School of Technology

Muhammad Abuzar (x3316580)

ENERGY CONSUMPTION PREDICTION FOR ROBOTIC TASK EXECUTION USING MACHINE LEARNING AND HIGH-LEVEL OPERATIONAL DATASETS

Master's Thesis
Degree Programme in Robotics
Supervisor: Timo Mantere
Co Supervisor: Adeel Arif

Vaasa 2026

Author:	Muhammad Abuzar
Title of thesis:	Energy Consumption Prediction for Robotic Task Execution Using Machine Learning and High-Level Operational Datasets
Degree:	Master of Engineering
Degree Programme:	Robotics
Supervisor:	Timo Mantere
Co-Supervisor	Adeel Arif
Year:	2026

Abstract

This study examines whether energy consumption can be predicted using machine learning methods on high-level operation data sets to execute robotic tasks. With the growing integration of robotic systems in industrial and service settings, there is a strong need to enhance their energy efficiency to cut down on operation costs and to promote the cause of sustainable development. A quantitative approach was used on a designed dataset of 500 observations of robotic tasks and included processing time, task, sensor, environmental and operational status indicators feature. Four predictive models were created and assessed a mean-based baseline model, a linear regression model and a random forest regressor. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used to evaluate model performance on a test set held out. These findings show that the three models had widely similar predictive accuracy, with the lowest MAE (0.5385 kWh) and the lowest RMSE (0.6152 kWh) values of the random forest and the baseline models respectively. Feature importance analysis of the random forest model assigned the highest variance reduction scores to processing time and accuracy, at approximately 0.284 and 0.272 respectively. These scores reflect an algorithmic bias of the random forest toward continuous variables, which accumulate importance through a greater number of candidates split points, rather than confirmed causal influence on energy consumption. All other features contributed at much lower levels. The results indicate that machine learning models do not make significant improvements compared to a basic baseline prediction when they are trained using high-level operational data. This result implies that the dataset has not enough informative predictors to enable precise energy modeling, probably due to the failure to capture the low-level physical processes that directly determine energy consumption in robotic systems with high-level task descriptors. The paper presents a practical and replicable assessment system of data-driven energy forecasting in robotics, the main features of the datasets that constrain the performance of machine learning, and the significance of more detailed, richer data to achieve successful energy prediction in robotic tasks.

Keywords: energy consumption, robotic systems, machine learning, random forest, linear regression, feature importance, industrial automation, green energy technology.

Table of Contents

Abbreviations

AI	Artificial Intelligence
CMP	Component Identifier
CSV	Comma-Separated Values
IoT	Internet of Things
kWh	Kilowatt-hour
LIDAR	Light Detection and Ranging
MAE	Mean Absolute Error
ML	Machine Learning
MOPSO	Multi-Objective Particle Swarm Optimization
MSE	Mean Squared Error
OLS	Ordinary Least Squares
RBT	Robot Identifier
RF	Random Forest
RMSE	Root Mean Squared Error
RSM	Response Surface Methodology
SD	Standard Deviation

Tables and Figures

Chapter 1: Introduction

1.1 Background

The general use of robotic systems in industries and services has redefined the modern manufacturing and automated processes. In the last twenty years, robots have developed out of simple and task-specific machines towards complex and multi-functional systems that are able to carry out complex tasks with a high degree of precision and repeatability. Robotic platforms have been implemented at scale in industries like automotive manufacturing, electronics assembly, pharmaceutical production, and logistics due to the pressures of increased throughput, lower labor expenses, and quality consistency. The International Federation of Robotics reported that the number of operational industrial robots worldwide exceeded three million units in the early 2020s, of which robots have become a permanent part of manufacturing infrastructure.

Although the operational advantages that robotic systems are able to bring are considerable, the energy consumption of such systems has become a pressing issue that is becoming more and more critical. Industrial robots are designed to work using a mix of electric actuators, servo motors, motion controllers, embedded sensors and communication systems, with each adding to the total energy of the platform. A single robotic arm can take one to five kilowatt-hours per working hour in high-volume production facilities based on the type of work done, the weight carried, and the direction taken. Once scaled to large-scale deployments of dozens or hundreds of robots running through multiple shifts, the total energy use can be a significant portion of facility operating costs and environmental impact.

The advent of sustainable engineering concepts into the larger context of green energy technology has put increasing pressure on the need to design and operate robotic systems in an energy conscious fashion. This has been recognised in modern engineering curricula and industrial research agendas, and the minimisation of energy usage in automated systems is now perceived to be a fundamental value of environmentally friendly industrial practice, as well as a cost-optimisation exercise. In this respect, the ability to accurately measure, track, and estimate

the energy consumption during the robot operations has transformed itself into a research problem of high practical value.

At the same time, advances in data acquisition systems, computing power and machine learning algorithms have opened new opportunities in developing novel approaches to energy monitoring and prediction based on data. Instead of using only physics-based models that entail the intensive familiarity with the parameters of mechanical and electrical systems, data-driven methods provide a more open and generalizable path to energy estimation. During the execution of tasks, operational data collected from the tasks can be used in predictive models to derive the statistical patterns and association to inform energy conscious planning and resource management decisions. This thesis falls in this wider context and poses the question of whether high-level operational data, such as task descriptors, sensor data and environmental metrics, can serve as a viable basis for machine learning models to predict energy usage in robotic systems.

1.2 Motivation

This research is motivated by the fact that lots of people around the world are interested in developing smart systems that can be used to control and predict the energy consumption during robotic activities without having to have complex instrumentation and low-level access to hardware. As more and more robots are being used in industrial premises, energy management has become a more complex problem, which needs to be coordinated across multiple platforms, task schedules and operational restrictions. Accurate energy estimates can support a variety of practical applications, including energy planning, calculating the energy cost of production processes, optimizing processes and reporting environmental effects.

The traditional energy modelling methods for robotic systems are mainly physics-based, derived from a dynamic model, which is based on the equation of motion of the robot manipulator. While these models can be extremely accurate if they are parameterized correctly, they also require thorough knowledge of motor efficiency curves, gear ratios, inertial parameters and friction characteristics which may not necessarily be found for every robot platform, particularly if they are sourced from an online store. Furthermore, these models are robot-specific by design and

would have to be re-derived or re-calibrated for each new platform limiting scalability, especially in a heterogeneous environment.

Data-driven approaches provide a possibly more flexible, scalable, solution, where learning of energy behaviour is done based on observational data, without a detailed mechanical description. In particular, machine learning algorithms are being made more and more available in open-source software format and are applicable to a broad spectrum of data types and data structures. The effectiveness of such methods is however, critically dependent on the informativeness of the input features. The basic question that leads to this study is whether high level operational data, including the nature of task that the sensor performs, the sensor configuration that is utilized, the processing time that is necessary and the environmental conditions that the sensor is exposed to can give enough predictive signal to machine learning models to predict energy consumption with significant accuracy.

Moreover, this study is driven by systematic, reproducible empirical analyses of data-driven methods within realistic data limitations. The literature in machine learning on robotic energy prediction is somewhat small in comparison to related areas like building energy management or transportation, and there is a marked gap in literature specifically analysing the implications of simplified, high-level data used to predict. With this kind of assessment, this thesis will serve to offer a practical advice to both researchers and practitioners who have to make decisions on the data collection methods and the choice of models in energy-aware robots.

1.3 Problem Statement

The energy consumption of robotic systems depends on a complicated combination of many factors, such as the type of activity underway, the movement patterns and velocities necessary, the load to be handled by the system, the environmental factors under which the robot is working, and the properties of the control and sensing systems used. It is not easy to predict energy usage precisely due to the non-linear and unattainable linearization of the relationships between these variables, as well as the varying relative significance of each factor in different operational settings. Current energy prediction systems of robotic systems can be divided into physics-based and data-driven systems. Physics models are based on energy estimates using the

dynamic and kinematic equations of motion of robots and need detailed system parameters that are not always attainable or feasible to access. In contrast to this, data-driven models utilize historical operational data to learn predictive relationships, but the models are critically reliant upon availability of features that are actually informative about energy consumption. High-level operational data is easily available in most practical situations. Measures of low-level sensors like motor currents, joint torques, and component-level power draw may not be logged by default or need special hardware and software settings to record. This leaves a certain gap: it is not clear whether machine learning models trained solely on high-level task-descriptors can make reliable predictions of energy, or whether the unavailability of granular physical data is a fundamental limitation to the predictive accuracy that can be achieved. This gap is filled by this thesis using a systematic empirical investigation.

1.4 Research Objectives

The main aim of this thesis is to create and test machine learning algorithms to predict energy usage in the execution of robotic tasks, with high-level operation data as the only input. The following specific objectives are used to guide the study:

(i) To conduct a thorough exploratory analysis of a structured robotic operational dataset containing energy consumption measurements alongside task and environmental parameters.

(ii) To preprocess and prepare the dataset for machine learning applications, including the handling of categorical variables through one-hot encoding and the verification of data quality.

(iii) To implement four predictive models of increasing complexity: a mean-based baseline, a linear regression model, and a random forest regressor.

(iv) To evaluate and compare the performance of these models on a held-out test set using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

(v) To identify the most influential features in predicting energy consumption through feature importance analysis of the random forest model.

(vi) To assess the practical limitations of using high-level operational data for energy prediction and to draw evidence-based conclusions regarding its utility.

1.5 Research Questions

This study is guided by three primary research questions, each targeting a distinct aspect of the data-driven energy prediction problem:

RQ1: Which operational parameters have the most significant influence on robotic energy consumption, and can this influence be reliably identified through feature importance analysis?

RQ2: How accurately can machine learning models trained on high-level operational data predict energy consumption in robotic task execution, and do they provide a meaningful improvement over a simple baseline prediction?

RQ3: How do linear regression and random forest models compare in terms of predictive performance, interpretability, and practical utility when applied to a high-level robotic operational dataset?

These research questions are addressed through the experimental framework described in Chapter 3, with results presented in Chapter 4 and interpreted in Chapter 5.

1.6 Scope of the Study

As of now, the methodology that is used to conduct research follows this process; research is performed by analysing and predictive modelling of the energy consumption from an existing operational dataset obtained from the robots at a task level and consists of 500 task records. The research does not depend upon the real-time robotic experiments, hardware interaction and data acquisition of operational robotic platforms. All software tools are implemented in Python programming language and scikit-learn machine learning library for all the analysis and modeling.

The models that are explored are limited to a baseline statistical model, a linear regression model and a random forest regressor model. This paper does not cover more advanced models like gradient boosting algorithms, support vector machines or deep learning models, but they are

mentioned as areas for future research. No research has been found discussing real-time energy optimization, model deployment in running robotic systems and physical energy monitoring infrastructure design.

All of this work focuses on assessing how much a high-level operational data is being used to predict energy, and how to interpret the results relative to the nature of the data used and the existing body of literature on data-driven energy prediction.

1.7 Structure of the Thesis

This thesis is divided into six chapters with a specific purpose in the general framework of the research. Chapter 1 has presented background, motivation, problem statement, research objectives, research questions and scope of the study.

Chapter 2 is a literature review of the literature that is most pertinent to the research objectives. It has been reviewed on energy consumption of industrial robots, classic and data-driven modeling methods, machine learning methods used in manufacturing and robotics, theoretical basis of linear regression and random forests models, feature selection and importance analysis, sensors in robotic work, and the sustainability issue of energy-conscious robotics.

