 doi:10.20944/preprints202205.0379.v1

*Article*

# Analyzing optimal battery sizing in microgrids based on the feature selection and machine learning approaches

**Hajra Khan [1,†,‡], Imran Fareed Nizami [1,‡], Saeed Mian Qaisar [2,3,*,‡], Asad Waqar [1,*,‡*] and Moez Krichen [4,‡]**

[1] Department of Electrical Engineering, Bahria University; imnizami.buic@bahria.edu.pk
[2] Electrical and Computer Engineering Department, Effat University, Jeddah, 22332, Saudi Arabia; sqaisar@effatuniversity.edu.sa
[3] Communication and Signal Processing Lab, Energy and Technology Center, Effat University, Jeddah 22332, Saudi Arabia
[4] Department of Information Technology, Faculty of Computer Science and Information Technology (FCSIT), Al-Baha University, Alaqiq, 65779-7738, Al-Baha, Saudi Arabia; mkreishan@bu.edu.sa
* Correspondence: sqaisar@effatuniversity.edu.sa; asadwaqar.buic@bahria.edu.pk
‡ These authors contributed equally to this work.

**Abstract:** Microgrids are becoming popular nowadays because they provide clean, efficient, and low-cost energy. To use the stored energy in times of emergency or peak loads, microgrids require bulk storage capacity. Since microgrids are the future of renewable energy, the energy storage technology employed should be optimized to generate electricity. Batteries play a variety of essential roles in daily life and are used at peak hours and during a time of emergency. There are different types of batteries i.e., lion batteries, lead-acid batteries, etc. Optimal battery sizing of microgrids is a challenging problem, that limits modern technologies such as electric vehicles, etc. It is important to know different battery features such as battery life, battery throughput, and battery autonomy to get optimal battery sizing for microgrids. Mixed-integer linear programming (MILP) is an established technique for the integration and optimization of different energy sources and parameters for optimal battery sizing. A new MILP based dataset is introduced in this work. Support vector machine (SVM) is the machine learning application used to estimate the optimum battery size. The impact of feature selection algorithms on the proposed machine learning-based model is evaluated. The performance of the six best-performing feature selection algorithms is analyzed. The experimental results show that the feature selection algorithms improve the performance of the proposed methodology. Ranker search shows the best performance with a Spearman's rank-ordered correlation constant of 0.9756, linear correlation constant of 0.9452, Kendall correlation constant of 0.8488 and root mean squared error of 0.0525.

**Keywords:** Battery autonomy; battery size; feature selection;

## 1. Introduction

### *1.1. Background*

For establishing an energy system in a microgrid, the high cost of batteries is currently the key limiting factor [1]. If the battery size is not optimal, it can result in higher costs. Batteries not only enable consumers to store energy for extended periods but also help in saving the customers money by charging storage devices during off-peak hours when the cost is less. Because of the growing relevance of batteries, researchers have been working hard to develop highly efficient and cost-effective storage devices. The method would require taking non-invasive measurements of the battery in real-time and analyzing the results.

The conventional energy systems such as fossil fuels being used are causing environmental pollution and depletion of fossil fuels. Due to the increase in demand for electricity, there is a need for a new energy distribution system such as batteries. Microgrids can

 doi:10.20944/preprints202205.0379.v1



supply load and provide backup when the main supply is insufficient with improved power quality. The different modes of operating in a microgrid use energy storage systems to meet the intermittent nature of load demand.

Battery Sizing directly deals with the frame of total cost in a microgrid. The basic aim is to minimize the size of the battery and regulate the constraints such as voltage, reliability, and frequency to maintain the performance of the microgrid with a much smaller battery bank. For developing an energy storage system in a microgrid, the high cost of batteries is another key limiting factor [2]. Battery sizing should be considered to make the energy storage system economical and affordable to any consumer. Because of the growing relevance of batteries, researchers have been working hard to develop highly efficient and cost-effective storage devices. The method would require taking non-invasive measurements of the battery in real-time and analyzing the results [3].

The customer's primary issue with installing batteries is the high cost, where batteries produced by LG Cam, Tesla, and the Trojan that is the most well-known battery manufacturers have prices per kWh ranging between $148 to $158. Batteries are more expensive in comparison to distributed generation market demand models (DGENs), but they have a faster response time when it comes to balancing. This offers a scenario for optimal battery sizing in microgrids, where renewable energy (RE) penetration is higher than traditional grids. As a result, in most cases, a compromised joint venture of batteries and DGENs is used. Grid restriction arises in developing countries when the load is continuously increasing due to population growth. This would necessitate entire distribution system reinforcement and does not appear to be economically possible. As a result, sizing appropriate batteries and distributed energy resources (DERs) in grid-connected mode with a limited grid is a difficult task. The microgrid market in the United States is expected to expand from 986 million in 2019 to 1.89 billion in 2022 [4]. Jayashree proposed the approach of Mixed Integer Programming (Mathematical Models) and professional tools like MATLAB for BESS (Battery Energy Storage System) optimization [5]. Generic Algebraic Modelling System (GAMS) and CPLEX Optimization Studio were used in the domain of BESS in the research of Jayashree. Decision-making and multiple system simulations were considered the main part of the research to yield results and to minimize the size of the battery by regulating the same criteria for a microgrid. Apart from the research done by Jayashree, there is still enough room available in this domain to use more efficient tools and come up with system optimization of BESS and explore multiple applications that include battery banks.

### *1.2. Related works*

The Techno-Economic Method for the optimization of the annual demand forecast and the use of HOMER Pro allowed the researchers to analyze the advantages of renewable systems as compared to conventional grid application [6]. The drawbacks of this research are that the research done by Ramesh et al does not include the future data and is only valid for one-year data of the plant at a rural site. The perfect optimization is still to be considered as an important parameter that is not yet catered in detail in the research done by Ramesh et al. The pattern search technique for the optimization of the RE hybrid system is done with the MATLAB Simulink Design Optimization (SDO) with the help following algorithms Latin Hypercube, GA, and Nelder-Mead. It was observed with the help of HOMER Pro software that following the Nelder-Mead Algorithm decreases the optimal penetration of DG. The energy consumption and the demand with time were not analyzed in detail in the above research and there is still enough room for research to be done in the long-term demand forecasting and the habits of consumption.

