Life

Figure 4.12 summarizes the distribution of battery life values. The central box represents the interquartile range (IQR), which contains the middle 50% of the data. The horizontal line within the box indicates the median battery life. The vertical lines (whiskers) extend from the box to show the typical range of the data, while individual points beyond these whiskers (if any) would represent outliers. In this particular plot, there are many data points above the upper whisker, suggesting a significant number of Battery Life values that are much higher than the majority of the data. This indicates the presence of extreme cases where battery performance exceeds the typical operating range. Such skewness highlights the variability in battery durability under different conditions.

4.6 Impact of Feature Selection and ML on BT and BA

TABLE 4.20: Predicted Values for Battery Autonomy: Comparison Before and After Outlier Removal

Feature Selection Algorithm	Before Outlier Removal				After Outlier Removal			
	LCC	SROCC	KCC	RMSE	LCC	SROCC	KCC	RMSE
LFS	0.9963	0.8793	0.7811	0.5423	0.9989	0.8858	0.7908	0.3038
RS	0.9985	0.9857	0.9123	0.0017	1.0000	0.9581	0.8514	0.0001
HS	0.0005	0.9998	0.9857	0.9123	1.0000	0.9656	0.8734	0.0000

TABLE 4.21: Predicted Values for Battery Throughput: Comparison Before and After Outlier Removal

Feature Selection Algorithm	Before Outlier Removal				After Outlier Removal			
	LCC	**SROCC**	**KCC**	**RMSE**	**LCC**	**SROCC**	**KCC**	**RMSE**
LFS	0.9999	0.9857	0.9124	0.0005	0.9999	0.9844	0.9096	0.0004
RS	0.9899	0.8753	0.7589	0.8978	0.9998	0.7713	0.6693	0.1078
HS	0.6287	0.9950	0.8789	0.7794	1.0000	0.8388	0.7422	0.0503

The tables 4.20 and 4.21 compare LFS, HS, and RS for predicting Battery Autonomy and Throughput before outlier removal. Results show clear differences in accuracy and correlation, with HS adapting better to data variations, while LFS and RS remain competitive. This highlights trade-offs between algorithm complexity and predictive performance.

4.6.1 Best Performing Algorithm: Harmony Search (HS)

Harmony Search (HS) is the most effective feature selection algorithm among the three for this problem, especially after outlier removal, offering superior accuracy and correlation across both output targets: **Battery Autonomy** and **Battery Throughput** as shown in table 4.22.

TABLE 4.22: Comparison of Feature Selection Methods for Battery Autonomy and Throughput (After Outlier Removal)

Metric	Feature Selection	Battery Autonomy	Battery Throughput	Remarks
LCC	HS	**1.0000**	**1.0000**	Perfect linear correlation
SROCC	HS	**0.9656**	0.8388	Best for autonomy, good for throughput
KCC	HS	**0.8734**	0.7422	Strong ordinal agreement
RMSE	HS	**0.0000**	**0.0503**	Lowest error in both targets
LCC	RS	1.0000	0.9998	Excellent linear correlation
SROCC	RS	0.9581	0.7713	Good for autonomy, weaker for throughput
KCC	RS	0.8514	0.6693	Lower ordinal agreement
RMSE	RS	0.0001	0.1078	Good, but not the lowest
LCC	LFS	0.9989	0.9999	Slightly lower than RS/HS
SROCC	LFS	0.8858	0.9844	Very good rank correlation for throughput
KCC	LFS	0.7908	0.9096	Best KCC for throughput
RMSE	LFS	0.3038	0.0004	Best RMSE for throughput only

4.7 Features Selected

To improve the prediction accuracy of battery performance metrics, different wrapper-based feature selection techniques were applied. These methods systematically evaluate subsets of features in combination with the prediction model to identify the most relevant inputs that contribute to higher accuracy. By filtering out redundant or less informative variables, the wrapper approaches not only reduce model complexity but also enhance generalization capability. In this study, techniques such as Linear Forward Search (LFS), Harmony Search (HS), and Ranker Search (RS) were employed to optimize the input space, ensuring that the selected features provide meaningful insights into predicting both battery autonomy and battery throughput. This comparative analysis highlights the importance of feature selection in balancing accuracy, robustness, and interpretability of the results.

4.7.1 Features Selected By HS

TABLE 4.23: Top features selected by Harmony Search for Battery Life, Battery Throughput, and Battery Autonomy.

Target Variable	**Selected Features**
Battery Life	Battery, Total Capital Cost, Tot. Ann. Cap. Cost, Operating Cost, COE, PV Production, Ren. Fraction, Battery Life
Battery Throughput	Battery, Converter, Total Capital Cost, Tot. Ann. Repl. Cost, Total Ann. Cost, DGEN Life, Battery Life, Battery Throughput
Battery Autonomy	Grid, DGEN, Tot. Ann. Repl. Cost, Total O&M Cost, Operating Cost, PV Production, Diesel, CO2 Emissions

Table 4.23 represents the best subsets identified by the Harmony Search method. The selected features combine both economic (capital and operating costs, COE) and technical parameters (battery, PV production, renewable fraction). This indicates that Harmony Search balances cost-efficiency with energy generation aspects for accurate prediction, highlighting its ability to capture both financial and operational dimensions of microgrid performance.

4.7.2 Features Selected By LFS

TABLE 4.24: Top features selected by Linear Forward Search for Battery Life, Battery Throughput, and Battery Autonomy.

Target Variable	**Selected Features**
Battery Life	Battery, AC Primary Load Served, Cap. Shortage, Excess Electricity, Battery Autonomy, Battery Throughput
Battery Throughput	Grid, DGEN, Battery, AC Primary Load Served, Battery Autonomy, Battery Life
Battery Autonomy	Grid, DGEN, Battery, Excess Electricity, Battery Life, Battery Throughput

Table **??** represents the best subsets identified by the Linear Forward Search (LFS) method. The selected features primarily emphasize technical parameters such as battery performance, PV production, and renewable fraction, while also incorporating certain economic indicators to a lesser extent. This shows that LFS tends to prioritize variables directly linked to system reliability and operational performance, making it effective for capturing the technical behavior of battery autonomy and life. Moreover, by sequentially adding features based on their contribution, LFS ensures a systematic and interpretable selection process, which helps in building models that are both efficient and transparent for practical microgrid applications.

4.7.3 Features Selected By Ranker Search

Table 4.25 represents the best subsets identified by the Ranker Search (RS) method. The selected features cover a wide range of technical, economic, and environmental variables, reflecting the method's comprehensive search strategy. This indicates that Ranker Search captures diverse system interactions, ensuring that no potentially relevant attribute is overlooked. However, the inclusion of a broad set of features can also lead to redundancy and reduced interpretability, as some variables may contribute overlapping information. Despite this, RS remains useful for uncovering hidden correlations and complex dependencies, making it particularly suitable when the goal is to maximize overall predictive accuracy rather than prioritize feature compactness.

TABLE 4.25: Top features selected by Ranker Search (F-score) for Battery Life, Battery Throughput, and Battery Autonomy.