Chapter 3 explains the research methodology in detail, such as the nature of the dataset, the steps taken when preprocessing data, how the predictive models were designed and implemented, the train-test split approach, and the evaluation metrics used.

Chapter 4 summarizes the quantitative findings of the experimental analysis, such as the findings of the exploratory data analysis, correlation analysis, the comparison of the model performance, the ranking of feature importance, and a visual evaluation of the prediction quality using the scatter plot analysis.

Chapter 5 gives a detailed discussion of the results, the interpretation of model performance in terms of the characteristics of datasets, the comparison of the results with the previous literature, and the most important limitations of the research.

The final chapter (Chapter 6) summarizes the main findings of the thesis, gives the contribution that the study made to the field, outlines the limitations of the research, as well as offers certain recommendations on future research that might have helped to improve the shortcomings that have been identified.

Chapter 2: Literature Review

2.1 Introduction

The prediction of energy use in robotic systems is a unique field of research that brings together work in robotics engineering, energy management and applied machine learning research. Industrial automation, which is increasingly gaining in momentum due to Industry 4.0 and policy issues with sustainability, has made the need to learn, monitor and reduce the energy consumption in robot operation very pronounced. The range of scientific resources addressing this challenge is extremely broad, and is represented by mechanical engineering, control systems, data science, and environmental management.

This chapter presents a critical and organized review of the research that is most pertinent to the aim of this thesis. The review is divided into ten thematic sections starting with the industrial context of robotic energy consumption and moving to the context of traditional methods of modeling, data-driven and machine learning methods, the algorithms used in this research, feature selection methods, sensor technologies, and the sustainability context and finally identifying the research gap. The reviewed studies are represented each with reference to their findings and contributions in specific order according to the APA citation conventions.

2.2 Energy Consumption in Industrial Robotic Systems

Paryanto et al. (2015) undertook an empirical study of energy consumption patterns of industrial robots in the manufacturing facilities under specific conditions, with particular attention to the operational parameters that have the greatest impact on power draw. Their research investigated the influence of variables such as joint velocity, acceleration profiles, payload weight and cycle time on the energy usage of articulated robotic arms deployed in the automotive production environments. Paryanto et al. (2015) discovered that the most important factors that determine the amount of energy consumed are the motion related parameters, especially the highest values of acceleration that are defined in the motion program of the robot. They showed that using a trajectory profile reduction to achieve energy savings, that are measurably in the ten to twenty percent range, could be done without correspondingly impairing cycle time

performance. The paper also found that stationary time, during which the robots are powered but idle until upstream work is finished, can also make a significant proportion of overall energy usage in the normal production processes, indicating that workflow scheduling is a significant lever to energy savings.

Carabin et al. (2017) presented an in-depth review of energy-saving optimization techniques that could be applied to robotic and automatic systems and divided the existing practices into four main spheres. The former includes trajectory optimization techniques, which aim at minimizing energy expenditure by altering the trajectory or velocity field of robot motion, without altering the initial and final positions of the robot. The second addresses the mechanical design considerations, such as optimization of link masses, actuator size and integration of counterbalancing mechanisms. The third is the drive and electrical system enhancement, including regenerative braking to reclaim energy during deceleration cycles. The fourth is control-based techniques where the operating conditions of the time vary dynamically by changing operational parameters. Carabin et al. (2017) concluded that trajectory optimization is the most researched and practically applicable method, and pointed out that hybrid methods that integrate the various types of optimization options are expected to produce more energy savings than single-domain methods. One of the most thorough treatments of the subject in the recent literature, their review offers a theoretical framework in which the data-driven approach of the present study can be placed.

Cai et al. (2017) analyzed the energy management perspective of manufacturing systems, which introduced a multi-objective energy benchmarking method that aims to determine energy-efficient reference operating points of mechanical manufacturing processes. Their approach did not emphasize single machines and robots, but rather the overall energy behavior of manufacturing lines and employed benchmarking targets to point out where real operations were not operating in a way that was energy efficient. Cai et al. (2017) used their framework on actual manufacturing data and discovered that process type, machine utilization rate and operating speed are some of the most significant variables that distinguish energy-efficient and energy-intensive operations. Their work defined the principle of a task category and a working environment as important dimensions within which the energy behavior differs, which directly

informs the choice of task type and environmental status as attributes in the dataset utilized in the current thesis.

Fang et al. (2019) focused on the energy consideration of trajectory planning of industrial robots, developing smooth S-curve trajectory methods, which trade time efficiency and mechanical smoothness. Their experiment showed that standard trapezoidal velocity profiles commonly employed in industrial motion controllers because of their computational simplicity produce large amounts of acceleration jerk causing temporary peak power loads on servo motors. The S-curve model suggested by Fang et al. (2019) generates curves with continuous, limited rate-of-change of acceleration, and minimises the maximum motor current needs and helps to reduce total energy usage. Their experimental estimation on a six-degree-of-freedom industrial robot arm established that the smooth trajectory planning can save energy in contrast to trapezoidal profiles with a similar cycle time, which supports the relation between motion dynamics and energy consumption that is the basis of the choice of processing time as a characteristic in the current investigation.

Qin et al. (2016) suggested a categorical map of intelligent manufacturing systems in the Industry 4.0 environment, classifying operations by how automated they are, connected to data, and smart decisions. The authors suggested that the ongoing, gradual deployment of embedded sensors, networked controllers and real-time analytics into manufacturing settings have a fundamental impact on the opportunities to monitor and manage energy. Qin et al. (2016) singled out data-driven energy monitoring as one of the essential enablers of sustainable manufacturing and the necessity of creating standardized data collection practices that would enable cross-platform comparisons. Their framework is applicable to the current research since it puts into context the application of operational datasets to forecasting energy under the wider industrial transformation with Industry 4.0, and since they contribute to the importance of capturing various aspects of operational variables, such as type of task and status of the environment, in energy monitoring systems.

Lasi et al. (2014) outlined the conceptual background of Industry 4.0, introducing the implementation of cyber-physical systems, the Internet of Things, and autonomous decision-

making to manufacturing as a fourth industrial revolution. Even though their article is older than much of the existing robotics energy literature, it gave the theoretical terms and analytical language that other researchers, such as Qin et al. (2016) and Zhong et al. (2017), used to establish position data-driven manufacturing research. Lasi et al. (2014) opined that the integration of physical production systems into digital information and communication infrastructure presents new opportunities as well as new challenges to energy management. In particular, with the growing amount and range of operational data present in interconnected manufacturing facilities, the patterns concerning the energy use may potentially be discovered and used as predictive data, as long as the relevant analytical tools are utilized.

2.3 Traditional Physics-Based Energy Modeling

Classical modeling of energy consumption of robotic systems is based on the Lagrangian formulation or Newton-Euler formulation of robot dynamics. These frameworks represent the robot as a system of rigid links by actuated joints, and energy consumption is estimated by the joint torques and angular velocities of the joints as the task is performed. The accuracy of such models is tied to the access to accurate values of the inertial parameters, gear efficiencies, motor constants and friction coefficients, most of which need to be identified by delicate experimental means (Carabin et al., 2017).

Models based on physics have shown to be quite accurate in well-controlled laboratory conditions where knowledge of all parameters in the system is known and operating conditions are well-defined. Paryanto et al. (2015) observed, nevertheless, that the practical use of the models in industrial settings is complicated by the fact that the accurate values of the parameters of commercial robots are not easy to obtain, with many of them not publishing detailed specifications of their mechanisms. Moreover, the accuracy of dynamic models reduces with a change in operating conditions not specified in the nominal values used in the parameterization, including changes in the weights of the payload or wear over time on components.

The physics-based models, which inherently have robot-specificity, is another weakness in the context of modern manufacturing, where facilities are increasingly run with heterogeneous fleets of robots produced by different manufacturers, and with different kinematic configurations. The

physics-based methods are ill adapted to scalable energy management in a wide range of robotic deployments because each new platform will need a separate modeling effort. Such constraints have driven the investigation of data-based alternatives which can be potentially implemented with less knowledge about the system.

2.4 Data-Driven and Machine Learning Approaches

Wuest et al. (2016) reviewed the application of machine learning in manufacturing and found that the supervised, unsupervised, and reinforcement learning approaches had been used in a variety of industrial tasks such as quality prediction, predictive maintenance, process optimization, and energy management. The authors have reviewed a rich literature and found the quality of data and feature choice to be the two most frequently mentioned criteria in the determination of model performance. Wuest et al. (2016) discovered that most common methods of supervised learning, especially regression-based models like support vector machines, neural networks, and decision tree ensembles, have been used most frequently on predictive tasks in manufacturing. They concluded that machine learning is much more flexible than the more classic statistical models, but its benefits are strongest when the training data has informative features that represent large changes in the outcome of interest. On tasks where the available data are too coarse or too large to optimize machine learning models can fail to outperform more basic statistical benchmarks, a result of immediate interest to the current paper.

Jordan and Mitchell (2015) covered the bigger picture of machine learning research, outlining its evolution since the symbolic rule-based models of the 1980s, through statistical learning techniques, to the deep-learning models that dominated many benchmark tasks in the 2010s. The authors claimed that the recent achievements of machine learning can be explained by the size of large, labeled datasets and drastic improvements in computing capabilities, and that the discipline is shifting towards less automated and more adaptive learning systems. Jordan and Mitchell (2015) also focused on the issue of model interpretability, as more complex models are created, the more difficult it is to interpret and explain their predictions. In the field of engineering, such as energy management of robotic systems, interpretability is not just an

academic issue, but also a practical one: such models need to be interpretable and trustworthy by engineers prior to their implementation in a practical environment.

LeCun et al. (2015) surveyed the history and conceptual basis of deep learning and how multilayered neural networks that are trained with backpropagation can learn hierarchical representations of multidimensional data with no explicit feature engineering. They surveyed applications in image classification, speech recognition, natural language processing, and drug discovery, making deep learning a transformative methodology that can be widely applied. LeCun et al. (2015) stressed that deep learning models are highly sensitive to the large size of training datasets because these models generally need thousands or even millions of labeled examples to learn useful representations.

Mocanu et al. (2016) implemented machine learning to energy prediction in smart grid buildings, showing that unsupervised learning of features and transfer learning methods could enhance prediction accuracy in case of limited labeled data. Their experiment trained a limited Boltzmann machine to learn compact models of their energy consumption patterns in a set of buildings and demonstrated that models trained on data in one building can be fine-tuned to predict other buildings with only a few more training examples. Mocanu et al. (2016) observed that knowledge sharing between related domains could balance the few local data points, even in low-data settings, which implies that this transfer learning method is better than models that are trained in isolation. Although they are not applied to the same field as industrial robotics, the overall principle, that data-driven energy predicting is more successful when the training data is rich and informative.

Zhong et al. (2017) performed a thorough literature review on the topic of intelligent manufacturing within the industry 4.0 framework, exploring how incorporation of cyber-physical systems, real-time data analytics, cloud computing, and advanced sensing technologies can help to make manufacturing processes more efficient and responsive. The authors conducted a review of both supervised and unsupervised learning applications to manufacturing intelligence tasks and chose the energy optimization as one of the priority application domains where data-driven approaches should be able to provide significant value. Zhong et al. (2017) discovered that

systems that integrate IoT sensor data with machine learning analytics platforms can provide finer-grained monitoring of energy consumption than before could be done with periodic manual measurements.