The sizing and allocation of the BESS storage system in a microgrid help in regulating the parameters of a microgrid. The PSCAD Grid Modelling Software is proposed by Jagdesh Kumar in his research in which he used the software efficiently to predict the sizing constraints of BESS in isolated Renewable Plants [7]. The sizing characteristics of the battery bank were also analyzed with the help of simulations that were made by MATLAB

 doi:10.20944/preprints202205.0379.v1



and the results were shared accordingly in the research done by Jagdesh. The drawback of the research done by Jagdesh is that he does not consider the battery aging phenomenon which can be considered along with designing strategies of BESS for future research.

Hannan proposed in his research certain methods and algorithms like filter-based battery sizing method, Discrete Fourier transform-based ESS sizing method, Multiperiod decision-making model for optimum sizing. Grey Wolf Optimization Algorithm and Swarm optimization technique helped in achieving optimization and sizing BESS. Model Predictive control algorithm is also considered by Hannan to address and explore the optimum sizing of BESS [8]. The research by Hannan et al gave a direction for many researchers to proceed in the domain of battery sizing for efficient and cost-effective functionality of microgrids. However, these algorithms are the basic tools for optimization of future microgrids and are very tools for the implementation of sizing techniques in modern microgrids.

El-Bidari proposed the Grey Wolf Optimizer Approach and the development of Grey Wolf Optimizer for the optimization of sizing parameters of the battery bank and regulating the constraints in a microgrid by reducing the battery size [9]. The Optimizer approach along with GWO Algorithm is used as an efficient tool for battery sizing in the research done by El-Bidari. GWO provided a high level of robustness and a meta-heuristic algorithm to deal with the issue of frequency deviation. digsilent/ POWERFACTORY software is used as a basic tool for simulation in the research conducted by El-Bidari. A higher penetration level of variables is proposed to work in the future and pursue with this research to optimize the system even more.

Yang in his research, while addressing the problem of battery sizing and fluctuations in the renewable systems proposed the use of Sodium Sulphur (NAS) batteries for size optimization and reducing the rate of fluctuations in the renewable system [10]. Yang also highlighted that Markov Decision Processes (MDP) can no longer address this problem of BESS therefore there is a need for sensitivity-based optimization theory for addressing this problem. An iterative optimization algorithm was developed, and the BESS optimization was addressed properly in the research of Yang. Although the research was a big step towards renewables stability there is still room for many rapid and dynamic iterations to minimize the computational time of the system.

Gao performed deep research on the optimal sizing of batteries and proposed deep learning and algorithmic approaches to solve the matter of battery sizing and to achieve the optimal size [11]. Auto Encoders Extreme Learning Machine (SDAE - ELM) is introduced by Gao for optimization of battery size. Similarly, the research also addressed the use of Single-Layer-Feed-Forward Neural Network (SLFNN) and Deep Neural Networks (DNN) for optimization purposes. One drawback of the deep learning algorithm is that it requires a large amount of training data. For CNN and RNN models this drawback of the deep learning algorithm may cause a decrease in training efficiency. The future work for this research is to continue machine learning with a high level of artificial intelligence and macro-scale numerical approaches must be considered for the future.

Boonluk in his research proposed a GA and PSO for optimizing the size of the battery bank to be used. Fourier Coefficients were used to be processed in the algorithms and simulations were made accordingly on MATLAB and MATPOWER 7.0 [12]. It was also highlighted by Boonluk that the lifetime of each algorithm GA and PSO was the same (8.8 years). PSO in terms of objective function optimization was more efficient than GA therefore the future studies must continue by involving PSO instead of GA for more functional optimization.

Talent in his research used the MILP and GAMS along with the CPLEX Algorithm for the sizing of the battery [13]. One drawback of the research is that it does not consider the temperature profile of the batteries for the calculation of panel efficiencies. For future work, it is recommended that temperature constraints must be considered in the computational data to carry on the research.

Optimal Battery Sizing was done with the help of GA, PSO and IEEE 30 Test System was used for the implementations for optimal BESS [14]. OpenDNS with COM interfaced

 doi:10.20944/preprints202205.0379.v1



by the integration of IEEE 30 Test System were used for optimization. Future work of this research recommends the time series analysis and control of BESS. The drawback of the research is that it does not include multi-type BESS in the system.

Gupta in his research proposed a technique for battery sizing where he designed a MATLAB algorithm where all the constraints regarding the comfort and need of the user are entered [15]. The sizing considerations are calculated, and output is taken from the algorithm, and LOLP (Loss of Load Probability) is also considered as an important parameter in the research done by Gupta. Higher reliability factors and economic benefit are the important constraints of battery sizing in this research. The future work for this research is to adopt more dynamic algorithms to increase the computational speed of the system and to further optimize the size of the battery.

In research done by Nouhaila Lazaar, the Genetic Algorithm (GA) was used for optimal sizing of the battery [16]. The main objectives for this research done by Nouhaila was to decrease the Net Present Cost (NPC) of the system and the consideration of the Equivalent Loss Factor (ELF) for the index amount of reliability. The future research that can be pursued in the same domain can be the use of advanced algorithms that can consider more factors while dealing with the sizing of batteries in a microgrid Shaobo et al. [17], described that convex programming is a mathematical optimization that examines the problem of minimizing convex functions over convex sets, and it is used to formulate co-optimization of battery size, energy management, and battery aging. The notion of battery modeling was explained in detail, however, it had one flaw: the battery model was inaccurate, and it ignored critical elements such as state of charge, etc. Peiman Mirhoseini [18], have MILP framework-based model to utilize and evaluate the operating and trading costs of a battery charging station, which increases the system's reliability. However, because the model concentrates on installing a charging station as an MG and delivering clean electricity to meet its demands, dispatchable units (diesel generators, fuel cells, and etc.) are omitted.

J. Sampietro et al. [19], have studied the optimum sizing of batteries and supercapacitors in automobiles is achieved at the lowest possible total cost. Dynamic programming is used to determine the best utilization of storage systems and fuel cells. This paper contributed to the investigation of the relationship between battery size and cost.T. Terzimehic et al. [20], work is related to battery degradation, by using Support Vector Regression. The paper described how data-driven techniques can be used for battery forecasting. The author used data from various batteries operating at various temperatures and used that data to validate the machine learning results. Ji Wo et al. [21], have researched the Feedforward neural network (FFNN) that is used to mimic the relationship between Remaining Useful Life and charge curve because of its simplicity and effectiveness. The assessment of RUL for the battery under various charge current rates was neglected.