Target Variable	**Selected Features**
Battery Life	Grid, DGEN, Battery, Converter, Total NPC, Tot. Ann. Repl. Cost, Total Fuel Cost, Total Ann. Cost, Operating Cost, COE, DGEN Production, Grid Purchases, Grid Net Purchases, Tot. Electrical Production, AC Primary Load Served, Cap. Shortage, Unmet Load, Unmet Load Frac., Excess Electricity, Diesel, CO Emissions, UHC Emissions, PM Emissions, SO2 Emissions, NOx Emissions, DGEN Fuel, DGEN Hours, Battery Autonomy, Battery Throughput
Battery Throughput	Grid, DGEN, Battery, Converter, Total NPC, Tot. Ann. Repl. Cost, Total Fuel Cost, Total Ann. Cost, Operating Cost, COE, DGEN Production, Grid Purchases, Grid Net Purchases, Tot. Electrical Production, AC Primary Load Served, Cap. Shortage, Unmet Load, Unmet Load Frac., Excess Electricity, Diesel, CO Emissions, UHC Emissions, PM Emissions, SO2 Emissions, NOx Emissions, DGEN Fuel, DGEN Hours, Battery Autonomy, Battery Life
Battery Autonomy	Grid, DGEN, Battery, Converter, Total Capital Cost, Tot. Ann. Cap. Cost, Tot. Ann. Repl. Cost, Total O&M Cost, Total Fuel Cost, DGEN Production, Grid Purchases, Grid Net Purchases, Tot. Electrical Production, AC Primary Load Served, Cap. Shortage, Unmet Load, Unmet Load Frac., Excess Electricity, Diesel, CO Emissions, UHC Emissions, PM Emissions, SO2 Emissions, NOx Emissions, DGEN Fuel, DGEN Hours, DGEN Starts, Battery Life, Battery Throughput

Table 4.25 represent the best subsets identified by the Ranker Search method. The selected features cover a wide range of technical, economic, and environmental variables. This indicates that Ranker Search captures diverse system interactions, but may include redundant features due to its broad selection strategy.

4.8 Statistical Analysis of Data

A scatter plot illustrates the relationship between two variables by plotting data points, making correlations, trends, and outliers easily observable. In this work, scatter plots are combined with Linear and Polynomial Regression models to compare prediction performance, assess whether system behavior follows a linear or nonlinear trend, and guide the selection of the most suitable regression approach for accurate battery performance estimation. Additionally, this visual analysis

reveals model limitations by highlighting deviations where regression lines fail to capture underlying data patterns.

4.8.1 Statistical Analysis With Linear Regression

Linear regression provides a baseline by assuming a straight-line relationship between selected features (via Harmony Search) and target variables such as Battery Throughput and Battery Autonomy.

4.8.2 Scatter Plots for Linear Forward Search

This figure contains two scatter plots, each demonstrating the predictive accuracy of a model using a Linear Forward Search method for different battery metrics.

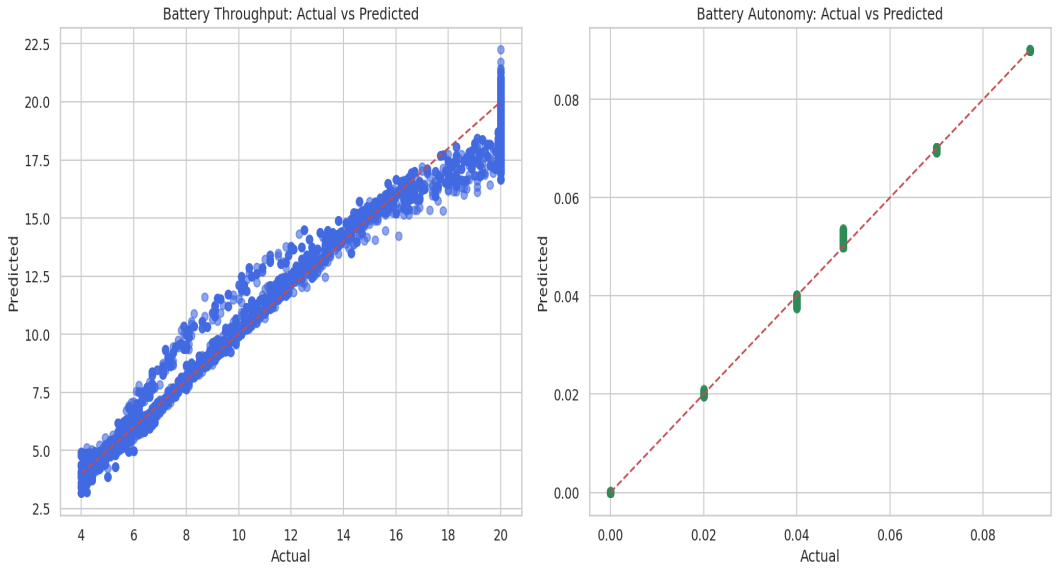

FIGURE 4.13: LFS Scatter Plot for BA/BT

Figure 4.13 presents two scatter plots comparing actual versus predicted values for battery metrics using the Linear Forward Search method. Left Plot (Battery Throughput): The x-axis shows true throughput values, and the y-axis shows predicted values. Blue points cluster tightly along the dashed red y=x line, indicating strong correlation and high prediction accuracy. Right Plot (Battery Autonomy): Similarly, green points align closely with the ideal line, reflecting highly accurate

predictions for battery autonomy. The figure also compares polynomial regression fits of varying degrees (3, 4, 5, and 8) alongside linear fits, assessing how well they capture the true relationship. Higher-degree fits reveal complex, nonlinear patterns, while simpler fits aim to balance accuracy with reduced risk of overfitting.

4.8.3 Outliers Detection

Figure 4.14 also presents two scatter plots, comparing actual versus predicted values for battery metrics derived using a Linear Forward Search method, with an added emphasis on outliers.

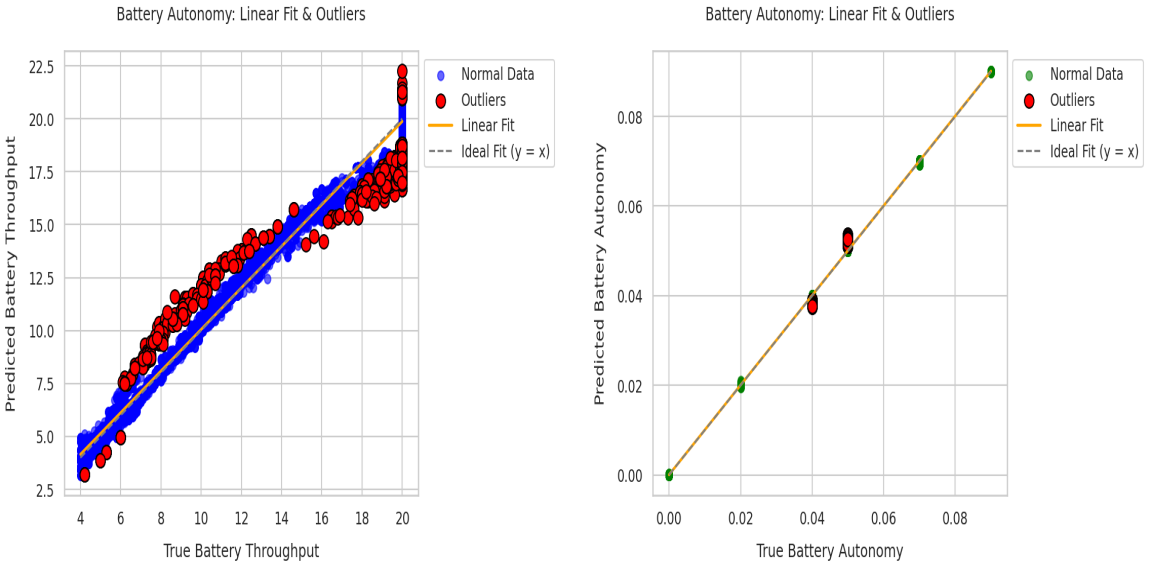

FIGURE 4.14: Showing Outliers for BA/BT using LFS

Left Plot: Predicted Battery Throughput vs True Battery Throughput: This plot evaluates the model's predictions for "Battery Throughput". The "True Battery Throughput" is on the x-axis, and "Predicted Battery Throughput" is on the y-axis. Most data points are shown in blue ("Normal Data"), clustering around an orange "Linear Fit" line and a dashed "Ideal fit (y = x)" line. Red-outlined points are labeled "Outliers," indicating values that deviate more significantly from the general linear trend. The overall tight grouping of points around the ideal line, despite some outliers, suggests the model generally makes accurate predictions for battery throughput. Right Plot: Predicted Battery Autonomy vs True Battery

Autonomy: This graph assesses the predictive accuracy for "Battery Autonomy". "True Battery Autonomy" is on the x-axis, and "Predicted Battery Autonomy" is on the y-axis. The green data points ("Normal Data") lie almost exactly on both the orange "Linear Fit" and the dashed "Ideal fit (y = x)" lines. A few red-outlined points ("Outliers") are also present, but they too are very close to the ideal prediction line. This remarkable closeness of all points to the ideal line signifies that the model achieves extremely high accuracy in predicting Battery Autonomy.