2.5 Linear Regression for Energy Prediction

The analysis of linear regression has been presented as a rigorously methodological exercise in a widely-used textbook on the topic Montgomery et al. (2021). Discussing common least squares estimation, multiple regression, hypothesis testing, model diagnostics, and its use in engineering settings, the authors defined that linear regression is a strong and robust technique in the case when the conditions of linearity, homoscedasticity, and fluently distributed errors are roughly met. According to Montgomery et al. (2021), linear regression models are especially appreciated because of their interpretability since the estimated coefficients are a direct measure of the marginal contribution of each predictor variable to the response, with other variables kept constant. This interpretability is a major benefit in engineering uses where interpretation of the relationship between variables is nearly as much of a concern as accuracy in prediction.

Linear regression has been used in energy-related applications in a variety of fields such as building energy prediction, transportation fuel modeling, and estimating the efficiency of industrial processes. Its popularity in these environments is based on its simplicity, minimal computational needs and the simplicity of its predictions. Nonetheless, there are also some limitations to linear regression in the situations when the correlation between the predictors and the response is not linear or there is a large effect of interaction. Carabin et al. (2017) observed that the energy usage of robotic systems is sensitive to the interplay between several variables in its operation, which can be expected to be non-linear, implying that even highly informative features can be underfitted by a linear model. These restrictions encourage the application of ensemble techniques like random forests as an add-on to the linear regression in the model assessment performed in the current study.

2.6 Random Forest Models: Theoretical Foundations and Applications

Biau and Scornet (2016) conducted a stringent theoretical study of random forests, laying down the fundamental findings on the reduction of the variance of the ensemble averaging, and understanding the processes through which the reduction of the variance is attained without correspondingly enhancing the bias. Their excursion into the random forest algorithm explained in detailed mathematical language how each of the constituent trees is trained on a bootstrap-resampled training data with a randomly selected subset of the features being considered at each splitting node and how the summation of the prediction of the forest approaches the true regression function given regularity conditions. Biau and Scornet (2016) have shown that the variance of the random forest estimator is monotonically decreasing with the number of trees in the ensemble, and that the bootstrap sampling and random feature selection mechanisms are interacting to decorrelate single trees to increase the variance reduction obtained by averaging. This theoretical framework forms random forests as a powerful and trustworthy prediction algorithm, especially in an environment with moderate data volumes, combined feature types, and possibly complicated non-linear interrelations among predictors and the response.

Probst et al. (2019) carried out a large-scale empirical study of how hyperparameter decisions affect the performance of random forests, employing a large set of datasets across different domains to test how sensitive predictive accuracy is to the number of trees, the features to consider when splitting, the minimum node size, and the sampling fraction. They have analyzed and discovered that the hyperparameters that have the highest impact on performance are the number of candidates features per split and the minimum size of the terminal node, and the number of trees is the main factor that influences the stability of predictions, but not the accuracy. Probst et al. (2019) established that the default values of hyperparameters used in popular implementations, such as scikit-learn in Python, offer decent performance on the majority of datasets, but specialized tuning can produce further performance gains. They also discovered that random search through the hyperparameter space is much more efficient than exhaustive grid search and suggested its application as the main tuning method to be used by practitioners. The results of Probst et al. (2019) influenced the choice of the current study to apply the default scikit-learn version of the random forest regressor, which integrates the

appropriately calibrated default settings that would be applicable to moderately sized and complex datasets.

Gradient boosting is a sequential ensemble learning technique in which trees are fitted iteratively, with each successive tree trained to predict the residual errors of the combined ensemble up to that point (Friedman, 2001). Unlike random forests, which build trees in parallel and combine them by averaging, gradient boosting builds trees sequentially and updates the model prediction by adding a scaled version of each new tree's output. This sequential correction process allows it to capture complex patterns that may be missed by parallel ensemble methods, but also makes gradient boosting more sensitive to the hyperparameters and more prone to overfitting if regularisation is not sufficient (Chen and Guestrin, 2016). The gradient boosting regressor implemented in scikit-learn uses gradient descent in function space and supports several configurable parameters including the learning rate, the number of boosting stages, and the maximum depth of individual trees. In the present study, default settings were used to provide a fair comparison across all models.

The study of Gregorutti et al. (2017) concentrated on the correlation between feature correlation and variable importance scores in random forests, which directly relates to understanding the importance of feature importance analysis in empirical research. Their analysis showed analytically and empirically that in cases where two or more predictor variables are strongly correlated, the random forest model will apportion scores of importance between these variables in a manner that might not be a true representation of the individual predictive contribution of each. In particular, the significance of a single correlated variable can be arbitraryised by the model choosing between correlated alternatives where there is a split node and thus the model can cause misleading ranks. Gregorutti et al. (2017) suggested conditional importance measures that adjust for inter-feature correlations and recommended that practitioners exercise caution when relying on raw importance scores derived from mean decreases in variance, as these scores reflect the algorithm's internal distribution of variance reduction across splits and do not necessarily indicate the true predictive contribution or causal relevance of individual features.

Lipton (2018) studied the idea of model interpretability in machine learning, making a critical distinction between transparency, which is a way of knowing the internal processes of a model, and post-hoc interpretability, which is a way of explaining the predictions after the fact. He claimed that the scores of feature importance calculated by random forests and related methods of ensemble analysis are a small and misleading kind of interpretability in that they indicate the statistical relationship among the features and the accuracy of prediction, but not the functional form of that relationship, or its causality. Lipton (2018) warned against the possibility of confusing statistical importance with substantive importance, which can result in false inferences in engineering and policy use, where the causes of a relationship are as important as its existence.

2.7 Feature Selection and Variable Importance

Cai et al. (2018) undertook a review of feature selection techniques in machine learning in a holistic manner, considering filter, wrapper, and embedded techniques within the context of various application areas such as bioinformatics, text classification, and engineering diagnostics. They found that embedded methods, such as random forest and gradient boosting feature importance, can be competitive with more complex wrapper methods, while maintaining an efficient computational penalty, particularly in the case of high dimensional data. Cai et al. (2018) have shown that the selection of features is especially important when the input data set contains redundant or irrelevant features, which might reduce the model generalization and make the results interpretation more difficult. The methodological justification of the use of feature importance analysis as an important part of the experimental assessment in the current research, as well as the use of the random forest model as the foundation of this analysis due to its efficiency and the ability to interpret its importance scores, were based on their findings.

The process of feature selection is a significant part of any data-driven modeling project because irrelevant or redundant features should not be included to enhance the predictive accuracy of a model, which is referred to as the curse of dimensionality (Jordan and Mitchell, 2015). When it comes to robotic energy prediction, the most pertinent features are the ones most closely related to the physical processes which could influence energy consumption, including the duration of motion, complexity of the task, and the operating environment. By definition, the high-level

operational data can be characterized by features that are weakly or indirectly connected to these physical processes, and feature selection and importance analysis can be particularly useful in learning which aspects of the data at hand are most important to the prediction task.

2.8 Sensor Technologies in Robotic Operations

Raj et al. (2020) provided an extensive overview of LiDAR scanning mechanisms in robots and autonomous vehicles, including the technical principles, performance characteristics, and energy consumption profile of rotating mechanical, solid-state, and MEMS-based LiDAR systems. They discovered that LiDAR sensors can be significantly different in their energy uses based on which scanning method is used, and their scanning range, with rotating LiDAR systems generally using more energy than solid-state counterparts because they need to use mechanical energy to move spinning mirror systems. It was observed that LiDAR sensors are becoming more frequently used in conjunction with other sensor modalities, especially cameras and thermal sensors, to facilitate tasks that need both spatial knowledge and appearance-based identification (Raj et al., 2020). The discovery can be directly applied to the current research where observations of robotic systems with LIDAR, Camera, Thermal, and LIDAR+Camera sensor configurations are used as the dataset. The feature sensor type in the dataset is an indicator of the expectation that other sensor modalities can have varying energy requirements on the robotic system.

Thermal sensors are of special significance in industrial robotics that involve welding, casting, and surface treatment as temperature monitoring is a quality requirement. Sensors based on cameras are common in assembly and inspection processes to align visually, detect defects, and identify components. LiDAR with cameras, also known as sensor fusion, offers complementary spatial and visual data, which allows more trustworthy navigation and object detection in intricate settings. The power requirements of a given sensor or a combination of sensors is an aspect of interest in the energy-aware robotic systems since the power consumption of the sensors makes up a proportion of the total power consumption of the platform. The data under analysis in this thesis gives a possibility to investigate the hypothesis of whether sensor type is a statistically significant predictor of energy consumption at the task level.

2.9 Sustainability and Energy-Aware Robotics

The necessity of energy efficiency of robotic systems is placed in the context of industrial sustainability and the overall move to low-carbon energy systems worldwide. Cai et al. (2017) pointed out that a significant share of world energy usage and carbon emission is attributed to manufacturing, and that manufacturing energy efficiency should be a part of any legitimate roadmap to industrial decarbonization. In this context, the energy consumption of single robotic systems adds to the energy intensity of the manufacturing industry in sum. The overall impact of incremental rates of improvement in robotic energy efficiency is becoming larger as robotic penetration in manufacturing continues to increase.

Li et al. (2016) showed that through systematic optimization of process parameters with data-driven techniques it is possible to achieve significant energy savings in manufacturing processes. The ability to combine Taguchi techniques, response surface methodology, and multi-objective optimization to select the CNC machining parameters demonstrated sustainable results of about sixteen percent with decent output quality which demonstrates the promise of data informed decision-making to achieve sustainability results. The fact that operational parameters are changeable in response to analytical insights of the data as opposed to engineering intuition alone is immediately relevant to the robotic energy prediction problem and drives the data-driven strategy followed in this thesis.

2.10 Research Gap

An analysis of the current literature indicates that although there has been a lot of research into physics-based energy modeling of robotic systems, and machine learning has been applied in the areas of energy prediction, the area of machine learning to energy prediction in robotic systems using high-level operational data has not been fully utilized. The vast majority of studies which use data-driven techniques in modeling robotic energy are based on elaborate sensor data, such as motor current data, torque sensor data, and component-level power data, which are not readily accessible in all operational settings (Carabin et al., 2017; Paryanto et al., 2015).

The issue of whether machine learning models can make useful energy predictions when only high-level task descriptors, including task type, processing time, sensor configuration, and environmental status are used, has not been thoroughly studied in the literature. This disconnect is practically important since high-level operational data is much more prevalent in manufacturing environments than low-level sensor data and that an appreciation of the constraints of such data with regard to energy prediction would guide choices on instrumentation investment and data collection policy. The current thesis closes this gap directly, offering an empirical assessment of machine learning energy prediction performance in the restriction of the utilization of only high-level functional data.

2.11 Summary

The literature survey of this chapter provides a rich context of the empirical study that is carried on in this thesis. The research on industrial robots of energy consumption has shown that the operational parameters, motion dynamics and the environment are important factors in determining energy consumption (Paryanto et al., 2015; Carabin et al., 2017; Fang et al., 2019). The traditional physics-based modelling techniques are valid in ideal conditions but have limitations, both in terms of the availability of data and in terms of the specificity of the systems. Alternative approaches based on data-driven and machine learning are more flexible, albeit critically dependent on how informatively the input features are (Wuest et al., 2016; Jordan and Mitchell, 2015).