The operation plan entailed the Harmonious operation of fuel-powered generators and batteries, multi-unit DGen operation constraints, and reserve capacity to limit the number of hours the diesel generators are used. Sahibzada Muhammad Ali proposed [22] Linear Support Vector Regression (LSVR)and Rational Quadratic are used to train LSVR, Gaussian Process Regression, and Rational Quadratic in this work. For qualitative examination of trained models, the RMSE is utilized as a critical performance metric. The basic design parameters of the battery are to be addressed to minimize the battery size to improve charge storage capacity in less space. Thus, making the model much more compact. The methodology used in [23] is K-means clustering on customer net electricity data to extract critical information from limited input net/gross energy data, which is then used in a techno-economic simulation model to estimate the best battery size.

In support vector regression integrated is used for photovoltaic renewable energy system's ideal size by lowering the Annualized Cost of the System [24]. In terms of prediction accuracy, both hybrid SVR algorithms exceeded the single SVR method. Renewable energy resources should be used as much as feasible when generators are running. They also considered the unpredictability of Wind speed and clearness index. Ahmed Elnozahy [25] proposed weather unpredictability, a probabilistic technique based on an artificial neural

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1



network is used. A complete eco-techno-economic optimization research is integrated with the established choices and strategy. The proposed model to use batteries instead of parallel generators in time of emergency reduces the overall pollution and number of emissions affiliated with the previous 13 model. The aim is to design a model which proves to help balance the system and can act as a backup power source in times of emergency.

### *1.3. Research gap*

The traditional way of calculating battery life and health autonomy is based on minor degradation mechanisms, active material loss, and other factors. Even though these attempts have improved the situation, the problem still exists. Datadriven techniques that employ data from a microgrid's home load are being offered to improve these calculations. The solution would involve taking non-invasive measurements of the battery in real-time and combining these readings with regression-based machine learning algorithms to provide an accurate estimate without the use of any physical mechanism. A full dataset is provided as an input to generate correct estimations by employing the essential variables applying machine learning techniques. The proposed methodology can serve as a warning system, allowing to prepare ahead for any battery changes and keeping the system stable. Using machine learning techniques, a reliable microgrid battery system that can provide backup for a longer period can be developed.

### *1.4. Contributions*

The major contributions of this work are,

- A dataset for the residential load of a microgrid with 24000 samples and 40 parameters are developed using mixed integer linear programming (MILP) technique.
- A machine learning-based approach to optimal battery sizing in microgrids is proposed.
- Feature selection algorithms are utilized to identify the parameters that are more relevant and have a higher impact on battery sizing in microgrids.

The remainder of the paper is structured as follows. Section 2 describes the MILP technique and how the dataset is formulated. Section 3 describes the proposed feature selection based methodology for predicting the optimal battery size in microgrids. Evaluation criteria and experimental results are presented on subjective IQA databases in Section 4 followed by conclusion in Section 5.

## 2. Dataset generation using mixed interger linear programming (MILP)

Mixed-integer linear programming (MILP) is an optimization technique that employs the solvers to work for the real numbers with the equations restricted for both the linear inequalities and equalities. The solver will work within the various objective function that is normally described for either minimization or a maximization function [23]. The use of the MILP has been widely adopted for numerous optimization problems. Delivery and distribution are one such sector that utilizes the MILP. In research by Anusuyasarkar, Ganesh and Mohandas, 2008 the use of the MILP has been widely adopted for the Vehicle Routing Problem (VRP). The data is classified into nodes and the node relationships can be formulated through an algorithm [24]. The percentage deviation for the MILP can be formulated as follows [25].

$$RD = \frac{\bar{y} - \hat{y}}{\hat{y}} \times 100, \tag{1}$$

where $\bar{y}$ is the solution obtained using MILP, $\hat{y}$ is the best known solution obtained from iterations and $RD$ represents the percentage deviation.

The use of MILP has also been used for the scheduling of tasks. Scheduling is a complicated task that is mainly theoretical and works for various real-world systems [25]. The MILP is applicable for the minimization of the costs associated with the handling of goods. The use of the MILP is useful for system analysis and optimization for both large

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1



and complex solving problems [26]. There have been some limitations that have arisen for the more complex problems, one such area is the industrial symbiosis [26]. To rise above the problems due to the analyses, other methods like non-linear deterministic optimization and process integration have been adopted. The use of process integration has made great possibilities for the improvement of the various process algorithms [24].

In the complex notation of the various problems, there is the use of the various tools that will suit the advanced mathematical programming methods. In most cases, the use of MILP is interconnected with the other program optimization tools. This means that more than one tool can be used for the computation of the various findings. The use of the MILP has been attributed to two main features namely, modelling flexibility and the use of solvers that are linear programming. In the linear problem below,

$$Z_{MP} = min \sum_{i=1}^{N} f_i(x_i) \qquad st, \tag{2}$$

$$E \leq h, \tag{3}$$

$$0 \leq x_i \leq \mu, \tag{4}$$

Where $Z_{MP}$ is the minimizing equation for the function $x$. in the function, the initial values $i$ are obtained from 1 to a possible number integer $n$. The expectation for the $x$ is less or equal to a certain parameter $h$. The values of $x$ are within the limits 0 to $\mu$ and belong to a set that has the values with a maximum value of $n$. In the MILP there is the need to transform the equation that would easily be modelled via MILP through the consideration of the given breakpoints [27]. The first step is the identification that can also be considered as a piecewise linear function for the equation above for instance. Afterwards, the line segments are considered for the univariate continuous piecewise linear function. This is the case that will also consider the bounded domain hence a finite union [27]. This is easily modelled for the MILP. The various computations for the $gr(f)$ can be arranged as follows

$$0\lambda_0 + 1\lambda_1 + 2\lambda_2 ... = x, \tag{5}$$

$$2\lambda_0 + 3\lambda_1 + 1\lambda_2 ... = z, \tag{6}$$

where $\lambda_i$ is the coefficient for the formulation of the various auxiliary variables with relation to the original variable. The use of the auxiliary variables that denote lifted, or higher-dimensional. The use of the auxiliary variables gives the MILP an advantage in the computation since the original variables remain intact. The formulation of the MILP can be said to be in the form of,

$$S \subseteq \hat{Q}, \tag{7}$$

Where $S$ is the basic form for the MILP and is a set of auxiliary variables that has a power to the nth value. The projection shown below for the main $x$ variables is exactly $S$

$$Ax + B\lambda + Dy \leq b, \tag{8}$$

$$x \in Q^n, \tag{9}$$

Where $A$, $B$ and $D$ are variable constants for the linear function that is used to describe the actual function. They are also considered as proportions of the function distribution. Also, $x$ and $\lambda$ are the constants for the equation with $y$ as the intercept for the linear expression. The variable $b$ is the computation proportion for the summation of the basic form of the equation. Afterwards, the next thing is the formulation of the generic MILP through the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1



replacement of the various occurrence variables like $(x_i, z) \in gr(f_i)$. This means that the $x$ and $z$ are a set in the $g_f$ [27]. The new formulation for the MILP becomes,

$$Z_{MILP} = min \sum_{i=1}^{m} Z_i, \tag{10}$$

where $Z_i$ is the formulation of the MILP after considering the minimization function.