4.8.4 Scatter Plots for Ranker Search

Figure 4.15 presents two scatter plots comparing actual values against predicted values for battery performance metrics.

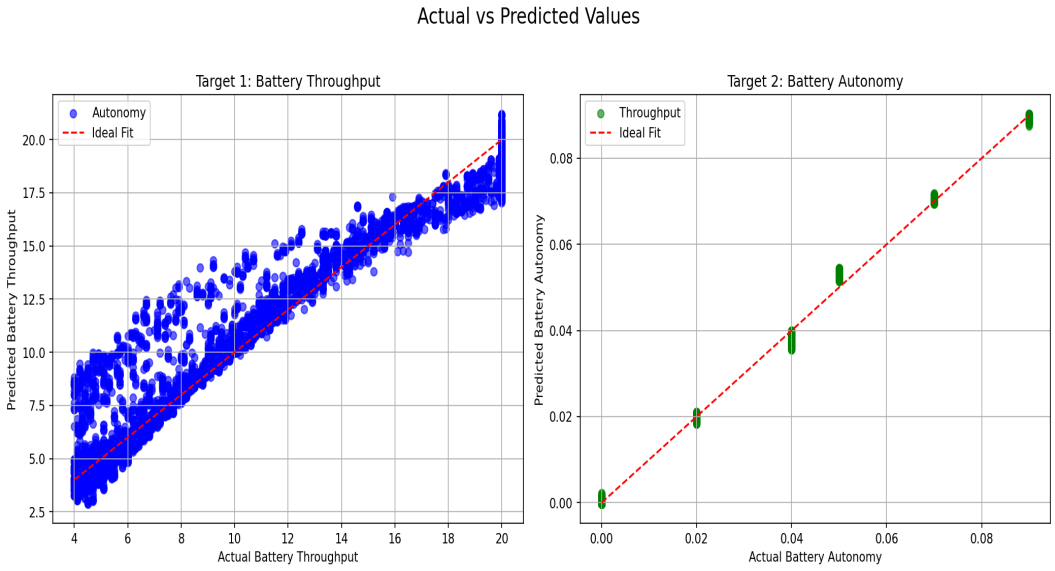

FIGURE 4.15: RS Scatter Plot for BA/ BT

The left plot illustrates the model's accuracy in predicting battery throughput. The x-axis displays the actual throughput values, while the y-axis shows the predicted values. Each blue dot represents a single data point, showing its actual and predicted pair. The dashed red line, labeled "Ideal Fit," serves as a benchmark for perfect prediction where predicted values equal actual values. The close clustering of the blue points around this ideal line indicates a strong positive correlation and

high accuracy in the model's predictions for battery throughput. The right plot evaluates the model's predictive performance for battery autonomy. The x-axis shows the actual battery autonomy values, while the y-axis shows the predicted values. The green data points are almost perfectly aligned with the dashed red "Ideal Fit" line. This close alignment indicates an exceptionally strong correlation and highly precise prediction accuracy for battery autonomy by the model.

4.8.5 Outliers Detection

Figure 4.16 also presents two scatter plots comparing actual versus predicted values for battery metrics derived using the Ranker Search method, with an added emphasis on outliers.

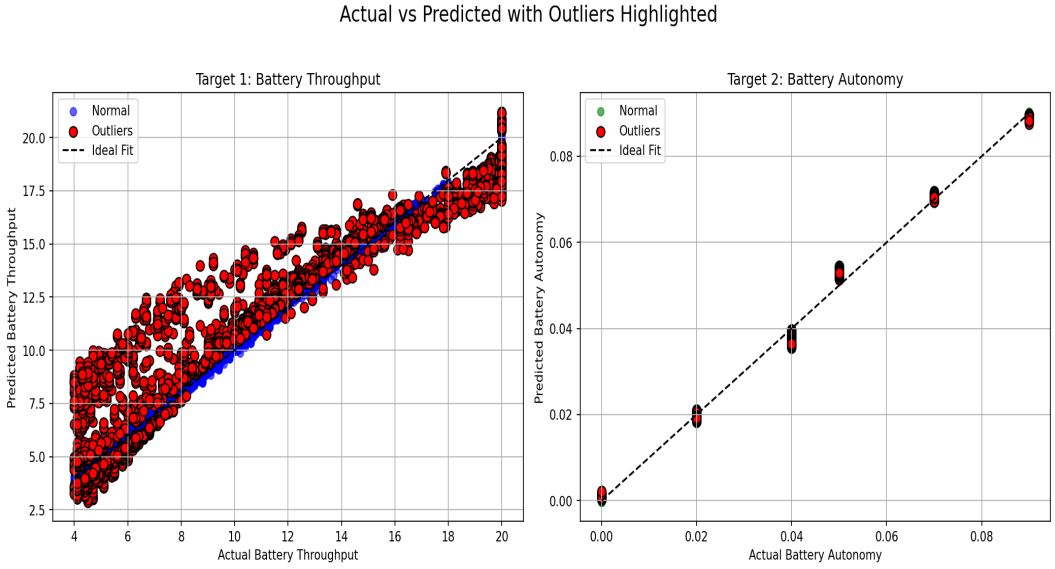

FIGURE 4.16: Showing Outliers for BA/BT Using RS

Figure 4.16 right plot assesses the predictive accuracy for battery autonomy. The left plot visualizes the model's predictions for battery throughput. The x-axis shows the actual battery throughput, and the y-axis shows the predicted values. Normal data points are shown in blue, while outliers are clearly marked in red with black borders. The dashed black line represents the ideal fit for perfect prediction. Despite the presence of some outliers, most points cluster closely around this ideal line, demonstrating overall strong predictive performance for battery throughput,

even though some data points show greater deviations.The x-axis shows the actual battery autonomy, and the y-axis shows the predicted values. Normal data points appear in green, while outliers are marked in red with black borders. Most points, including those identified as outliers, lie very close to the dashed black ideal fit line. This strong alignment indicates that the model delivers highly precise predictions for battery autonomy, with even the outliers showing only minimal deviation from the ideal performance.

4.8.6 Scatter Plots for Harmony Search

Figure 4.17 contains two scatter plots, each demonstrating the predictive accuracy of a model using a Harmony Search method for different battery metrics.