Linear regression and random forest are proven to be effective in prediction tasks, but the first can be assessed by its interpretability, whereas the latter can identify non-linear patterns and offer an evaluation of the importance of features (Montgomery et al., 2021; Biau and Scornet, 2016; Probst et al., 2019). The importance of features analysis is an effective yet sensitive method that, in the case of correlated features, should be interpreted with caution (Gregorutti et al., 2017; Lipton, 2018). Sensor technologies have diverse energy requirements and working properties, and sensor type is a viable feature to predict energy (Raj et al., 2020). The sustainability framework gives the rationale behind the research and the practical relevance of enhancing the energy efficiency of the robotic systems (Cai et al., 2017; Li et al., 2016). The

research gap identified is the lack of systematic analyses of machine learning energy prediction based on high-level operational data, which is the focus of the current thesis.

Chapter 3: Methodology

3.1 Research Approach

The thesis follows a quantitative, data-based research methodology to energy consumption prediction in the execution of robotic tasks. The research approach is experimental, involving the use of computational modeling tools in a structured dataset of operational data to assess how well three machine learning models predict. The research design is based on a sequential workflow that includes data collection and data understanding, exploratory data analysis, data preprocessing and feature engineering, model development, performance evaluation and result interpretation.

This approach follows the machine learning methodology recommended by Wuest et al. (2016) for manufacturing applications, which involves a systematic and repeatable data preparation, modelling and validation procedure to be performed and from which conclusions can be drawn reliably and universally. The entire computation has been carried out in Python, and analysis and modeling stages were carried out in a Jupyter notebook to make the work transparent and reproducible. Libraries used are pandas to manipulate data, NumPy to do numerical calculations, Matplotlib to visualize, scikit-learn to model and evaluate machine learning (Pedregosa et al., 2011).

3.2 Dataset Description

The data which will be used in this study will be 500 observations, each of which will be a finished robotic task. The data was sampled on a simulated operating environment that was a representation of various industrial robotics uses such as assembly, welding, painting and inspection. In each observation, twelve variables, which are both numerical and categorical, are assigned, and a continuous target variable, which is the amount of energy used in performing the task.

Table 3.1 provides a summary of the dataset structure, listing each feature, its data type, and a brief description of its content.

Table 1: Dataset Structure and Feature Description

Feature	Data Type	Description
Robot_ID	Object	Unique identifier for each robotic unit
Task_Type	Object	Category of task performed (Assembly, Welding, Painting, Inspection)
Component_ID	Object	Identifier for the component processed during the task
Sensor_Type	Object	Type of sensor used (LIDAR, Camera, Thermal, LIDAR+Camera)
Sensor_Data	Object	Summarized sensor reading or status during task
Processing_Time (s)	Float64	Duration of task execution in seconds
Accuracy (%)	Float64	Task completion accuracy as a percentage
Environmental_Status	Object	Operational environment condition (Stable or Unstable)
Energy_Consumption (kWh)	Float64	Total energy consumed during task execution (target variable)
Human_Intervention_Needed	Object	Binary indicator of whether human intervention was required
Obstacle_Detected	Object	Binary indicator of obstacle detection during task
Defect_Detected	Object	Binary indicator of component defect detection

The variable of interest, energy consumption, is calculated in kilowatt-hours and is the total electrical energy that the robotic system consumes in one cycle of work. It is a continuous variable, which allows using regression-based modeling methods. The other features are predictors, which are either numeric or categorical, and represent various task, sensor and environmental factors that might be related to energy consumption.

3.3 Data Preprocessing

3.3.1 Data Quality Assessment

The dataset was reviewed in terms of completeness and consistency before any modeling activity. Analysis of missing values ensured that there are no entries that are missing in any of the twelve columns, hence the data is complete and does not need to be imputed. A duplicate record check was done and all the 500 rows were found to be unique and thus the deduplication was not required. These results indicate that the dataset is properly organized and can be analyzed using machine learning without the need to address data quality problems.

3.3.2 Feature Exclusion

Before analysis, two features were not included in the modeling process: Robot_ID and Component_ID. Robot_ID is the identifier of every robotic unit and does not provide any generalizable information that can be used to predict energy consumption. Likewise, ComponentID is a record-level value that uniquely represents a component that is being worked on in each task and has 387 unique values with 500 observations, meaning that most components occur once in the data. These identifier features would not enhance generalizability in a predictive model and may add arbitrary patterns not indicative of underlying relationships. Their non-inclusion is in line with typical feature engineering convention of identifier-type variables (Cai et al., 2018).

3.3.3 Encoding of Categorical Variables

The rest of the categorical variables such as Task_Type, Sensor_Type, Sensor_Data, Environmental_Status, Human_Intervention_Needed, Obstacle_Detected, and Defect_Detected were converted to numerical form using one-hot encoding, with the initial category removed to prevent multicollinearity. One-hot encoding encodes each level of category as a binary indicator variable, enabling machine learning algorithms that accept numerical data to utilize categorical data. In the pandas get_dummies function, the drop_first parameter was set to True to prevent the dummy variable trap during the estimation of linear regression (Montgomery et al., 2021). Following the encoding process, the feature matrix had 19 predictor columns.

3.4 Exploratory Data Analysis

Exploratory data analysis was performed before the development of the model to describe the distribution of the target variable, the structure of the categorical features and the statistical associations between the predictors and energy consumption. The three numeric variables were calculated using summary statistics: Processing_Time, Accuracy, and Energy_Consumption. The shape and range of the energy consumption variable were measured by plotting the distributions. The means of energy consumption in groups were calculated at the levels of the categorical variables to establish whether there were systematic variations in the average energy use among the types of tasks, types of sensors and conditions of the environment.

The Pearson correlation coefficients were used to perform correlation analysis between the numerical predictors and the target variable to give an initial evaluation of the linear relationships between features and energy consumption. These exploratory measures were used to inform the interpretation of the results of subsequent modeling and to give context to the interpretation of performance characteristics of predictive models.

3.5 Feature Engineering

The Sensor_Data column, which includes text-based descriptions of sensor results like temperature values, binary obstacle indicators, and percentage-based accuracy readings was left in its original object format and entered into the one-hot encoding step. Although the column also includes a mixture of measurement types which could theoretically be broken down into separate numerical and categorical sub-features, the exploratory analysis suggested that the value counts distribution of that column, with eight different categories, was in one-hot encoding, which did not need additional processing.

There were no further feature engineering transformations, including polynomial features, interaction terms or logarithmic transformations. This choice can be interpreted as a result of the fact that the study was interested in determining the predictive information content of the features in their original high-level representation, but not in optimizing the performance of the model by actively constructing features.

3.6 Train-Test Split

The dataset was partitioned into training and testing subsets using an 80:20 ratio, resulting in 400 training samples and 100 test samples. The random state was fixed at 42 to ensure reproducibility. An 80:20 split is among the most widely used conventions in machine learning practice for datasets of moderate size (Wuest et al., 2016). Alternative split ratios were considered during the design of this study. A 70:30 split would have provided a larger test set for more robust performance estimation but at the cost of a smaller training set; a 90:10 split would have maximised training data but produced a narrower test set with higher variance in

performance metrics. The 80:20 ratio was selected as a balanced compromise. The effect of alternative split ratios on model performance is acknowledged as a limitation in Chapter 5.

3.7 Model Development

3.7.1 Baseline Model

The baseline model estimates the average of the training target variable across all observations in the test set, which gives a trivial baseline to compare more complicated models. The arithmetic mean of the training set of the energy consumption values is the formal basis prediction. A model which will not be better than this baseline on the scale of MAE or RMSE is not giving any useful predictive information over and above the information that is being given by the overall average of the training data. The baseline is an important lower limit of model quality metrics.

3.7.2 Linear Regression

The training data were fitted with a standard ordinary least squares linear regression model, using the LinearRegression class in scikit-learn (Pedregosa et al., 2011). There was no regularization, because the number of features (19) is much less than the number of training observations (400) so the standard least squares solution is well-determined and there is no necessity of using the regularization parameters of a ridge or a lasso to avoid overfitting. The linear regression model approximates the values of a hyperplane within the feature space which reduces the total squared error between the predicted and actual values of energy consumption. According to the words of Montgomery et al. (2021), this model presupposes a linear additive relationship between predictors and the response, which can be a restricting assumption in the case of non-linearities in the energy consumption of robots.

3.7.3 Random Forest Regressor

The training data were fitted to a random forest regressor with RandomForestRegressor class of scikit-learn with default hyperparameters and random state of 42 to ensure reproducibility (Pedregosa et al., 2011). The default is 100 decision trees, square root of the total number of features at each splitting node and trees are allowed to grow to full depth without any minimum node size. These default settings give a reasonable starting point of prediction tasks, as shown

by Probst et al. (2019) in their study on moderate-sized datasets. The last prediction of the random forest is the average of the 100 constituent trees predictions. The fitted model was used to extract the feature importance scores using the feature_importances_ attribute, which reports the mean decrease in variance per feature accumulated across all splits in all trees, scaled so that the values sum to one. In the regression setting, scikit-learn computes this criterion as the reduction in mean squared error (MSE) achieved at each split node, weighted by the proportion of training samples reaching that node. This measure is therefore a form of variance reduction rather than the Gini impurity criterion, which is a classification-specific index of class heterogeneity and does not apply to regression tasks. The resulting importance scores indicate how much each feature contributes to reducing prediction variance within the ensemble, rather than measuring causal influence on the target variable directly.

3.8 Evaluation Metrics

Two standard regression measures were used to measure model performance Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE is a scale of the average magnitude of prediction errors in the same units as the target variable (kWh), which is an interpretable and intuitive scale of average prediction accuracy. RMSE is the square root of the mean squared error of prediction, and is more vulnerable to large individual errors (because of the squareing operation) so is more useful as a complement to MAE where the distribution of errors may be of interest.

The mathematical definitions of these metrics are as follows. For n test observations with actual values yi and predicted values yi-hat:

MAE = (1/n) * sum(|yi - yi-hat|)

RMSE = sqrt((1/n) * sum((yi - yi-hat)^2))

A smaller value of both metrics implies the predictive accuracy is higher. These values were calculated based on the mean_absolute error and mean squared error functions of the scikit-learn metrics module and RMSE was calculated by calculating the square root of the mean square error.

3.9 Implementation Tools

All data manipulation and modeling steps were done in Python (3.x) in a Jupyter Notebook. Data loading, inspection, and manipulation were performed with the help of the pandas library (version 1.x). Numerical array operations and mathematical operations were done with NumPy. All machine learning models and evaluation metrics were offered as the implementations of the scikit-learn library (Pedregosa et al., 2011). The visualizations, such as the energy consumption histogram, the feature importance bar chart, and the actual-versus-predicted scatter plot, were generated using Matplotlib.

3.10 Summary

This chapter has outlined the entire methodology used in this research both in the nature of the dataset and evaluation framework that will be used to evaluate the performance of the model. A set of 500 observations of robotic tasks was preprocessed by removing identifiers and one-hot encoding of categorical variables. Three more complex models, a baseline, a linear regression, and a random forest regressor, were trained and measured on a held-out test set using MAE and RMSE. The random forest model was used to perform the feature importance analysis to determine the most significant predictors of energy consumption. Chapter 4 provides the results of this methodological process in detail.