This is obtained for the summation of the values from the initial, $i$ to the maximum, $m$. The function is set to minimize the values set for the new formulation as shown. The workability for the MILP will be based on size & strength for the (Linear programming) and the formulation branching. The simple branch and bounding for the MILP will go in a long way to ensure faster resolution of the problems presented before it [27]. In the minimization formulation, the Linear programming relaxation bound is obtained and is easier to formulation to give the various trivial solutions. The finding of the trivial solutions can be formulated and done on a computer[28]. This will also ensure that the various variables can be adjusted for future computations. The modern world today has numerous complex problems that are also dynamic and would often require the adaptability of the various designs. This means that the modern algorithms that are used must have the flexibility to adapt to the various types of designs to be implemented. The MILP has great and wide flexibility and offers solution stability [29]. Moreover, the algorithm has gained a global presence through the search capability.

## 3. Proposed methodology

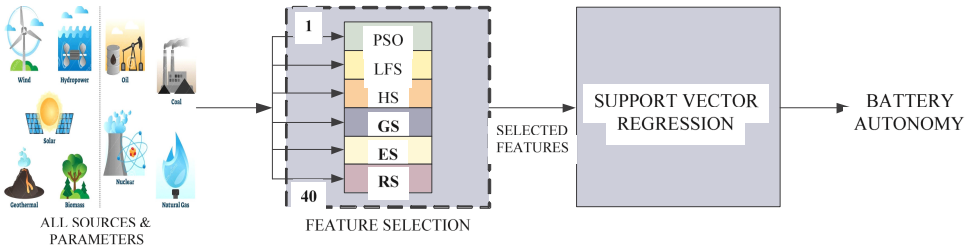



**Figure 1.** Proposed methodology for optimum performance.

Fig. 1 shows the two-step methodology proposed for optimal battery sizing of microgrids. The input to the system is power generation source factors and external parameters. In the first step, a feature selection algorithm is applied to the input to select the most relevant factors and parameters for optimal battery sizing in microgrids. The feature selection is performed using various search methods that try to navigate different combinations of attributes in the dataset to arrive at a shortlist of chosen features by keeping battery parameters as the target value. Many feature selection algorithms were evaluated but six top-performing search methods are considered here, that include ranker search, harmony search, evolutionary search, PSO search, genetic search, and linear forward search. In the second step, the selected features by the respective feature selection algorithm are given as input to the support vector regression (SVR) to predict the battery autonomy.

### *3.1. Generation techniques factors and external parameters*

#### 3.1.1. Photovoltaic power generation (PV) KW

The PV in kilowatts (KW) is the power that would be made available by the Photovoltaic cells for a given project. Normally the rating is given in watts, in the large-scale power production the rating can be given in kW [28].

#### 3.1.2. Distributed generation market demand (DGEN) KW

It is a tool that was developed for the analysis of the factors that affect the future market demands for energy resources [29]. The use of the dGen is vital for the focus of the

 doi:10.20944/preprints202205.0379.v1



future power demands and can be quantified as demand in kilowatts. This can therefore lead to the foreseen planning of the power demand and hence the battery size of microgrids.

#### 3.1.3. Hoppecke 6 OPzS 300

The solar-based grid power systems heavily rely on battery systems that can provide power to the systems when the sun peak is low. The Hoppecke 6 OPzS 300 is a 300Ah battery, with dimensions of about 147 $\times$ 208 $\times$ 420 mm, and weighs about 24.9 kg [30]. The battery is a vented lead-acid type of cell that has a longer life span of up to 18 years on the model [30]. The applicational use of the cell includes the mobile phone stations and the BTS stations. These form the mobile and signal applications in the communication sector due to their high reliability.

#### 3.1.4. Converter (KW)

It is the typical rating of the power inversion of an operation inverter. The inversion of the DC to AC would normally lead to some loss of current. The max peak power scope for the converter is considered when designing a PV system. This means that when the nominal converter demand is at 300 W (0.3 kW) an inverter of around 0.5 kW can be considered for the design. The design would consider the losses, zero voltage switching as well as the waveforms [31]. The converter affects the sizing of the battery since the higher the converter the larger the battery sizing.

#### 3.1.5. Total Capital Cost

The grid design is a capital-intensive project that in most cases requires the use of huge amounts of investment. The project overhead costs and the running costs would sum up the Total Capital Cost and can be estimated in USD ($) for the international standards due to the consideration of the US currency as a global monetary term [32]. This costing in the USD would in many cases safeguard the initial cost of projects against the local inflation of the prices of the forex as the projects can take a while for the implementation.

#### 3.1.6. Unmet Load Fraction

It is termed as the proportion of the total annual electrical load that went unserved because of insufficient generation for the system. In hybrid energy systems, variation in the power fluctuations leads to poor supply [33]. However, the compensation of the grid during the fluctuations would not only increase the reliability but also the efficiency of the system. This is often done using an optimization algorithm to compensate for the loss and to increase the dependency.

#### 3.1.7. Total net present cost (TNPC) $

TNPC is an economic parameter that is used for decision-making when doing a feasibility study on the power models. The NPC would present itself as an indicator in the terms of the United States dollar ($) to make the project be considered for development or stalled. The use of the NPC would also ensure that the overhead costs and other running costs are also considered in the economic projection of the estimates for the costs [34]. This makes the use of estimator to give a more reliable projection of the working costs of the design.

#### 3.1.8. Total Emissions

The total emissions are the volume of effluents that are released by a project into the environment [35]. The total emissions are considered for the design of most modern projects for the assessment of the environmental impact of the project. The impact assessment would be able to determine if the project would need to be adjusted or adopted as it is.

 doi:10.20944/preprints202205.0379.v1



#### 3.1.9. Total annual capital cost (TACC)

It is the annual cost of the capital that is considered for energy projects. This is considered for the project lifetime cost of operations that can be projected for the duration, the project will be operational. In most cases, the cost is also factored within the maintenance const for running the grid systems [36].