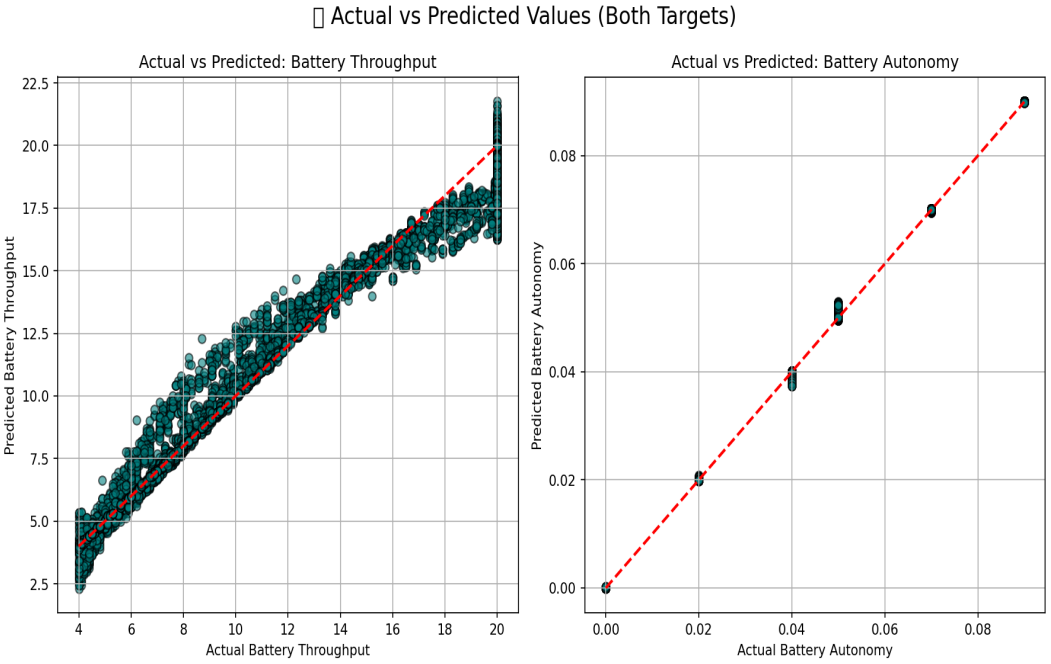

FIGURE 4.17: HS Scatter Plot for BA/ BT

This figure shows two scatter plots of the model's predictions for battery throughput and battery autonomy. In the first plot, actual versus predicted throughput values cluster tightly around the dashed red "Ideal Fit" line, indicating strong correlation and high accuracy. In the second plot, predicted autonomy values also align closely with the ideal line, demonstrating very strong correlation and precise predictive performance.

4.8.7 Outliers Detection

Figure 4.18 also presents two scatter plots, comparing actual versus predicted values for battery metrics derived using a Harmony Search method, with an added emphasis on outliers.

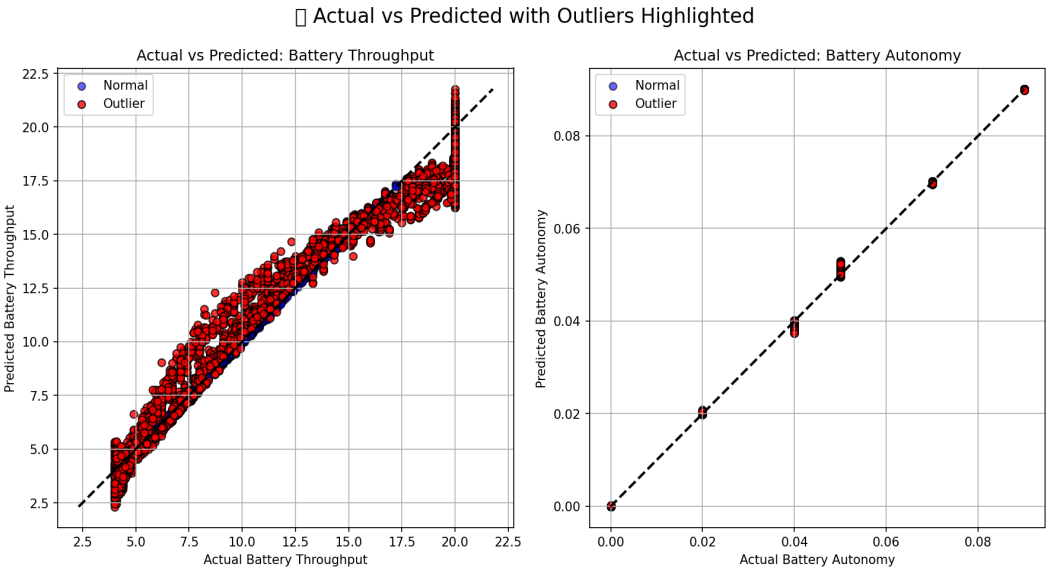

FIGURE 4.18: Showing Outliers for BA/BT Using HS

Figure 4.18 presents scatter plots of actual vs. predicted values for battery throughput and autonomy. For throughput, most points (blue) cluster near the dashed "Ideal Fit" line, with a few red outliers showing minor deviations. For autonomy, points (green) lie very close to the ideal line, indicating highly accurate predictions with minimal error.

4.9 Statistical Analysis With Polynomial Regression

The polynomial regression model extends linear regression by incorporating non-linear terms (e.g., x^2, x^3). This allows the model to capture curved and more complex patterns in the relationship between the selected features and battery performance.

4.9.1 Scatter Plot with Linear Forward Search

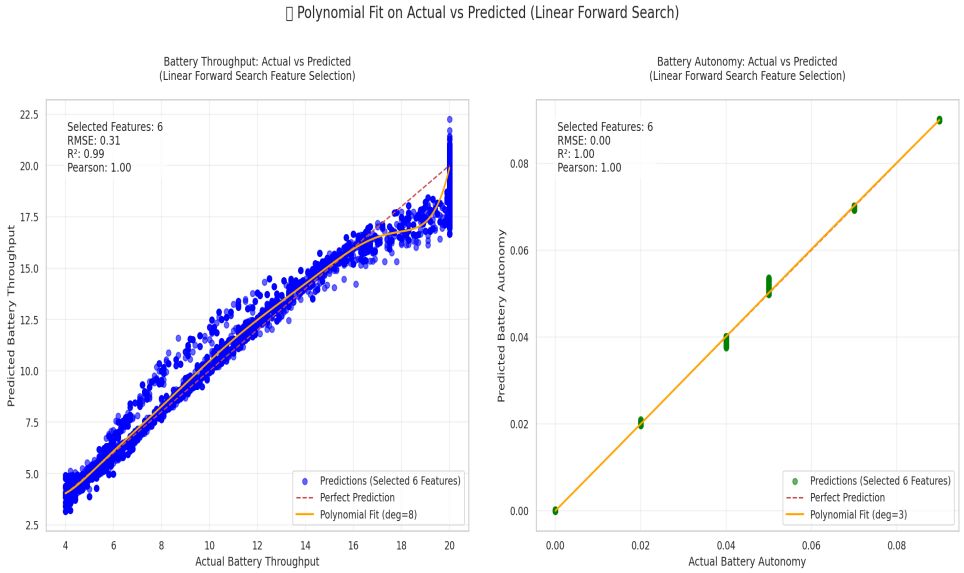

FIGURE 4.19: Scatter Plot for LFS with Polynomial Regression

Figure 4.19 shows the actual vs. predicted values for Battery Throughput and Battery Autonomy using Linear Forward Search (LFS). The results demonstrate that the selected features achieve near-perfect prediction accuracy, as indicated by low RMSE and high R^2 values, while the polynomial fit effectively captures nonlinear relationships.