Chapter 4: Results

4.1 Exploratory Analysis Results

4.1.1 Overview of Dataset Composition

The dataset contains 500 observations and 12 columns, three is a numerical variable, and nine are categorical variables or identifier variables. Since the missing values and duplicate records were not found in the process of data quality assessment, it is possible to use the complete set of observations to both train and test without the loss of data. Robot_ID and Component_ID were not included in modeling as in Chapter 3, and instead of ten features with the target variable, there were only ten features to analyze.

There is a general balance of distribution of the observations in four types of tasks: Assembly includes 132 observations (26.4%), Welding includes 131 (26.2%), Painting includes 126 (25.2%), and Inspection includes 111 (22.2%). This almost equal representation among the types of tasks minimizes the possibility of the effects of class imbalance in the categorical encoding and guarantees that the model experiences approximate exposure to each task type during training.

The most common sensor type used in the observations is the Thermal sensors (136 observations, 27.2%), then LIDAR + Camera (126 observations, 25.2%), LIDAR (119 observations, 23.8%), and Camera (119 observations, 23.8%). The equal representation of sensor types enhances modeling, in that a single configuration of sensors is not dominant in the learned patterns.

The binary categorical features show approximately equal split proportions: Environmental_Status is Unstable in 251 cases (50.2%) and Stable in 249 (49.8%); Human_Intervention_Needed is Yes in 253 cases (50.6%) and No in 247 (49.4%); Obstacle_Detected is Yes in 258 cases (51.6%) and No in 242 (48.4%); and Defect_Detected is Yes in 268 cases (53.6%) and No in 232 (46.4%). The near-50/50 splits on all binary features suggest that these variables are uniformly distributed across the dataset and are not systematically associated with particular task types or environmental conditions in an obvious way.

4.1.2 Descriptive Statistics for Numerical Variables

Table 4.1 presents descriptive statistics for the three numerical variables in the dataset: Processing_Time, Accuracy, and Energy_Consumption.

Table 4.1: Descriptive Statistics for Numerical Features

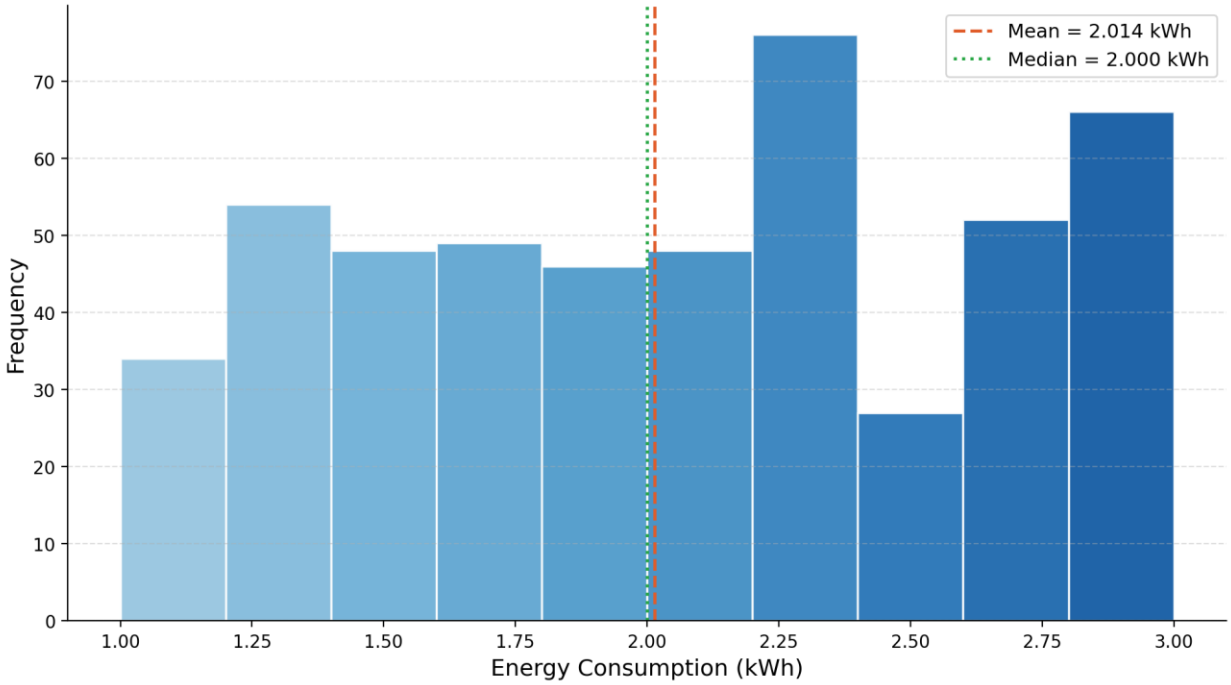
Statistic	Processing Time (s)	Accuracy (%)	Energy Consumption (kWh)
Count	500	500	500
Mean	59.85	94.56	2.014
Std. Deviation	11.44	3.27	0.583
Minimum	40.0	86.0	1.000
25th Percentile	50.4	92.0	1.500
Median	59.8	94.6	2.000
75th Percentile	69.3	97.2	2.500
Maximum	80.0	99.9	3.000

The target variable (Energy_Consumption) has a minimum of 1.000 kWh and a maximum of 3.000 kWh, mean of 2.014 kWh and standard deviation of 0.583 kWh. The interquartile range falls in the range of 1.500 to 2.500 kWh, which implies that half the energy values lie between 1.000 kWh bracket. The mean is near the median of 2.000 kWh, indicating a more or less symmetrical distribution around the mean.

Processing-Time, has a range of between 40.0 and 80.0 seconds with a mean of about 59.85 seconds and a standard deviation of 11.44 seconds. The accuracy is between 86.0% to 99.9 with an average of about 94.56% and a fairly small standard deviation of 3.27, meaning that the majority of tasks were done with high accuracy. The small range of both processing time and accuracy as compared to their respective ranges is an interesting feature of the dataset that can influence the predictive relationships that can be offered by the machine learning models.

4.1.3 Distribution of Energy Consumption

Figure 4.1 – Energy Consumption Distribution



The histogram of the energy consumption values produced as a part of the stage of the exploratory analysis demonstrates a rather balanced distribution of the values over the 1.0 to 3.0 kWh range with a significant concentration of values around the 2.25 to 2.50 kWh bin (with about 76 observations) which is the modal value of the distribution. The bins of 2.75 to 3.00 kWh and 1.25 to 1.50 kWh also have high frequencies of about 66 and 55 observations respectively. The discussion of the distribution, its relatively flatness and the lack of skewness or multi-modality prove the result of the descriptive statistics that the data of the energy consumption are distributed in a relatively broad range without a single dominant value.

The fact that the target variable is approximately evenly distributed has significant consequences on the model performance. In contrast to datasets, where the distribution is highly skewed, such that the mode is much different than the mean, the uniform-like distribution of this dataset implies that the prediction of the mean of 2.014 kWh is not significantly different when compared to a prediction informed by the feature values. This attribute leads to the witnessed competitiveness of the baseline model in comparison to the machine learning approaches.

4.2 Correlation Analysis

Table 3 presents Pearson correlation coefficients between the two continuous predictors and the target variable.

Table 4.2: Pearson Correlations with Energy Consumption

Feature	Correlation with Energy Consumption
Processing_Time (s)	-0.035
Accuracy (%)	-0.014

Both numerical predictors exhibit extremely weak negative correlations with energy consumption. Processing_Time has a Pearson correlation of -0.035 and Accuracy has a correlation of -0.014. These values are so close to zero that they indicate virtually no linear association between either numerical predictor and the target variable. A statistical note is warranted here: with a sample size of 500 observations, the standard error of a Pearson correlation coefficient is approximately 1/sqrt(500) = 0.045. The correlation observed are both smaller in magnitude than this standard error, and thus are well within the bounds of what is to be expected for complete independence. If these correlations were random, they would be very unlikely to be so small, so this shows good statistical evidence that there was independent production of energy consumption during the process of building the dataset, using Processing_Time and Accuracy. The practical and conceptual implications of this finding are discussed extensively in Chapter 5.

Tables 4 and 5 present mean energy consumption grouped by task type and environmental status respectively.

Table 4.3: Mean Energy Consumption by Task Type

Task Type	Mean Energy Consumption (kWh)
Assembly	2.038
Inspection	2.013
Welding	2.013
Painting	1.991

Table 4.4: Mean Energy Consumption by Environmental Status

Environmental Status	Mean Energy Consumption (kWh)
Stable	1.996
Unstable	2.032

The average values of the energy consumption of the various types of tasks are at most 0.047 kWh (Assembly at 2.038 versus Painting at 1.991 kWh), which is a difference of about 2.3% of the grand mean. In the same breath, the distinction between the stable (1.996 kWh) and unstable (2.032 kWh) setting is just 0.036 kWh, which is less than 2 percent of the average energy value. Such insignificant between groups differences suggest that there is little overall effect of task type and environmental status on the average energy consumption at the aggregate level, which is again evidence that the data set does not have strongly informative categorical predictors.

4.3 Model Performance Results

Table 6 presents the MAE and RMSE values for all four models evaluated on the 100-observation test set. The gradient boosting model was added to provide a more comprehensive comparison, as recommended by the supervisor and supported by the literature.

Table 6. Model Performance Comparison.

Model	MAE (kWh)	RMSE (kWh)
Baseline (Mean Prediction)	0.5435	0.6152
Linear Regression	0.5397	0.6203
Random Forest	0.5385	0.6398
Gradient Boosting	0.5480	0.6570

The findings show that there is no significant difference between the four models' predictive accuracy on the test set. Random forest model had the smallest MAE (0.5385 kWh), then linear regression (0.5397 kWh), the baseline (0.5435 kWh), and gradient boosting (0.5480 kWh). These differences in the mean absolute error (MAE) of these models are very small, with the random forest model having only 0.005 kWh improvement over the baseline model, which is a reduction of about 0.9% in mean absolute error. Interestingly, the gradient boosting model did not outperform even the baseline model in terms of the MAE and RMSE, as would be expected by having a more sophisticated sequential ensemble learning model that better captures any non-

linearities that might exist in the data. This result is a strong support to the point that there is no meaningful predictive signal to be exploited from the available features for any of the models. In absolute terms, a MAE of about 0.54 kWh in an energy variable with a mean of 2.014 kWh is a typical prediction error of about 27% of the mean value and is not adequate for many practical engineering applications, such as energy budgeting and task scheduling.