#### 3.1.10. Total annual replacement cost (TARC)

It is considered as the annual cost of the replacement for the various components that will be used for the grid system. The cost is factored for the working days throughout the year, and this will present a case to the investors on the overhead costs for the investment for the working duration of 1 year [36]. The capital cost would affect the battery sizing if more resources were allocated for the battery purchase.

#### 3.1.11. Total operations & maintenance cost (TOMC)

It is the annual cost of the operation and maintenance that is slotted for a particular project. The cost is given in USD for the working year from the inception of the plant/system [37]. The use of the USD would safeguard against the inflation of the local currencies. Higher O&M costs are suitable for larger sizing batteries for microgrids.

#### 3.1.12. Total fuel cost (TFC)

Total Fuel Cost is a feature that would look at the cost of operation of the fuel for the project the fuels can be fossil-based fuels and gases [37]. This means that working operation costs of the fuel is considered for the monetary terms i.e., cost of fuel.

#### 3.1.13. Total annual cost (TAC)

It is the total annual cost of operations for the project. The costs in most cases would include both the operations and other expenses. This means that the gross costs can later be worked from the cost to get an accurate projection of the working costs for the project [32].

#### 3.1.14. Operating cost

It is the cost of the factors of production that are used for the generation of the power for 1 year [32]. This is useful for the estimation and the evaluation of the project for future reference and helps in the planning of the other maintenance and costs that may rise.

#### 3.1.15. Cost of energy (COE)

It is the average cost per kWh of useful electrical energy that a system would produce [34]. This can be expressed as a $/kWh. This means that the higher the COE the better performing the system is in terms of the cost per power production hence more useful energy can be harnessed at a cheaper cost.

#### 3.1.16. Photovoltaic (PV) Production

It is the average projection of the photovoltaic cell power production in a given duration normally a year [28]. This means that this can be given for the average months and the working hours for the PV. This is useful for the estimation of the power production for the PV cell-based grid systems this can later be useful to enhance the dependability of the system. Higher PV production would require higher battery sizing and vice versa for the microgrids.

#### 3.1.17. Distributed generation production (dGENP)

It is a tool is developed for the analysis of the factors that affect the future market production of energy resources [29]. The production estimates are done in a given duration namely a year.

 doi:10.20944/preprints202205.0379.v1



3.1.18. Grid purchases (GP)

It is the cost incurred for the acquisition of power into the grid this is payable to the production companies that are involved [29]. In the independent grids, the purchases are often evaluated as the production costs for the power that would be produced. This is expressed as kWh/yr. and can be used for the projection of the future working operations for the grid.

3.1.19. Grid net purchases (GNP)

It is the cost minus the working expenses for the grid production of power. The net purchases are the real estimation of the costs after the working operational cost has been removed [29]. It is vital to determine the actual working costs for the grid to determine the profitability or the loss of the grid in each duration. The cost can be expressed in kWh/yr.

3.1.20. Total electrical production (TEP)

Total electrical production is the peak value of the power produced for the grid system that can be converted to useful power [37]. The total electrical production can be estimated for a plant or grid-based system. This can be used for the planning of the grid peak and base power demands. The higher electrical production would require higher storage and hence a larger battery sizing for the microgrids.

3.1.21. AC primary Load Served (AC-PLS)

It is the total amount of energy that can be used towards serving the AC primary load(s) for a year. The working loads for the duration are used as an estimator the power to enhance the AC loads for future considerations [38].

3.1.22. Deferrable load served (DLS)

It is the electrical load that requires a certain amount of energy for a given time. The time, in this case is not specific and can therefore wait for the availability of the power hence can be differed to a later time [39]. This means that the loads can be classified as deferrable when they are associated with the method of storage.

3.1.23. Renewable fraction (RF)

It is the renewable fraction i.e., the ratio of the nonrenewable to the total electrical energy served to a specified load [39]. This means that the ren fraction can determine the workability of the power over the similar produced through other means for renewable energy.

3.1.24. Capacity shortage (CS)

It is the total amount of capacity energy shortage that occurs throughout the year [28]. This is expressed as kWh/yr. and can be used for the planning of the alternative sources for both the peak and the shortfalls that may arise on the method or energy production. The capital shortage is a vital factor in the grip system to enhance reliability through the diversification of the other means of power production. The higher cap shortage would require a slightly larger battery sizing to cater to the shortfalls.

3.1.25. Unmet load (UL)

It is termed as the fraction for the proportion of the total annual electrical load that arises from the insufficient generation, hence it goes unserved [33]. The use of the Unmet load in a grid system is vital for operations and planning. The insufficient power can be planned to reduce the value of the Unmet Load. The expression for the term is kWh/yr.

 doi:10.20944/preprints202205.0379.v1



3.1.26. Unmet load fraction (ULF)

It is the ratio of the working power load to the unmet load [40]. The use of the fraction will aid in supply reliability since the vital components of the loads can be determined and help in increasing the dependability of microgrids [40].

3.1.27. Excess electricity (EE)

Excess Electricity is the surplus power that is produced by a system based on the estimated base loads for a given duration [28]. In the grid system, it is important to plan with the excess electricity that will help in setting up future loads.The value can be expressed in Kwh/yr. The excess electricity can be stored for a later use hence, the need for a larger battery sizing of the microgrid.

3.1.28. Diesel

It is a feature that evaluates the volumetric consumption of diesel (fuel) for a given duration. For renewable grid systems, the value for this is zeros since they are not dependent on fossil fuels to produce power [39].

3.1.29. Carbon dioxide (CO2) Emissions

: It is the deposition of CO2 as effluent into the air in terms of weight per given duration. The duration for a year will give the value in terms of kilograms per year. The emissions of carbon dioxide are useful for the determination of the carbon footprint of the industry and can also be useful for the total evaluation of a country/region's carbon emissions [40]. This provides a vital part in the environmental impact assessment for the plant.

3.1.30. Carbon mono oxide (CO) Emissions

It is the deposition of CO (carbon monoxide gas) as effluent into the air in terms of weight per given duration. The duration for a year will give the value in terms of kilograms per year. The emissions will help in the assessment of the plant in terms of the impact. Also, renewable energies have 0 emissions of CO [36].

3.1.31. UHC Emissions

It is the emission of the Unburned Hydrocarbons as effluent into the air in terms of weight per given duration. The emissions of unburned hydrocarbons can lead to toxic products and affect health. Renewable energies have 0 emissions on Unburned Hydrocarbons [37].