4.9.2 Outliers with Polynomial Regression

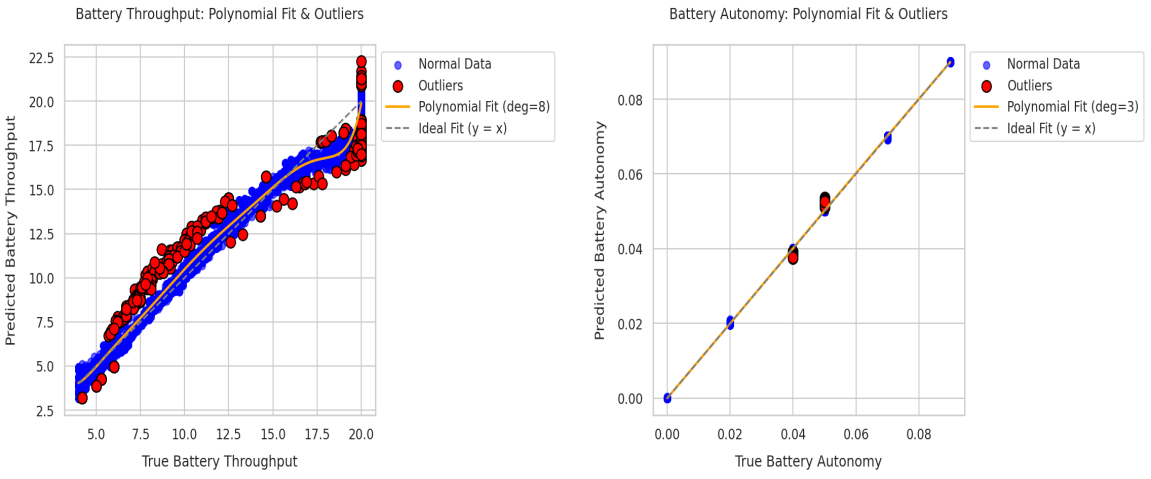

FIGURE 4.20: Scatter Plot for LFS with Polynomial Regression

The scatter plot in Figure 4.20 shows the actual versus predicted values for Battery Throughput and Battery Autonomy using LFS with polynomial regression. The model effectively captures nonlinear patterns, but noticeable outliers deviate from the prediction line, indicating regions of reduced accuracy and the need to assess their impact on overall performance.

4.9.3 Scatter Plot with Harmony search

Figure 4.21 presents the polynomial regression fitting results for Battery Throughput and Battery Autonomy after applying Harmony Search (HS) for feature selection.In the case of Battery Throughput, the polynomial regression of degree 8 closely follows the distribution of predicted points, with only minor deviations observed at higher values, indicating that the selected features contribute to improved model accuracy. For Battery Autonomy, the polynomial regression of degree 3 almost overlaps with the ideal reference line, highlighting a near-perfect agreement between the actual and predicted values.

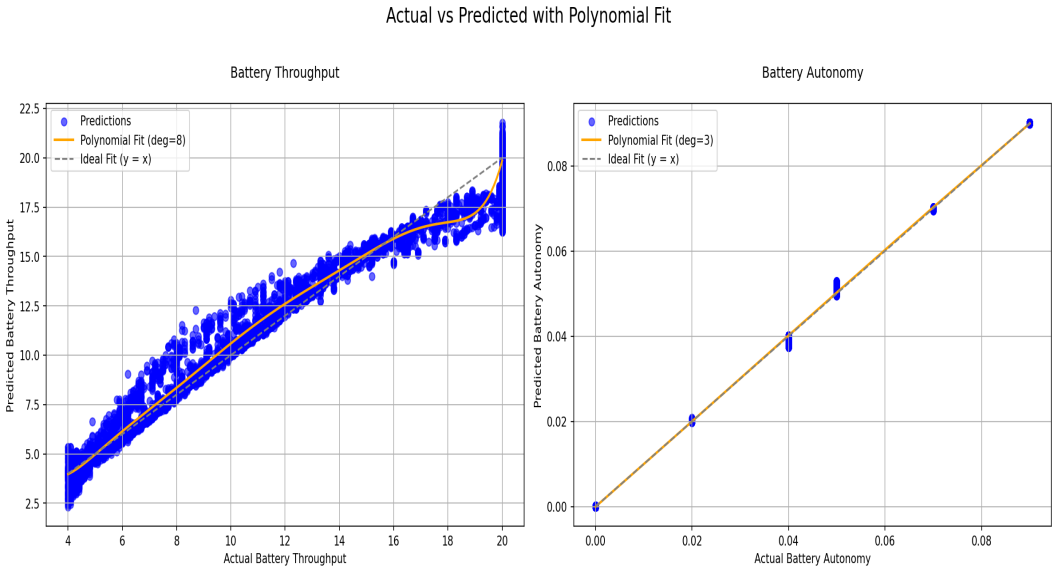

FIGURE 4.21: Scatter Plot for HS with Polynomial Regression

These results confirm that the HS-based feature selection effectively eliminates irrelevant and redundant variables, allowing the model to better capture the underlying patterns in the data. Consequently, HS enhances both the predictive

accuracy and robustness of the model, demonstrating its effectiveness in optimizing feature subsets for battery performance prediction.

4.9.4 Outliers of HS with Polynomial Regression

Figure 4.22 illustrates the outlier detection results using polynomial residual analysis for Battery Throughput and Battery Autonomy after Harmony Search feature selection. For Battery Throughput, the polynomial fit of degree 8 captures the general trend of the predictions, while several points (highlighted in red) deviate significantly from the fitted curve, indicating outliers. These outliers are primarily located at lower and higher throughput values, suggesting that extreme operational conditions lead to deviations from expected predictions.

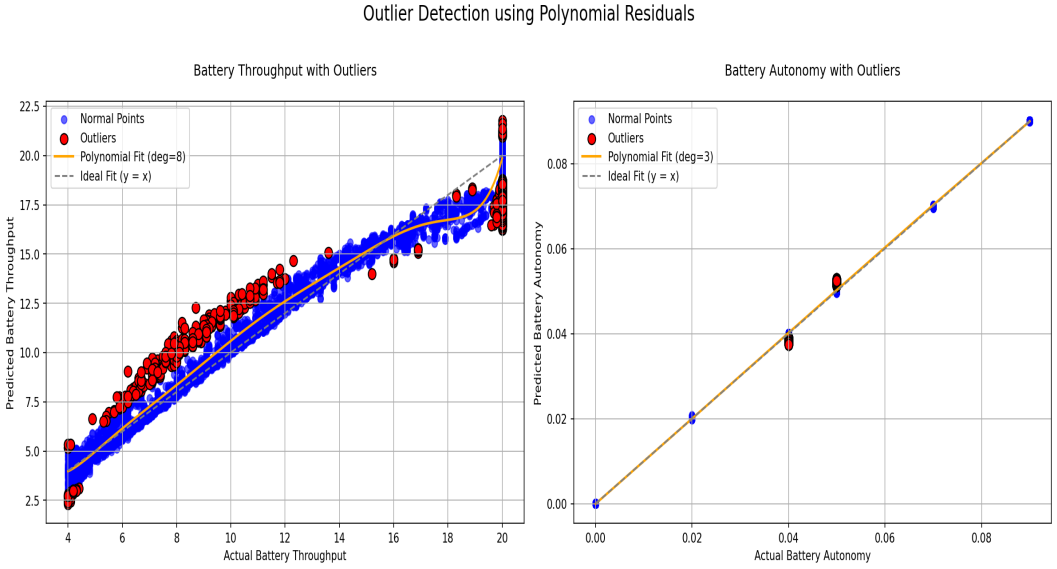

FIGURE 4.22: Scatter Plot for HS with Polynomial Regression

In contrast, the Battery Autonomy plot with a degree 3 polynomial fit shows only a few scattered outliers, while the majority of points align closely with the ideal fit line. This indicates that Battery Autonomy predictions are more stable and less sensitive to extreme variations. Overall, the analysis confirms that while polynomial regression improves fit accuracy, residual-based outlier detection is crucial for identifying abnormal data points that may otherwise bias the model's performance evaluation.

4.9.5 Scatter Plot with RS

Figure 4.23 shows polynomial regression results for Battery Throughput and Autonomy using Ranker Search features. For throughput, a degree-8 polynomial fit captures the nonlinear trend well, with only slight boundary deviations, suggesting that most key features are captured while some minor variance remains.