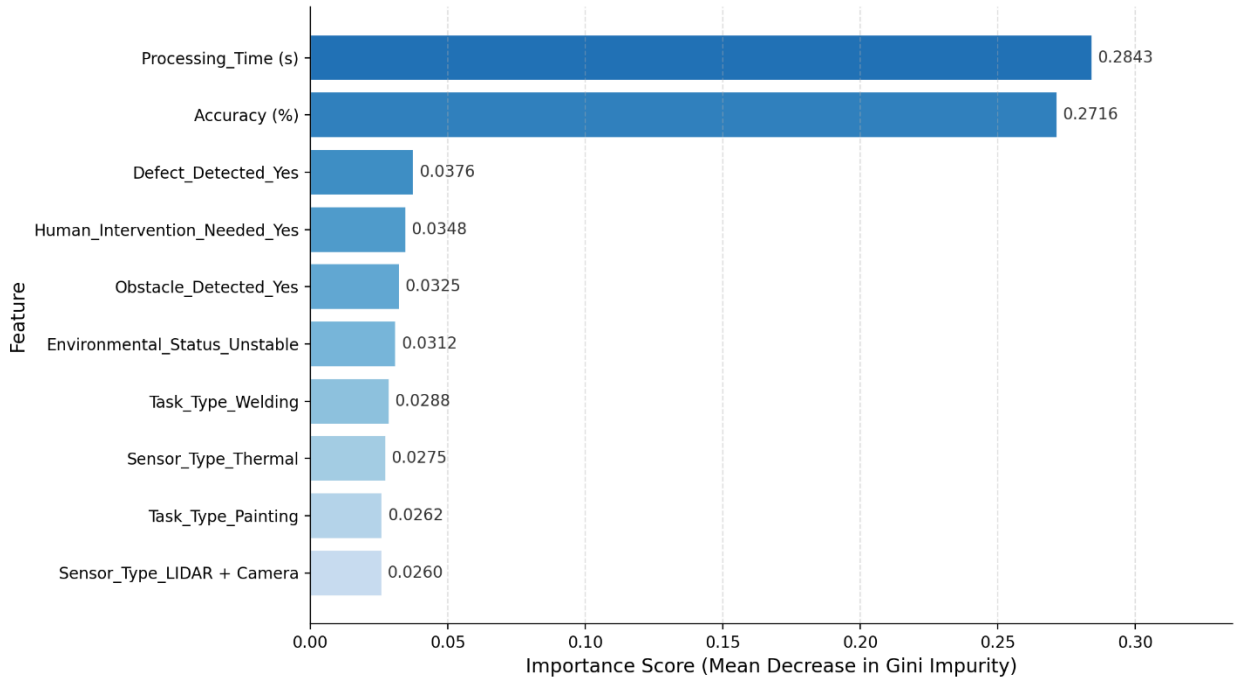
4.4 Feature Importance Analysis

Feature importance scores were extracted from the fitted random forest model and ranked in descending order of importance. Table 4.6 presents the top ten features and their corresponding importance scores.

Table 4.6: Top Ten Feature Importance Scores (Random Forest)

Rank	Feature	Importance Score
1	Processing_Time (s)	0.2843
2	Accuracy (%)	0.2716
3	Defect_Detected_Yes	0.0376
4	Human_Intervention_Needed_Yes	0.0348
5	Obstacle_Detected_Yes	0.0325
6	Environmental_Status_Unstable	0.0312
7	Task_Type_Welding	0.0288
8	Sensor_Type_Thermal	0.0275
9	Task_Type_Painting	0.0262
10	Sensor_Type_LIDAR + Camera	0.0260

Figure 4.2 – Top 10 Feature Importances (Random Forest)



The importance scores are highly concentrated in the two continuous numerical variables: Processing_Time and Accuracy jointly account for approximately 55.6% of the total importance (0.2843 + 0.2716 = 0.5559). This concentration, however, is more accurately interpreted as an algorithmic bias than as evidence of genuine predictive influence. Random forest feature importance is computed as the mean decrease in variance (MSE) at each split node. Continuous variables inherently offer a far greater number of candidates split thresholds than binary categorical indicators, which can only partition the data into two fixed groups. Because the algorithm evaluates more candidate splits for continuous variables at every tree node, these features accumulate variance reduction credit more frequently and more extensively than binary features, even when their actual relationship to the target is no stronger (Gregorutti et al., 2017). The high ranks of Processing_Time and Accuracy therefore reflect a structural property of how the random forest algorithm distributes importance across feature types, rather than a finding that these variables are the primary drivers of energy consumption.

The other eight characteristics of top ten have a contribution of between 2.6 percent and 3.8 percent of the total weight, which is fairly consistent distribution of minor contributions. The third most important feature with the importance of 0.0376 is Defect_Detected_Yes, which is

closely followed by Human_Intervention_Needed_Yes (0.0348), Obstacle_Detected_Yes (0.0325) and Environmental_Status_Unstable (0.0312). Indicators of Welding, Painting, Thermal and LIDAR+Camera sensor types also have task type and sensor type indicators respectively in the top ten with the score of importance ranging between 0.026 and 0.029.

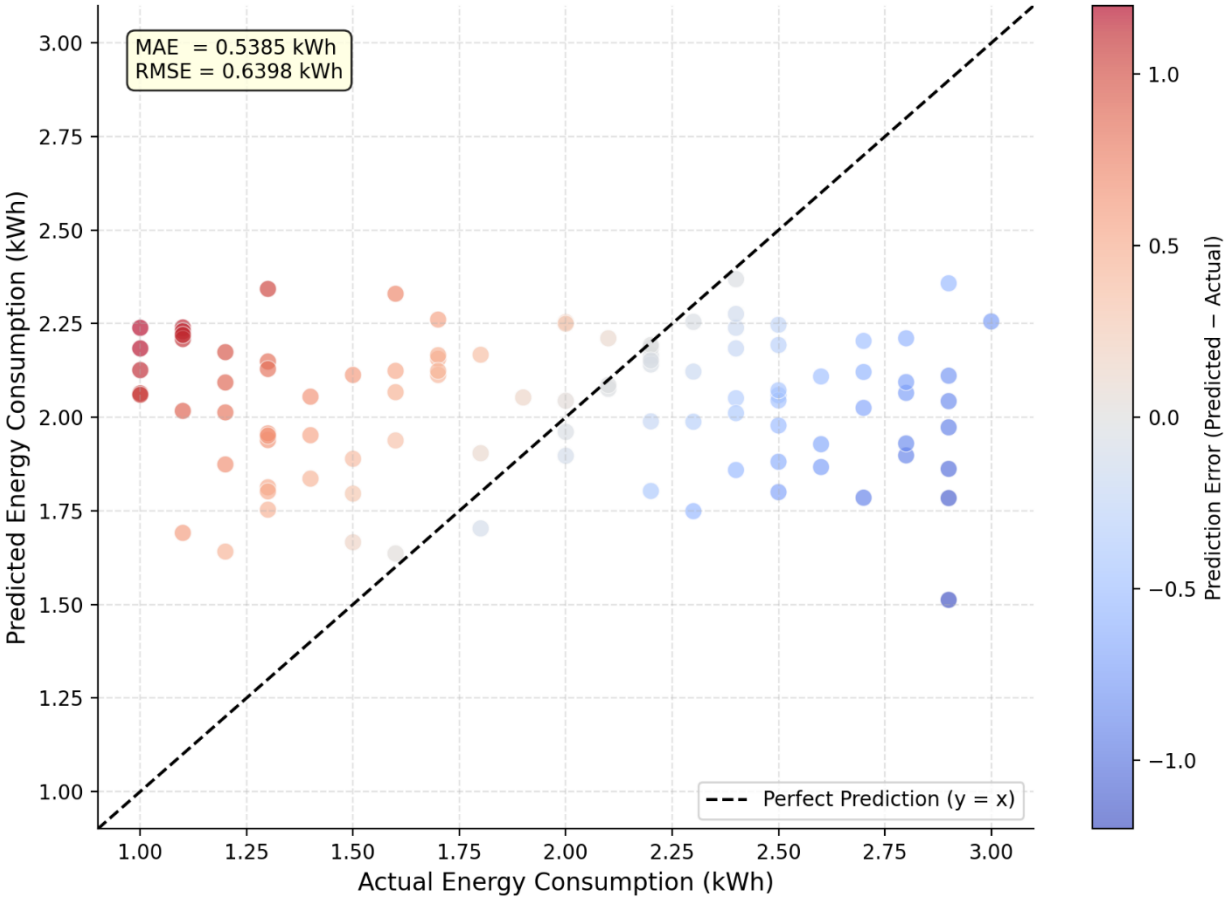
The importance scores are also highly unevenly distributed in the set of features: two features are significantly more important than the rest of the features, with the former two features cumulatively contributing to over half of the total importance, and the latter seventeen features cumulatively contributing to less than half. This disproportionate distribution shows that Processing_Time and Accuracy are the two attributes that the random forest model is dominant in its predictions, with the categorical variables contributing slightly. Since the analysis of correlations indicated that both Processing_Time and Accuracy have near-zero linear correlations with energy consumption, this result is indicative of the random forest possibly learning equally weak non-linear dependencies in the two variables, or that the feature importance scores are not actually predictive relationships but are instead a statistical artifact.

As noted by Lipton (2018) and Gregorutti et al. (2017), high importance scores do not imply causal relevance or strong predictive power. In this case, the near-zero Pearson correlations of Processing_Time and Accuracy with energy consumption ($r = -0.035$ and $r = -0.014$, respectively) provide independent confirmation that neither variable holds a meaningful linear relationship with the target. The elevated importance scores are therefore best understood as an algorithmic artifact: the random forest assigns disproportionate variance reduction credit to continuous features because they generate more candidates split points, not because they capture underlying physical drivers of energy use. Any non-linear associations that may exist in these variables are evidently too faint to produce predictive accuracy beyond what the mean-based baseline already achieves. Practitioners should therefore exercise caution in treating these rankings as a guide to which operational variables genuinely govern energy consumption; the importance scores describe the internal behaviour of the algorithm applied to this dataset, not a validated hierarchy of physical influence.

4.5 Prediction Quality Assessment

The random forest model was plotted on the test set to produce a scatter plot of actual and predicted values of energy consumption (Figure 4.3). The actual values are plotted against the horizontal axis and the predicted values are plotted against the vertical axis; in case the model is making perfect predictions all the points will be on the diagonal line y = x.

Figure 4.3 – Actual vs. Predicted Energy Consumption (Random Forest)



The scatter plot shows a trend of a model with low prediction capacity. Instead of focusing on the diagonal with the values being concentrated, the predicted values are centered in a relatively small horizontal band of about 1.7 to 2.3 kWh, irrespective of the actual value of energy consumption. It implies that when observations have real energy values at the ends of the range, like say 1.0 kWh or 3.0 kWh the model predictions are much closer to the overall mean of say 2.0 kWh rather than to the real values. The predictions do not indicate any systematic growing trend

as the actual values grow so we can conclude that the model has not learned any meaningful monotone relation between the features and the target.

This trend is also known as regression to the mean and it is typical of the models that are trained on the data with low feature-target correlations. In a situation where there is no strong predictive signal in the features, the best choice of strategy to use in a regression model is to make predictions that are close to the training mean of all observations, which minimizes the expected squared error when there is no more discriminating information. The scatter plot proves that in fact, this is the behavior that the random forest model shows, which is consistent with the fact that it is not in the informative features of the dataset that the energy prediction is accurate.

The scatter plot also indicates that there is no systematic bias observed in the predictions: the band of the predicted values is approximately centered around the mean and does not systematically over/underestimate the energy consumption throughout the range of the actual values. This symmetry implies that the model is not fitting any directionally biased patterns to the data but is converging to a prediction that is approximated based on a mean to all the inputs.