3.1.32. Particulate matter (PM) Emissions

It is a feature that investigates the emissions of the PM as effluent into the air in terms of weight per given duration. The PM comprises smaller particles like dust and other undissolved substances that can be found in flue gases of industries or plans that require combustion [41]. Fine PM is of the highest concern in the emissions due to volatility and the inability of disposal. In renewable energy, the PM Emissions are greatly reduced.

3.1.33. Sulfur dioxide (SO2) Emissions

It is the deposition of SO2 gas as effluent into the air in terms of weight per given duration. The duration for a year will give the value in terms of kilograms per year. In renewable energy, the SO2 Emissions are greatly reduced [37].

3.1.34. Nitrogen oxide (NOx) Emissions

It is a feature that investigates the emissions of the Nitrogen Oxides (NOx) as effluent into the air in terms of weight per given duration. The nitrogen oxides would result in the volatile gases that can result from the combustion of most of the fuels [41].

 doi:10.20944/preprints202205.0379.v1



3.1.35. Distributed generation market demand (DGEN) model Fuel

It is a feature that evaluates the distributed generation of the fuel, and it is expressed as liters per year. This is vital for the forecasting of the fuel requirements for the power system and hence the costs [32].

3.1.36. Distributed generation market demand (DGEN) model Hours

DGEN hours is a feature that evaluates the Distributed Generation in terms of active hours to produce electricity for a given duration. This can be expressed in a year. For instance, in PV production the active hours for the availability of the sun rays can be considered [28]. Lower DGEN Hours would require adequate storage of the power, and this would increase the battery sizing of microgrids.

3.1.37. Distributed generation market demand (DGEN) model starts/yr

It is the feature that looks at the working statistics for the Distributed Generation of the power system that can be analyzed for a given year. This would include the breakdowns and other issues that may arise.

3.1.38. Distributed generation market demand (DGEN) model Life

It is the feature that looks at the working life for the Distributed Generation of the power system that can be analyzed for a given year. The longer the life span of the system the better the operation ability of the system that can be evaluated.

3.1.39. Battery Throughput

It is the lifetime of the battery in years that is worked out by dividing the energy level by the duration [30]. A good performance battery would be able to give a better Battery Throughput. This means better dependability and power dispensation over the years for the system. The dependable batteries are suitable for the workability of a robust power grid especially the microgrid hence affecting the battery sizing positively for the system.

3.1.40. Battery Life

It is the estimated working life of the battery under which it can operate under normal terms with the power capacity storage and dispensation [38]. The battery life is also vital for the operational estimation of the power systems before a complete overhaul and helps in the maintenance operations. The higher the battery life the lower the sizing since the few installed capacity would take a while before changes can be done and hence the positive effect on the battery sizing of microgrids.

*3.2. Feature Selection*

The input features are subjected to feature selection. Various feature selction algorithms were analyzed but the performance of only six top feature selection algorithms is reported in this work. Feature selection algorithms select the most relevant features for optimal battery sizing. The details of each feature selection algorithm are given below.

3.2.1. Harmony search (HS)

Harmony Search (HS) is an optimization algorithm that utilizes the metaheuristic method. HS offers the advantage of search efficiency; algorithm simplicity and it converges quickly to the optimal solution. The resolution time for the method is generally low [42]. HS has been used on numerous engineering problems and has shown great application adaptations leading to the different versions of the algorithm to be adopted. In most engineering optimization problems, there is the consideration for the nonlinear and in some cases nonconvex functions that have intense equality. This has led to the increasing difficulties that arise from the solving of optimization problems using the traditional methods. HS is better suited for complex optimization problems. The HS method tries to

 doi:10.20944/preprints202205.0379.v1



search for the perfect harmony that is analogous to the optimal solution. This has led to harmonious improvisation.

$$x_{new} = x_{bold} + b_{\omega}, \tag{11}$$

where $x_{new}$ is the new harmony vector, $x_b old$ is the old harmony vector and $b_{\omega}$ is a constant. The random walk adjustment from the pitch can be illustrated to,

$$x_{new} = x_{old} + b(2\epsilon - 1), \tag{12}$$

where $x_{old}$ is the fixed variable for the pitch and $b$ is constant for the pitch displacement.

3.2.2. Evolutionary search (ES)

Evolutionary search utilizes the mechanisms that are inspired by nature for the solution of the various problems through processes that emulate the various behaviors of the living things. The mechanisms used for the development of the algorithm would therefore use the biological terms and evolution like reproduction, and recombination. The main working principle of the algorithm is the use of solutions that eliminate the weakest links and preserve the strong links i.e., the Darwin-based model. This helps in achieving a more viable solution [43]. The major benefits of the algorithm include increased flexibility, better optimization bandwidth, and unlimited solutions.

$$pr_i = \frac{F(k^i)}{\sum_{i=1}^{M} F(k^i)}, \tag{13}$$

where $F(k^i)$ is a function of random variables.

3.2.3. Genetic Search (GS)

The algorithm uses a set of terms named fitness function, initial population, mutation, selection, and crossover. The algorithm uses Darwin's model with genetic operators that form a key part of the problem-solution finding [44]. Some of the key benefits to the algorithm are the complex problems solving approach and its parallelism application. The diversification of the optimization to be able to deal with functions are stationary or. It can also deal with random noise. The ability of the algorithm to investigate various directions simultaneously in feature space makes it appropriate for the scientific field [42]. However, given its dynamism, the algorithm is one of the widely used in optimization that involves nonlinear data computations. The genetic search algorithm can be considered as a probability function for the chosen selector operator. In the case of chromosome, C the algorithm would be,

$$P = \left| \frac{f(C)}{\sum_{i=1}^{N} f(C)} \right|, \tag{14}$$

where $f(C)$ is the function for the chromosome and N is the total number of outcomes which depicts the nominal value.

3.2.4. Linear forward search

The method would use the sequential method that is key for the finding of the desired element in the list from a group. Upon the successful location of the searched item, the index would often be returned. The movement is in the forward direction when the search is performed [5]. The application for the linear search is mainly for the discrete values of data that would involve many elements. In n models, the function for the linear standard regression model can be written as.

$$y = Q\theta + \epsilon, \tag{15}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1



where $Q$ is the regression constant and $\theta$ is the variable for the regression. The error $\epsilon$ is used to make up for the second order diffrential equation. The main assumption is that the variance ($\sigma^2$) is additive. This means we can get the parameter, $\theta$ through the least square method [5].