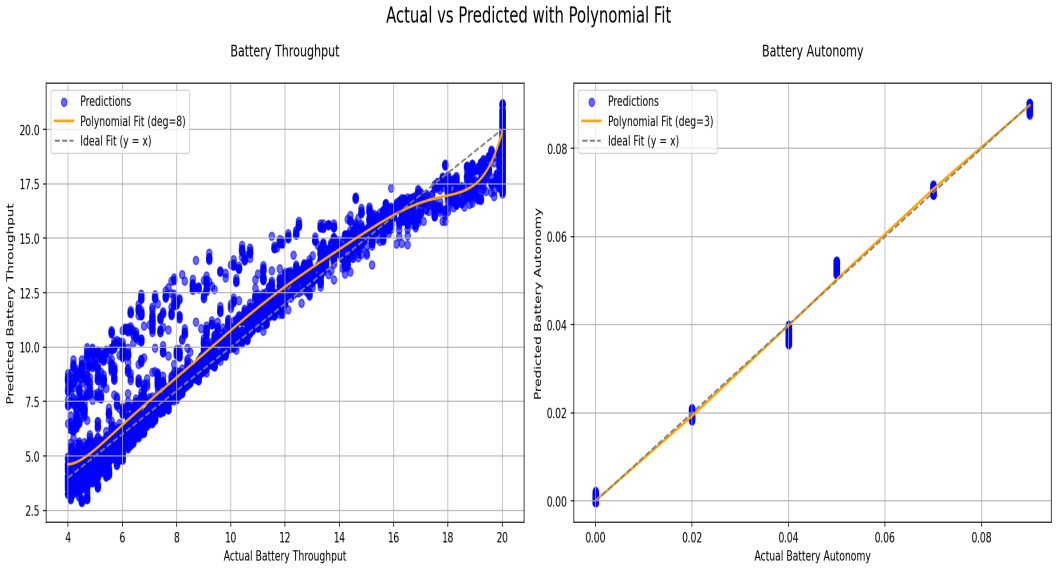

FIGURE 4.23: Scatter Plot for RS with Polynomial Regression

For Battery Autonomy, a degree-3 polynomial fit aligns closely with the ideal line, showing that Ranker Search selects features yielding stable and accurate predictions. Overall, it performs robustly, though nonlinearities in Battery Throughput still need careful handling.

4.9.6 Outliers of RS with Polynomial Regression

Figure 4.24 presents the outlier detection results using polynomial regression residuals under the Ranker Search feature selection method. For Battery Throughput, a degree-8 polynomial fit was applied, and while the majority of data points follow the expected trend, several outliers (highlighted in red) deviate significantly from the fitted curve.These outliers indicate instances where Ranker failed to capture certain nonlinear relationships, leading to localized prediction errors. In contrast,

Battery Autonomy with a degree-3 polynomial fit shows almost no noticeable outliers, with predictions aligning closely to the ideal line.

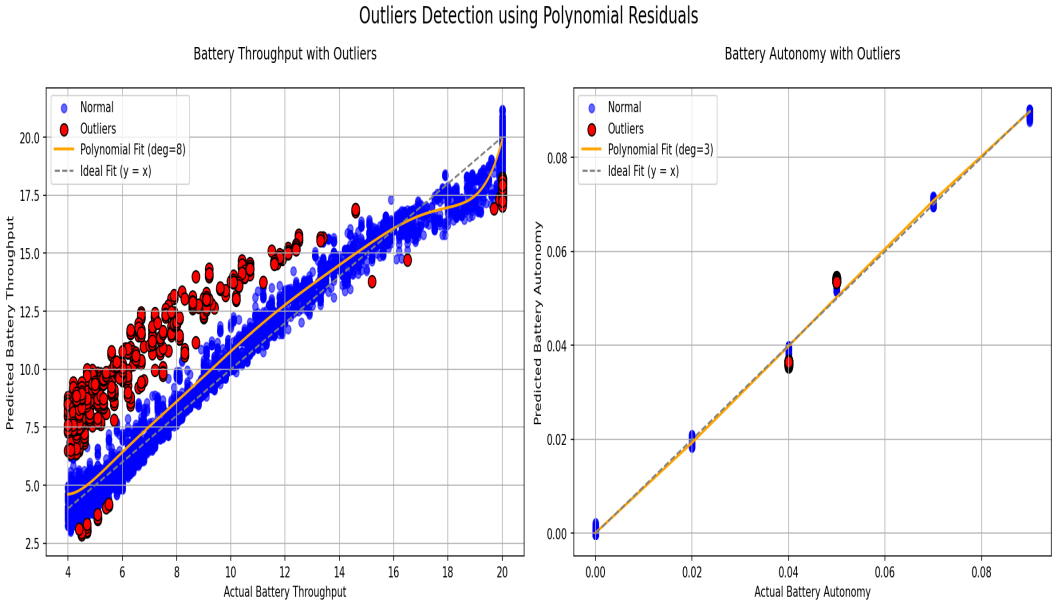

FIGURE 4.24: Scatter Plot for RS with Polynomial Regression

This confirms that Ranker Search provides a highly reliable feature set for stable predictions of Battery Autonomy, while additional refinements may be required to minimize deviations in Battery Throughput.

The threshold for identifying outliers is defined as $\pm 2\sigma$, where σ represents the standard deviation of the residuals (errors). Data points lying above $+2\sigma$ or below -2σ are considered outliers. This criterion is based on the statistical property of the normal distribution, in which approximately 95% of values fall within the range of $\pm 2\sigma$. Therefore, any point outside this boundary is regarded as unusual or extreme. For stricter detection, a threshold of $\pm 3\sigma$ may be used, which covers nearly 99.7% of the data and results in fewer outliers. Conversely, for more sensitive detection, a threshold of $\pm 1.5\sigma$ may be adopted, which covers around 86% of the data and identifies more points as outliers. Outliers were defined as points where the residual error exceeded $\pm 2\sigma$ of the residual distribution. This criterion corresponds to excluding approximately 5% of the data under the assumption of a normal error distribution, thereby identifying predictions that deviate significantly from the polynomial regression fit.

By comparing both models, it is evident that while linear regression provides a simple and interpretable baseline, polynomial regression better captures the nonlinear behavior of battery throughput and autonomy. This highlights the importance of considering nonlinear models when dealing with energy storage systems, where performance often depends on complex interactions among features.

Chapter 5

Conclusion and Future Work

This research presents a comprehensive study on battery sizing and performance prediction based on three key features: Battery Autonomy, Battery Life, and Battery Throughput. Initially, each of these features was independently analyzed using various data visualization techniques, including scatter plots, box plots, and linear regression analysis, to detect patterns and identify outliers. To enhance prediction performance, multiple feature selection techniques were employed, namely Harmony Search (HS), Linear Forward Search (LFS), and Ranker Search (RS). These algorithms were applied individually to each feature and later jointly to predict two target variables simultaneously: Battery Autonomy and Battery Throughput. Support Vector Regression (SVR) with a radial basis function (RBF) kernel was used as the regression model. The models were evaluated using four key performance metrics: Root Mean Squared Error (RMSE), Spearman's Rank Correlation Coefficient (SROCC), Kendall's Tau (KCC), and Pearson's Linear Correlation Coefficient (LCC). Based on the cleaned data performance at Epsilon ($\epsilon = 0.01$), the evaluation of separate battery metrics reveals the following:

- Battery Autonomy: LFS performs best (RMSE = 0.0005, SROCC = 0.9851).
- Battery Life: LFS again leads (RMSE = 105.4754, SROCC = 0.9852).
- Battery Throughput: RS dominates (RMSE = 0.000262, SROCC = 0.9852).