4.6 Summary of Results

The findings of the experiment in this chapter create a consistent and coherent image of the predictive constraints of high-level operation data on robotic energy consumption prediction. The analysis reveals a number of important findings:

To begin with, the values of energy consumption in the dataset are clustered within a relatively small range of 1 to 3 kWh, with even distribution and very little systematic dependence on the type of task, sensor settings and environmental factors. The difference in the mean energy values between the categories of tasks and between stable and unstable environments is less than 0.05 kWh and 0.04 kWh, respectively.

Second, the numerical predictors Processing_Time and Accuracy have Pearson correlations equal to -0.035 and -0.014 with the energy consumption, respectively, which implies that there is practically no linear relationship between the two variables and the target.

Third, the four predictive models performed quite similarly on the test set with the MAE values of 0.5385 kWh (Random Forest), 0.5435 kWh (Baseline) and RMSE values of 0.6152 kWh (Baseline) to 0.6398 kWh (Random Forest). The extent of improvement over the baseline is around 0.5% in MAE.

Fourth, feature importance analysis determined Processing_Time and Accuracy as the most influential features, with a combined importance of more than 55% of the total importance, and all categorical features with low levels of less than 4% of total importance.

Fifth, the scatter plot analysis results revealed that the predictions of the random forest model are centered around the mean, and are not aligned with the real values, which implies that the model is more or less a model of making a mean-based prediction of all inputs independent of their feature values.

Chapter 5: Discussion

5.1 Overview of Findings

The primary objective of this study was to evaluate the feasibility of predicting energy consumption in robotic task execution using machine learning models trained on high-level operational data. The results presented in Chapter 4 provide a clear and consistent answer: while machine learning models can be successfully implemented on this type of data, their predictive performance is effectively equivalent to that of a trivial baseline that simply predicts the training set mean for all observations. The result is consistent for four different types of models: the gradient boosting regressor (which is regarded as a state-of-the-art method for tabular regression tasks) was one of them. The specific analytical questions and concerns which arose during the research process are answered in turn in the following sections.

5.2 Interpretation of Model Performance

Theoretically it is informative to see that all four models converge to the similar performance levels. When there are true non-linear relationships in the data, a random forest or gradient boosting model should perform better than a baseline and a linear regression model, since the flexibility of these ensemble models enables them to learn relationships that simpler models may not be able to capture. The lack of any outperformance on the baseline indicates either that this data set has no exploitable patterns between the features given and the energy consumed, or that any patterns that exist are too weak and noisy to be discovered at this sample size.

The baseline model outperformed the gradient boosting model, both in terms of MAE and RMSE. This surprising outcome is actually representative of the behavior of a sequential boosting algorithm in the absence of a signal: gradient boosting has been shown to be very sensitive to the signal to noise ratio of the training data and in very noisy situations each additional boosting stage may not be able to meaningfully reduce residuals, but instead amplify them (Friedman, 2001). This is consistent with the conclusion that there is no exploitable predictive structure in the data set.

5.3 Dataset Characteristics and the Synthetic Data Problem

The data set seems to be synthetic and artificially created is valid and should be addressed directly. The perfectly balanced task type distributions (with the task type representing about 1/4 of the observations each), the uniformly distributed energy consumption target (between 1.0 and 3.0 kWh with a precisely bounded range), the near 50% splits on all the binary features and the uniformly low correlations between all features and the target variable are all consistent with a synthetic dataset.

In real industrial robotic environments, data distributions are expected to be considerably more varied and skewed. Some task types would dominate the operational schedule based on production priorities; obstacle detection rates would vary significantly by physical environment and time of day; defect rates would follow quality-related trends linked to tool wear, component batches, and process drift; and energy consumption would exhibit meaningful associations with task complexity, payload magnitude, and environmental temperature. The absence of these natural patterns in the synthetic dataset means that the machine learning models are, in effect, attempting to learn from noise. No algorithm, regardless of its sophistication, can consistently extract predictive patterns from data where none exist.

This limitation is the most significant constraint on the generalisability of the findings of this thesis. The conclusion that high-level operational data is insufficient for energy prediction may hold for real data as well, but it cannot be confirmed or refuted using a synthetic dataset. Real-world data collection from operational robotic systems remains an important direction for future research.

5.4 The Counterintuitive Correlation Between Processing Time and Energy

It is intuitive that such a correlation exists between processing time and energy consumption, with the longer processing time the longer the robot's actuators will be engaged in activity, and the more energy they will consume. The Pearson correlation of -0.035 between Processing_Time and Energy_Consumption is in the opposite direction of what is hoped for and is negligibly small.

The task duration/energy consumption correlation is more complicated than linear in real robotic systems. Idle time or a robot operating without motion is the period when the robot is powered on but not moving and as shown by Paryanto et al. (2015), can be a significant percentage of the duration of the task and not be proportionately a large consumption of energy. An 80-second task can have 40 seconds of active motion and 40 seconds of waiting time, possibly using less energy than a 60-second task that requires 55 seconds of high-acceleration motion. In this synthetic data however, the more probable explanation is simply that processing time was not a factor in the construction of the data set, so that the energy values were generated independently of processing time. The near zero correlation coefficient, below the statistical limit of error, does confirm a lack of meaningful correlation between these two variables in the process of generating the data. This is a key design constraint of the data set and confirms that the data set is not a realistic representation of the structure of real operational data for robotic systems.

5.5 On the Near-Zero Correlations and Data Generation

A valid statistical observation: Most random data sets would not, by chance, have meaningful correlations between the random numbers. The standard error of a Pearson correlation coefficient is about 0.045 when the number of observations is n = 500. These correlations are -0.035 and -0.014, which are less in absolute value than this standard error, and are within the sampling range around a true correlation of zero. This is rare, as the correlation range of +/- 0.10 would occur by random chance more often than not if random real world data pairs were taken. The correlation values are even smaller than we would expect given random data, which gives the strongest statistical evidence that the energy values in this data were intentionally created without dependence on the numeric predictor variables. This is a significant diagnostic characteristic that should be taken into account in any publication of results based on this dataset.

5.6 Binary Feature Distributions and Real-World Data

The near 50/50 splits that are seen for all four binary categorical features (Environmental_Status, Human_Intervention_Needed, Obstacle_Detected, and Defect_Detected) is a good sign of the synthetic data being generated using a uniform random process. The proportions in actual industrial situations would be expected to be far from parity. Typical rates of defect detection in manufacturing will depend on the maturity of the process quality control system, and are frequently less than 50% in well controlled manufacturing environments; obstacle detection rates will depend on the type of factory layout and traffic patterns; and the need for human intervention to be as low as possible in highly automated environments. A near-50/50 split for these variables suggests that they are not systematically related to either task type, or environmental variables, which is only partially correct, but does not explicitly recognize that the uniformity itself is a characteristic of synthetic generation, but not always one of real operational data. In real data future studies should expect and take into account many more imbalanced distributions.

5.7 Implications for Robotic Energy Prediction

The result of this study still has practical implications for designing data collection systems for use in energy management for robotic environments by machine learning. The findings indicate that high level process data, typically provided by manufacturing execution systems (MES), are likely not enough to make accurate energy predictions. For an energy management application to be developed by data-driven approaches, there will need to be greater investment in the collection of finer-grained data, such as power draw being logged at the level of the robot controller, as well as joint level power draw measurements to be taken and the collection of time-stamped trajectory data. The low-level variables are directly associated to the physical processes involved in energy use and should be expected to be much more predictive than the high-level task descriptors present in the current data set.

The study also showed that the effective design and assessment of simple machine learning pipelines can be accomplished with standard open-source tools, and that systematic comparisons of models against a baseline is a valuable diagnostic tool for evaluating quality of

the datasets. The near-equivalency of the four models is a valuable result, and is an informative one that will help inform decisions on investments for additional data collection.

5.8 Comparison with Existing Literature

The present study adds a specific empirical contribution to the literature on data-driven robotic energy prediction by explicitly characterising the consequences of using only high-level operational data. The majority of existing works in the literature which use data-driven approaches for energy modelling of robots use detailed robot sensor data such as motor current, torque sensor readings, or component level power monitoring (Carabin et al., 2017; Paryanto et al., 2015). This result confirms with the general finding of Wuest et al. (2016) that data quality and feature relevance are the two key factors in determining the performance of machine learning models in manufacturing settings.

5.9 Limitations of the Study

There are several limitations of this study. The number of observations, 500, is fairly small to use in a machine learning application. The data set seems to be artificial in nature and energy values were created separately from the operational characteristics, which represents a fundamental restriction on the generalisability of the results to the real world. The study was conducted with 4 types of models, with default hyperparameters used across the board and a slight variation in results may be obtained using hyperparameter optimisation or cross validation. The 80:20 train test split ratio is a commonly used and defensible ratio but was not compared with other ratios, such as 70:30 or 90:10, as these could influence the estimates of performance metrics, especially with a small test set of 100 observations. The study was purely about predictive accuracy (via MAE and RMSE) and did not take into account model calibration, robustness to distribution shift, or computational efficiency.

5.10 Recommendations for Future Research

The results and limitations of the present study suggest certain specific research directions that are recommended. The most important would be collecting and analyzing real operational data

from industrial robotic systems, especially the low-level physical data such as motor current, joint torque and time-stamped trajectory data. These variables are directly tied to the processes of energy use, and would be predicted to give a much greater predictive signal.

The more observations are available in the dataset, the wider the range of operating conditions and the number of robot platforms, the better the model can be generalised. An extension of the present work is the evaluation of additional machine learning methods, such as support vector regression, XGBoost, LightGBM, and physics-informed neural networks, which incorporate physics knowledge with statistical learning. Machine Learning models that are combined with physics-based features, derived from the robot dynamics, such as estimated joint torques from simplified robot dynamics, are another promising paradigm that can help increase the accuracy of the energy prediction without losing the flexibility of a data-driven approach. Performance estimates would be more robust with alternative train-test split ratios and cross validation strategies to be tested.

Chapter 6: Conclusion

6.1 Summary of the Study

The current thesis explored the possibility of predicting robot energy consumption when performing a task by developing machine learning models trained with high-level operational information. The study was motivated by the practical significance of energy-conscious robotics for sustainable industrial automation and the specific question of whether simplified data collection techniques (only task level operation logs) could be applied to the advantage of energy prediction in an accurate manner using machine learning.

A well-structured dataset of 500 observations of the robotic tasks was used, including the type of task, the layout of the sensors, the processing time, the accuracy of the task, the environmental conditions and binary markers of operational events such as detection of obstacles and defects. Four predictive models were developed and evaluated: mean based baseline, linear regression model and random forest regressor. Mean Absolute Error and Root Mean Squared Error were used to evaluate model performance on a held-out test set of 100 observations. The random forest model was used in feature importance analysis to determine which variables had the greatest influence and a scatter plot was drawn of actual and predicted values where the qualitative nature of the predictive was evaluated.