3.2.5. Particle Swarm Optimization (PSO)

This is an optimization tool used for finding the optimal solutions to the specific parameters for a design requirement with a consideration of the lowest possible cost, optimization. The application is vast in various scientific fields. Since its introduction in 1995, the method has quickly gained several useful applications in various fields [6]. The adoption of the algorithm was based on social behavior especially the bees and insects that move in a swam (group). In nature, it is a stochastic novel-based population and key in solving complex nonlinear optimization problems [7]. PSO uses three parameters i.e., the number of dimensions, lower and the upper boundaries. The function can be elaborated as a function, where the minimum function can be seen below.

$$Min, f(x), x = (x_1, x_2, ..., x_N), \tag{16}$$

where f(x) is the function for the variable x, subject to several inequalities.

$$Subj = g_m(x) \leq 0 \qquad for\ values\ of\ m = 1,2,3..., n_g \tag{17}$$

where $g_m$ is the inequality function.

$$h_m(x) = 0 for\ m = ng + 1, ng + 2..., n_g + n_h, \tag{18}$$

where $n_h$ is the final value of equality. The algorithm mainly adopts five main principles. First, the proximity means the ability for the space and time computational adjustments incorporated into the model. Next, the quality refers to the swarm's ability to sense the changing quality for the environment and hence appropriate response. Thirdly, the diverse response is the ability of the smarm to change in a broad way and not in a narrow manner. Stability refers to the swarm not being able to change with all aspects of change but rather a controlled environment. Finally, adaptability refers to the change that is most suitable hence the worthy adjustment [7].

3.2.6. Ranker search

This is a search algorithm that uses the evaluation metrics to be able to retrieve the information mainly from various data sources. For instance, Google uses a ranger search algorithm; PageRank that will rank the various URL pages depending on the importance of the various web pages. So, the main function is the frequency that is considered by the search algorithm for the ranking of the web pages. The case study into the google search engine can be sued for the benchmark of the operations of the ranker search algorithm.

The algorithm would work in 3 stage process namely crawling, indexing, and serving. In the crawling stage, the use of other information gathering techniques like the bots will be used to get the updated changes to the various URL. Next, at the indexing stage, the categorical ranking of the various web pages is based on the content of either the images or the various texts. This is done by the identification of the various headers and the tags. Finally, at the serving stage also known as the ranking stage, the various URL will be listed based on the most relevant to the search parameter that is obtained for the search. The use of a similar concept is adopted by the various search engines with a few minor adjustments to the attributes of search like the price in some cases or the frequency of visits in some (inbound traffic) [8]. In the case of the ranker algorithm, the basic ranker search can be formulated as,

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}, \tag{19}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1



where A, B and C and web pages that are lined together, and L() is the outbound links. The various probability functions to the pages can be used. The overall function would become define PR,

$$PR(U) = \sum_{\forall \in B_u} \frac{PR(v)}{L(v)}, \tag{20}$$

The notation, Bu is the set that contains all links to the URL page u. and L(v) number of links to URL v. the damping factor is also considered for the algorithm.

The SVR uses tools like sparse solution and V-C control in the margins and the various support vectors [24]. The use of the SVR will consider a Hyperplane; this is a separation through alike that will aid in the prediction of the target value. The Kernel in an SVR model would be a function that would be suitable for the mapping of the data points into a higher dimension. The commonly used kernels are the sigmoidal, polynomial, and Gaussian radial basis function kernels. Finally, the boundary line margin that separates the hyperplane for the data points [24]. The illustration can be seen below for the SVR. The normal vector's magnitude relative to the surface can be estimated as

$$Min_w \quad 0.5||w||^2 \tag{21}$$

Where $w$ are the weights and the error is compensated in the constrains by constrains,

$$|y_i - W_i X_i| \leq \epsilon \tag{22}$$

where $y_i$ is the initial y constrains for the variable and $x_i$ is the initial $x$ constrains for the variable.

## 4. Experimental Results and Discussion

### *4.1. Data set and Evaluations Parameters*

The dataset is one of the basic requirements for the quantitative evaluation of a system. The data set is of the residential load of a microgrid. It is self-developed by using MILP. The data set has 24,000 samples and 40 features. Each row in the data set presents generation sources, external factors, and battery parameters.

#### 4.1.1. Spearman Ranked Correlation Coefficient (SROCC)

This is a nonparametric measure of the strength and direction of association that can be established between two variables. The assumptions used for SROCC are mainly three. The first assumption is that the two variables under study should be measured on an ordinal, interval, or ratio scale. The second assumption is that the two variables should present paired observation-based criteria. The last assumption is that there should be a monotonic relationship between the two variables [9]. The SROCC score is given as,

$$SROCC = \frac{\sum_i((x_i - \bar{x}) - (y_i - \bar{y}))}{\sqrt{\sum_i (x_i - \bar{x})^2 (\sum_i (y_i - \bar{y}))}} \tag{23}$$

where $x_i$ is the $i^{th}$ value of $x$, $\bar{x}$ is the mean value of $x$, $y_i$ is the $i^{th}$ value of y, $\bar{y}$ is the mean value of $y$.

#### 4.1.2. Kendal Correlation Constant

The KCC is used for the measurement of the ordinal association between two measured quantities. The correlation would test for the similarities in the ordering of data when it is ranked by quantities. The coefficient value of 1 means that the elements in the two sets are ordered in a similar manner i.e., high correlation. When the coefficient value is -1 ($\tau = -1$) means that the two sets are ordered oppositely. Finally, when $\tau = 0$ it means there is no relationship between the two sets [10]. The rank correlation can be expressed by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1



**Table 1.** Number of features selected by each feature selection algorithm

| Feature Selection algorithm | number of features selected |
|---|---|
| All | 40 |
| Ranker search | 29 |
| Particle swarm optimization | 14 |
| Linear forward search | 6 |
| Harmony search | 8 |
| Evolutionary search | 12 |
| Genetic search | 11 |

$$\tau = \frac{n_c - n_d}{n(n-1)}, \tag{24}$$

where $n_c$ is the number of cordant pairs, $n_d$ is the number of discordant pairs and $n$ is the total number of pairs.

4.1.3. LCC: Linear Correlation Constant

This is a measure of the strength of the linear relationship between two variables like x and y. The LCC values, r shows the strength of the relationship. When the value is near 1 or -1, the linear relationship is strong. Contrary, the value is near 0, which shows a weak relationship [26]. The assumption used is that the correlation coefficient would require the underlying relationship between the two variables to be linear. The concussions are mainly drawn from observable variables in most cases for the tests. The formulation of the linear coefficient can be seen below

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i^2)}} \tag{25}$$

where $x_i$ is the $i^{th}$ values for the $x$ and $y_i$ is the $i^{th}$ values of y.