These results indicate that there is no single feature selection method that universally performs best across all three battery metrics. Instead, the effectiveness of each method depends on the specific characteristic being predicted. Overall, Linear Forward Search (LFS) proves most effective for predicting Battery Autonomy and Battery Life, while Random Search (RS) excels in predicting Battery Throughput.

In contrast, when predicting both Battery Autonomy and Battery Throughput jointly using multi-output SVR, Heuristic Search (HS) delivers the most balanced and effective performance across all evaluation metrics (RMSE, SROCC, KCC, LCC). This indicates that HS is more suited for scenarios requiring simultaneous prediction of multiple battery-related outcomes.

The optimal feature selection strategy should be chosen based on the prediction objective,i.e. LFS and RS for isolated targets, and HS for joint target modeling. the comparative analysis demonstrates that while linear regression serves as a useful baseline, it lacks the flexibility to fully capture the nonlinear dynamics present in battery performance prediction. Polynomial regression, on the other hand, provides a more accurate representation of both battery throughput and autonomy, effectively addressing the complex interactions among system parameters. Furthermore, the incorporation of Ranker Search for feature selection significantly enhances model robustness, ensuring that the most influential variables are prioritized while minimizing redundancy. These findings confirm that nonlinear modeling approaches, combined with systematic feature selection, are essential for reliable battery performance prediction and can serve as a foundation for optimizing energy storage systems in practical applications

5.1 Future Work

Future research may focus on the following directions:

- Deep Learning Models: Explore LSTM or Transformer models for improved prediction accuracy over time, building on the current SVR-based battery performance prediction framework.
- Hybrid Feature Selection: Combine heuristic and statistical techniques for enhanced robustness, complementing the current analysis of HS, LFS, and RS methods.
- EV Integration: Apply the methodology to Electric Vehicles for battery sizing and range estimation, adapting the SVR approach for diverse battery technologies.
- Environmental Variables: Incorporate temperature, load patterns, and real-world conditions into the dataset to further improve the SVR model's prediction accuracy for microgrid battery performance.
- Multi-objective Optimization: Use algorithms like NSGA-II to balance multiple performance targets, extending the current single-objective SVR regression to handle trade-offs among battery autonomy, throughput, and life.
- Explainable AI: Apply SHAP or LIME to interpret feature importance and model behavior in the existing SVR-based prediction pipeline, enhancing transparency and trust in feature selection outcomes.

Bibliography

[1] M. Abou Houran, X. Yang, and W. Chen, "Energy management of microgrid in smart building considering air temperature impact," in *2018 IEEE Applied Power Electronics Conference and Exposition (APEC)*. IEEE, 2018, pp. 2398–2404.

[2] A. C. Luna, N. L. Diaz, M. Graells, J. C. Vasquez, and J. M. Guerrero, "Mixed-integer-linear-programming-based energy management system for hybrid pv-wind-battery microgrids: Modeling, design, and experimental verification," *IEEE Transactions on Power Electronics*, vol. 32, no. 4, pp. 2769–2783, 2016.

[3] H. Shahinzadeh, M. Moazzami, S. H. Fathi, and G. B. Gharehpetian, "Optimal sizing and energy management of a grid-connected microgrid using homer software," in *2016 Smart Grids Conference (SGC)*. IEEE, 2016, pp. 1–6.

[4] S. Kanwal, B. Khan, and S. M. Ali, "Machine learning based weighted scheduling scheme for active power control of hybrid microgrid," *International Journal of Electrical Power & Energy Systems*, vol. 125, p. 106461, 2021.

[5] B. Lu and M. Shahidehpour, "Short-term scheduling of battery in a grid-connected pv/battery system," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 1053–1061, 2005.

[6] R. M. Elavarasan, G. Shafiullah, S. Padmanaban, N. M. Kumar, A. Annam, A. M. Vetrichelvan, L. Mihet-Popa, and J. B. Holm-Nielsen, "A comprehensive review on renewable energy development, challenges, and policies of

leading indian states with an international perspective," *Ieee Access*, vol. 8, pp. 74 432–74 457, 2020.

[7] D. C. Momete, "Analysis of the potential of clean energy deployment in the european union," *IEEE Access*, vol. 6, pp. 54 811–54 822, 2018.

[8] H. Khan, I. F. Nizami, S. M. Qaisar, A. Waqar, M. Krichen, and A. T. Almaktoom, "Analyzing optimal battery sizing in microgrids based on the feature selection and machine learning approaches," *Energies*, vol. 15, no. 21, p. 7865, 2022.

[9] M. Abou Houran, X. Yang, and W. Chen, "Energy management of microgrid in smart building considering air temperature impact," in *2018 IEEE Applied Power Electronics Conference and Exposition (APEC)*. IEEE, 2018, pp. 2398–2404.

[10] D. Papadaskalopoulos, D. Pudjianto, and G. Strbac, "Decentralized coordination of microgrids with flexible demand and energy storage," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 4, pp. 1406–1414, 2014.

[11] S. Ganesan, U. Subramaniam, A. A. Ghodke, R. M. Elavarasan, K. Raju, and M. S. Bhaskar, "Investigation on sizing of voltage source for a battery energy storage system in microgrid with renewable energy sources," *IEEE Access*, vol. 8, pp. 188 861–188 874, 2020.

[12] C. Klansupar and S. Chaitusaney, "Optimal sizing of utility-scaled battery with consideration of battery installtion cost and system power generation cost," in *2020 17th international Conference on electrical engineering/electronics, computer, Telecommunications and information technology (ECTI-CON)*. IEEE, 2020, pp. 498–501.

[13] J. Sobon and B. Stephen, "Model-free non-invasive health assessment for battery energy storage assets," *IEEE Access*, vol. 9, pp. 54 579–54 590, 2021.

[14] Y. Wang, Y. Li, L. Jiang, Y. Huang, and Y. Cao, "Pso-based optimization for constant-current charging pattern for li-ion battery," *Chinese Journal of Electrical Engineering*, vol. 5, no. 2, pp. 72–78, 2019.

[15] X. Peng, C. Zhang, Y. Yu, and Y. Zhou, "Battery remaining useful life prediction algorithm based on support vector regression and unscented particle filter," in *2016 IEEE International Conference on Prognostics and Health Management (ICPHM)*. IEEE, 2016, pp. 1–6.

[16] M. Abou Houran, X. Yang, and W. Chen, "Energy management of microgrid in smart building considering air temperature impact," in *2018 IEEE Applied Power Electronics Conference and Exposition (APEC)*. IEEE, 2018, pp. 2398–2404.

[17] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

[18] Y. Masoudi-Sobhanzadeh, H. Motieghader, and A. Masoudi-Nejad, "Feature-select: a software for feature selection based on machine learning approaches," *BMC bioinformatics*, vol. 20, pp. 1–17, 2019.

[19] Y. Li, C.-Y. Chen, and W. W. Wasserman, "Deep feature selection: theory and application to identify enhancers and promoters," *Journal of Computational Biology*, vol. 23, no. 5, pp. 322–336, 2016.

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[22] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Feature selection for high-dimensional data," *Progress in Artificial Intelligence*, vol. 5, pp. 65–75, 2016.

[23] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & electrical engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[24] T. Hasanin, T. M. Khoshgoftaar, J. Leevy, and N. Seliya, "Investigating random undersampling and feature selection on bioinformatics big data," in *2019*

IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, 2019, pp. 346–356.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[26] L. Gao, J. Song, X. Liu, J. Shao, J. Liu, and J. Shao, "Learning in high-dimensional multimedia data: the state of the art," *Multimedia Systems*, vol. 23, pp. 303–313, 2017.

[27] R. E. Neapolitan *et al.*, *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, 2004, vol. 38.