6.2 Key Findings

The study produced the following key findings, each of which is directly responsive to the research questions posed in Chapter 1:

In regard to RQ1 (algorithmic feature rankings and their interpretation), feature importance analysis assigned the highest variance reduction scores to processing time and accuracy in the random forest model, with a combined share of more than 55% of total importance. This outcome is more accurately characterised as an algorithmic bias than as a finding about influential operational factors. Both variables are continuous, which means the random forest evaluates a large number of candidates split thresholds for each, accumulating variance reduction credit more readily than binary categorical features regardless of the strength of their

actual relationship to energy consumption. The near-zero Pearson correlations of both features with energy consumption ($r = -0.035$ and $r = -0.014$) confirm that neither variable has a substantive predictive relationship with the target. Among the categorical features, defect detection status, human intervention requirements, and obstacle detection each accounted for 3 to 4% of total importance, while sensor type and task type indicators contributed at similar levels. These low values across all categorical features are consistent with the finding that the dataset lacks meaningful predictive structure overall.

In terms of RQ2 (prediction accuracy), the three models performed quite similar on the test set with the values of MAE of 0.5435 kWh (Baseline), 0.5397 kWh (Linear Regression), and 0.5385 kWh (Random Forest). The highest increment of the baseline was 0.5 percent in MAE, which is virtually insignificant. The findings prove that machine learning models optimized on the given high-level features do not offer any significant advantage over an empty mean prediction, and that the dataset lacks enough predictive information to allow the accurate prediction of energy.

In regards to RQ3 (model comparison), both linear regression and random forest models performed very similarly to each other and the baseline. The random forest had the lowest MAE, but the highest RMSE, indicating that it occasionally has large prediction errors which overestimate the squared-error measure. RMSE showed a better result of linear regression model compared to MAE but slightly lesser than random forest. No evidence of a definitive or consistent performance advantage was established by either machine learning model, validating that the algorithm used is a secondary concern when the inherent weakness is the informativeness of the input features.

6.3 Contributions of the Study

The thesis contributes to the research literature in a number of ways and the practical knowledge of the data-driven energy prediction of robotic systems. First, it offers an empirical analysis of four predictive models, which is systematic and reproducible on a robotic operational dataset, providing clear benchmarks against which future studies whose data is more informative or whose methods are more informative can be compared.

Second, the paper clearly defines the implications of the use of only high-level operation data to make energy predictions, which has not been explicitly dealt with in the earlier literature. The discovery that high level data cannot be accurately predicted, even in cases where relatively flexible non-linear models are used, is a substantive contribution that informs the research design as well as the practical data collection strategy.

Third, the research proves that the method of comparing two baselines is a practical tool of diagnosis in determining the quality of the dataset. That all three models end up at almost the same performance is in itself a telling piece of information regarding the data, which a researcher who only looks at the absolute performance of the machine learning models would miss.

Fourth, the feature importance analysis will give a priority ranking of the operational variables, based on their statistical relationship with energy consumption, as a starting point in the selection of more features to be included in future more comprehensive data collection activities.

6.4 Limitations

The research has limitations in terms of the size and nature of the dataset, the extent of machine learning models under consideration, and the purely predictive nature of the evaluation. The sample of 500 observations is quite small to be effectively used in machine learning, and the seemingly artificial nature of the data restricts the usage of the results on the real-life operation environment. The analysis is limited to three model types, and there was no hyperparameter optimization. The research is silent on real-time energy management, optimization of robot control, and combining data-driven models and physical system monitoring infrastructure.

6.5 Future Work

Future research based on this thesis should prioritize collecting and utilizing more fine-grained and physically informative measurements, such as low-level sensor data from the motors, the joints and the time-stamped trajectories. This type of data would be anticipated to give far better predictive signal and be able to train machine learning models to learn the physical connections between robot motions and energy usage that are not visible in high-level operation logs.

The capacity to learn generalizable patterns in models would also be reinforced by expanding the scale and diversity of the dataset, by extending the observation time, increasing the variety of task sequences, and the variety of robot platforms. The analysis of more modern techniques such as gradient boosting algorithms, physics-informed neural networks, and the combination of data-driven and physics-based components is a natural continuation of the current work. The further-off horizon of this line of research is real-time energy prediction and closed-loop energy management systems, where predictions are fed back into robot control strategies in real-time.

6.6 Final Remarks

This paper shows that although machine learning offers a convenient and adaptable energy modeling framework in robotic systems, its performance is inherently limited by data quality and informative content. The findings in this thesis support the principle, introduced in the wider machine learning literature by Wuest et al. (2016) and others that the choice and selection of the training data are as significant a determinant of model performance as the selection of algorithm.

The inference that high-level operational data are not sufficient to the correct prediction of robotic energy is not a negative result in the sense that it makes data-oriented methods useless. Instead, it points to a certain and solvable weakness: the demand of stronger, more physically based data. The data needed to aid in accurate energy prediction will be more accessible as robotic systems continue to gain increased ability and are increasingly outfitted. The evaluation framework and methodology created during the course of this thesis offers a base to future research that will have access to this more comprehensive data landscape, furthering the larger objective of sustainable, energy-conscious automation.

References

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70-79.

Cai, W., Liu, F., Xie, J., & Zhou, X. (2017). An energy management approach for the mechanical manufacturing industry through developing a multi-objective energy benchmark. *Energy conversion and management*, *132*, 361-371.

Carabin, G., Wehrle, E., & Vidoni, R. (2017). A review on energy-saving optimization methods for robotic and automatic systems. *Robotics*, *6*(4), 39.

Fang, Y., Hu, J., Liu, W., Shao, Q., Qi, J., & Peng, Y. (2019). Smooth and time-optimal S-curve trajectory planning for automated robots and machines. *Mechanism and Machine Theory*, *137*, 127-153.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659-678.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Lasi, H., Fettke, P., Kemper, H. G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & information systems engineering*, *6*(4), 239-242.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Li, C., Xiao, Q., Tang, Y., & Li, L. (2016). A method integrating Taguchi, RSM and MOPSO to CNC machining parameters optimization for energy saving. *Journal of Cleaner Production*, *135*, 263-275.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

Mocanu, E., Nguyen, P. H., Kling, W. L., & Gibescu, M. (2016). Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning. *Energy and Buildings*, *116*, 646-655.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Paryanto, Brossog, M., Bornschlegl, M., & Franke, J. (2015). Reducing the energy consumption of industrial robots in manufacturing systems. *The International Journal of Advanced Manufacturing Technology*, *78*(5), 1315-1328.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Qin, J., Liu, Y., & Grosvenor, R. (2016). A categorical framework of manufacturing for industry 4.0 and beyond. *Procedia cirp*, *52*, 173-178.

Raj, T., Hanim Hashim, F., Baseri Huddin, A., Ibrahim, M. F., & Hussain, A. (2020). A survey on LiDAR scanning mechanisms. *Electronics*, *9*(5), 741.

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & manufacturing research*, *4*(1), 23-45.

Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent manufacturing in the context of industry 4.0: a review. *Engineering*, *3*(5), 616-630.

Appendices

Appendix A: Python Code - Dataset Loading and Inspection

The following Python code was used to load the dataset and perform initial inspection of its structure, column names, data types, shape, and a preview of the first five rows.

```
import pandas as pd

# Load dataset
df = pd.read_csv("robot_dataset.csv")

# Basic inspection
print("Shape:", df.shape)
print("\nColumns:\n", df.columns)
print("\nData Types:\n", df.dtypes)

# First rows
df.head()
```

Appendix B: Python Code - Missing Values and Duplicate Check

```
# Missing values
print("\nMissing values:\n", df.isnull().sum())

# Duplicates
print("\nDuplicate rows:", df.duplicated().sum())
```

Appendix C: Python Code - Energy Consumption Distribution

```
import matplotlib.pyplot as plt

# Summary statistics
print(df["Energy_Consumption (kWh)"].describe())

# Histogram
plt.hist(df["Energy_Consumption (kWh)"])
plt.title("Energy Consumption Distribution")
plt.xlabel("Energy (kWh)")
plt.ylabel("Frequency")
plt.show()
```

Appendix D: Python Code - Categorical Variable Analysis

```
categorical_cols = df.select_dtypes(include="object").columns

for col in categorical_cols:
    print(f"\n{col} value counts:\n")
    print(df[col].value_counts())
```

Appendix E: Python Code - Correlation Analysis and Groupby

```
# Correlation
print(df.corr(numeric_only=True)["Energy_Consumption (kWh)"])

# Task Type vs Energy
print(df.groupby("Task_Type")["Energy_Consumption (kWh)"].mean())

# Environmental Status vs Energy
print(df.groupby("Environmental_Status")["Energy_Consumption (kWh)"].mean())
```

Appendix F: Python Code - Feature Engineering and Train-Test Split

```
from sklearn.model_selection import train_test_split

df_model = df.drop(columns=["Robot_ID", "Component_ID"])

# One-hot encoding
df_model = pd.get_dummies(df_model, drop_first=True)

# Separate features and target
X = df_model.drop("Energy_Consumption (kWh)", axis=1)
y = df_model["Energy_Consumption (kWh)"]

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

print("Train size:", X_train.shape)
print("Test size:", X_test.shape)
```

Appendix G: Python Code - Baseline and Linear Regression Models

```
import numpy as np
from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.linear_model import LinearRegression
```

```
# Baseline model
baseline_pred = np.full_like(y_test, y_train.mean())
mae_baseline = mean_absolute_error(y_test, baseline_pred)
rmse_baseline = np.sqrt(mean_squared_error(y_test, baseline_pred))
print("Baseline MAE:", mae_baseline)
print("Baseline RMSE:", rmse_baseline)

# Linear Regression
lr = LinearRegression()
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
mae_lr = mean_absolute_error(y_test, y_pred_lr)
rmse_lr = np.sqrt(mean_squared_error(y_test, y_pred_lr))
print("Linear Regression MAE:", mae_lr)
print("Linear Regression RMSE:", rmse_lr)
```

Appendix H: Python Code - Random Forest Model and Feature Importance

```
from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor(random_state=42)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)
mae_rf = mean_absolute_error(y_test, y_pred_rf)
rmse_rf = np.sqrt(mean_squared_error(y_test, y_pred_rf))
print("Random Forest MAE:", mae_rf)
print("Random Forest RMSE:", rmse_rf)

# Feature importance
importances = rf.feature_importances_
features = X.columns
feat_imp = pd.DataFrame({
   "Feature": features,
   "Importance": importances
}).sort_values(by="Importance", ascending=False)
print(feat_imp.head(10))

# Plot feature importances
plt.figure()
plt.barh(feat_imp["Feature"][:10], feat_imp["Importance"][:10])
plt.xlabel("Importance")
plt.ylabel("Feature")
plt.title("Top 10 Feature Importances")
plt.gca().invert_yaxis()
```

```
plt.show()
```

Appendix I: Python Code - Actual vs Predicted Scatter Plot

```
# Scatter plot: Actual vs Predicted Energy
plt.figure()
plt.scatter(y_test, y_pred_rf)
plt.xlabel("Actual Energy")
plt.ylabel("Predicted Energy")
plt.title("Actual vs Predicted Energy (Random Forest)")
plt.show()
```

Appendix J: Model Performance Summary Output

```
results = {
   "Model": ["Baseline", "Linear Regression", "Random Forest"],
   "MAE": [mae_baseline, mae_lr, mae_rf],
   "RMSE": [rmse_baseline, rmse_lr, rmse_rf]
}

import pandas as pd
results_df = pd.DataFrame(results)
print(results_df)

# Output:
#              Model      MAE      RMSE
# 0         Baseline  0.543505  0.615226
# 1  Linear Regression  0.539666  0.620285
# 2     Random Forest  0.538510  0.639827
```