[28] A. Kazemtarghi and A. Mallik, "Techno-economic microgrid design optimization considering fuel procurement cost and battery energy storage system lifetime analysis," *Electric Power Systems Research*, vol. 235, p. 110865, 2024.

[29] A. El Shamy, P. Aduama, and A. Al-Sumaiti, "Chance constrained optimal sizing of a hybrid pv/battery/hydrogen isolated microgrid: A life-cycle analysis," *Energy Conversion and Management*, vol. 332, p. 119707, 2025.

[30] R. Kolluri and J. de Hoog, "Adaptive control using machine learning for distributed storage in microgrids," in *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, 2020.

[31] K. Shivam, J. Tzou, and S. Wu, "A multi-objective predictive energy management strategy for residential grid-connected pv-battery hybrid systems based on machine learning technique," *Energy Conversion and Management*, vol. 237, p. 114103, 2021.

[32] F. Pilati, G. Lelli, A. Regattieri, and M. Gamberi, "Intelligent management of hybrid energy systems for techno-economic performances maximisation," *Energy Conversion and Management*, vol. 224, p. 113329, 2020.

[33] M. Mehrtash, F. Capitanescu, and P. Heiselberg, "An efficient mixed-integer linear programming model for optimal sizing of battery energy storage in smart sustainable buildings," in *2020 IEEE Texas Power and Energy Conference (TPEC)*, 2020.

[34] M. Bagheri-Sanjareh, M. Nazari, and G. Gharehpetian, "A novel and optimal battery sizing procedure based on mg frequency security criterion using coordinated application of bess, led lighting loads, and photovoltaic systems," *IEEE Access*, vol. 8, pp. 95 345–95 359, 2020.

[35] G. Almeida, A. Souza, and P. Ribeiro, "A neural network application for a lithium ion battery pack state-of-charge estimator with enhanced accuracy," *Proceedings*, vol. 58, no. 1, p. 33, 2020.

[36] S. Jayashree and K. Malarvizhi, "Methodologies for optimal sizing of battery energy storage in microgrids: A comprehensive review," in *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 2020.

[37] A. Cano, P. Arévalo, and F. Jurado, "A comparison of sizing methods for a long-term renewable hybrid system. case study: Galapagos islands 2031," *Sustainable Energy & Fuels*, vol. 5, no. 5, pp. 1548–1566, 2021.

[38] J. Kumar, C. Parthasarathy, M. Västi, H. Laaksonen, M. Shafie-Khah, and K. Kauhaniemi, "Sizing and allocation of battery energy storage systems in Åland islands for large-scale integration of renewables and electric ferry charging stations," *Energies*, vol. 13, no. 2, p. 317, 2020.

[39] M. Hannan, M. Faisal, P. J. Ker, R. Begum, Z. Dong, and C. Zhang, "Review of optimal methods and algorithms for sizing energy storage systems to achieve decarbonization in microgrid applications," *Renewable and Sustainable Energy Reviews*, vol. 131, p. 110022, 2020.

[40] K. El-Bidairi, H. Nguyen, T. Mahmoud, S. Jayasinghe, and J. Guerrero, "Optimal sizing of battery energy storage systems for dynamic frequency control in an islanded microgrid: A case study of flinders island, australia," *Energy*, vol. 195, p. 117059, 2020.

[41] B. Ratner, W. Wang, and Y. Lu, "Analysis of the mean absolute error (mae) and the root mean square error (rmse) in assessing rounding model," *IOP Conference Series: Materials Science and Engineering*, p. 5, 2018.

[42] T. Gao and W. Lu, "Machine learning toward advanced energy storage devices and systems," *iScience*, vol. 24, no. 1, p. 101936, 2021.

[43] P. Boonluk, A. Siritaratiwat, P. Fuangfoo, and S. Khunkitti, "Optimal siting and sizing of battery energy storage systems for distribution network of distribution system operators," *Batteries*, vol. 6, no. 4, p. 56, 2020.

[44] O. Talent and H. Du, "Optimal sizing and energy scheduling of photovoltaic-battery systems under different tariff structures," *Renewable Energy*, vol. 129, pp. 513–526, 2018.

[45] P. Prabpal, Y. Kongjeen, and K. Bhumkittipich, "Optimal battery energy storage system based on var control strategies using particle swarm optimization for power distribution system," *Symmetry*, vol. 13, no. 9, p. 1692, 2021.

[46] Y. Gupta, R. Vaidya, H. K. Nunna, S. Kamalasadan, and S. Doolla, "Optimal pv – battery sizing for residential and commercial loads considering grid outages," in *2020 IEEE International Conference on Power Electronics, Smart Grid and Renewable Energy (PESGRE2020)*, 2020.

[47] N. Lazaar, E. Fakhri, M. Barakat, J. Sabor, and H. Gualous, "A genetic algorithm based optimal sizing strategy for pv/battery/hydrogen hybrid system," in *Artificial Intelligence and Industrial Applications*, 2020, pp. 247–259.

[48] S. Xie, Q. Zhang, X. Hu, Y. Liu, and X. Lin, "Battery sizing for plug-in hybrid electric buses considering variable route lengths," *Energy*, vol. 226, 2021.

[49] P. Mirhoseini and N. Ghaffarzadeh, "Economic battery sizing and power dispatch in a grid-connected charging station using convex method," *Journal of Energy Storage*, vol. 31, p. 101651, 2020.

[50] J. L. Sampietro, V. Puig, and R. Costa-Castelló, "Optimal sizing of storage elements for a vehicle based on fuel cells, supercapacitors, and batteries," *Energies*, vol. 12, no. 5, p. 925, 2019.

[51] A. Y. Ali, A. Basit, T. Ahmad, A. Qamar, and J. Iqbal, "Optimizing coordinated control of distributed energy storage system in microgrid to improve battery life," *Computers & Electrical Engineering*, vol. 86, p. 106741, 2020.

[52] J. Li, "Optimal sizing of grid-connected photovoltaic battery systems for residential houses in australia," *Renewable Energy*, vol. 136, pp. 1245–1254, 2019.

[53] M. Sufyan, N. A. Rahim, M. M. Aman, C. K. Tan, and S. R. S. Raihan, "Sizing and applications of battery energy storage technologies in smart grid system: A review," *Journal of Renewable and Sustainable Energy*, vol. 11, no. 1, p. 014105, 2019.

[54] Y. Liu, X. Wu, J. Du, Z. Song, and G. Wu, "Optimal sizing of a wind-energy storage system considering battery life," *Renewable Energy*, vol. 147, pp. 2470–2483, 2020.

[55] C. D. Rodríguez-Gallegos *et al.*, "A siting and sizing optimization approach for pv battery–diesel hybrid systems," *IEEE Transactions on Industry Applications*, vol. 54, no. 3, pp. 2637–2645, 2018.

[56] B. Li, X. Wang, S. Liu, S. Li, and T. Xu, "Probability for conditional noncoherency of microgrid due to transmission overloading," pp. 6–6, 2015.

[57] N. V. Otten, "Support vector regression (svr) simplified & how to tutorial," Spot Intelligence, May 2024, accessed: 2025-07-01. [Online]. Available: https://spotintelligence.com/2024/05/08/support-vector-regression-svr/

[58] Lund Research Ltd, "Spearman's rank-order correlation," Laerd Statistics statistical guide, 2025, accessed: 2025-07-01. [Online]. Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

[59] P. Saul McLeod, "Correlation: Meaning, types, examples & coefficient," Simply Psychology, Jul. 2023, updated July 31, 2023; Accessed July 1, 2025. [Online]. Available: https://www.simplypsychology.org/correlation.html

[60] Great Learning, "Rmse – what does it mean?" Medium, 2024, accessed: 2025-07-01. [Online]. Available: https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e