4.1.4. RMSE: Root Mean Square Error

The RMSE is the statistical tool that is used for the prediction of standard deviation error (Residual). The residuals are the measure of how far predicted data points are from the original values [26]. This would mean that the RMSE would sow the concentration of the data around the line of best fit for a statistical finding [29]. The use of the RMSE is vital for statistical data to illustrate data relationships and establish the variation of the data from the set under study. The RMSE can be formulated as follows

$$RMSE = \sqrt{\frac{\sum(P-O)^2}{n}} \tag{26}$$

where P is the predicted values in the observations and O is the observed values for the observations with a sample size n.

*4.2. Performance Analysis*

TABLE 1 shows the number of features selected by each feature selection algorithm. The total number of features are 40. Ranker search selects 29 and particle swarm optimization selects 14 number of features. 6 number of features are selected by linear forward selection and 8 number of features are selected by harmony search algorithm. Evolutnary search and genetic serach algorithms select 12 and 11 number of features respectively. It can be observed that ranker search feature selection algorithm selects the largest number of features i.e., 29, whereas linear forward features selection algorithm selects the least

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1



**Table 2.** Performance comparison of feature selection algorithms.

| Feature Selection algorithm | LCC | SROCC | KCC | RMSE |
|---|---|---|---|---|
| All | 0.3760 | 0.2893 | 0.2165 | 0.2152 |
| Ranker | 0.9756 | 0.9452 | 0.8488 | 0.0525 |
| PSO | 0.9645 | 0.9252 | 0.7983 | 0.0608 |
| Linear forward | 0.9443 | 0.8846 | 0.7369 | 0.07518 |
| Harmony search | 0.9528 | 0.9076 | 0.7685 | 0.0701 |
| Evolutionary search | 0.9639 | 0.9229 | 0.7954 | 0.0613 |
| Genetic search | 0.9640 | 0.9237 | 0.7959 | 0.0613 |

number of features i.e., 6. particle swarm optimization algorithm selects the second largest number of features i.e., 14. Third least number of features is selected by harmony search i.e., 8.

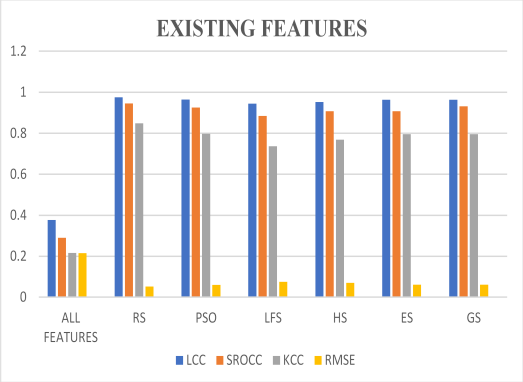



**Figure 2.** The performance comparison of feature selection algorithms in terms of LCC, SROCC, KCC and RMSE scores.

Fig. 2 shows the comparison of all the feature selection algorithms in terms of SROCC, LCC, KCC, and RMSE in the form of a bar graph. The feature selection algorithms help in improving the performance of the proposed methodology. There is a large difference between the performance of the system with and without feature selection, which can be observed that the SROCC score utilizing all the features is 0.2893, which is still lower than the SROCC for the worst-performing feature selection algorithm i.e., the linear forward selection is 0.9443.

The performance analysis of the proposed methodology in terms of SROCC, LCC, KCC, and RMSE is shown in TABLE 2. It can be observed that the performance of the proposed methodology improves when feature selection algorithms are utilized. The SROCC score is 0.2893 LCC score is 0.3764, KCC score is 0.2165 and RMSE score is 0.2152, when all features are utilized. Ranker search is ranked top with a SROCC score of is best performing as compared to all other algos in terms of SROCC.

Fig. 3 shows the box plot of the SROCC scores over 1000 iterations for the proposed methodology, when using all features and utilizing the top-performing six feature selection algorithms. It can be observed that the median value for the box plot of ranker search is the highest, which shows that it performs best. It can also be observed that the box plot for the ranker search is more compact as compared to all features, evolutionary search, genetic search, and harmony search, which shows that there is a lower standard deviation in the case of ranker search. A lower value of standard deviation shows that the results are more consistent over the 1000 iterations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1



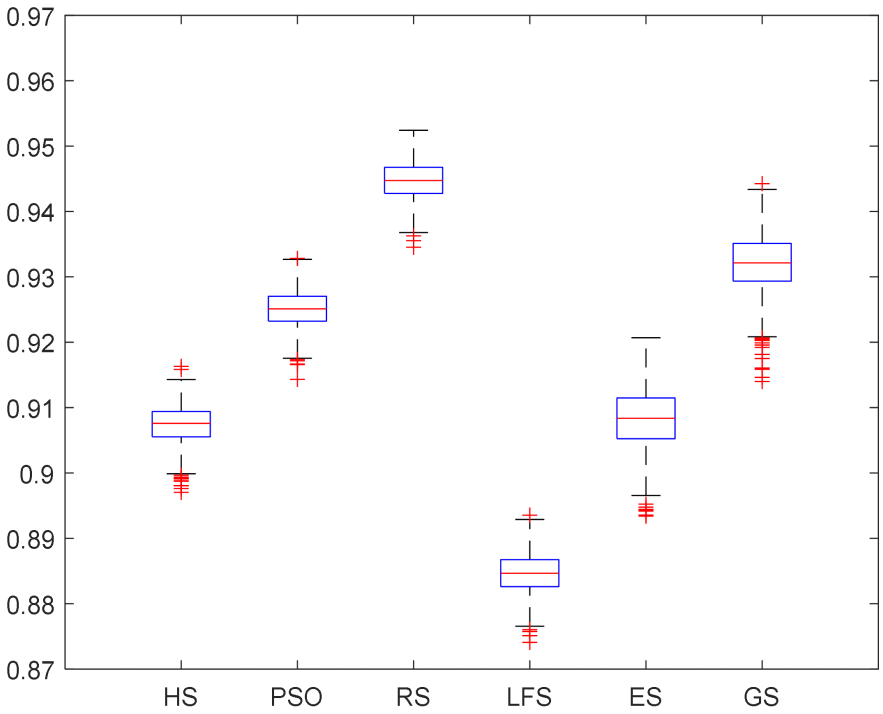



**Figure 3.** Box plot.

**Table 3.** Perforance comparison of proposed methodology with state-of-the-art techniques

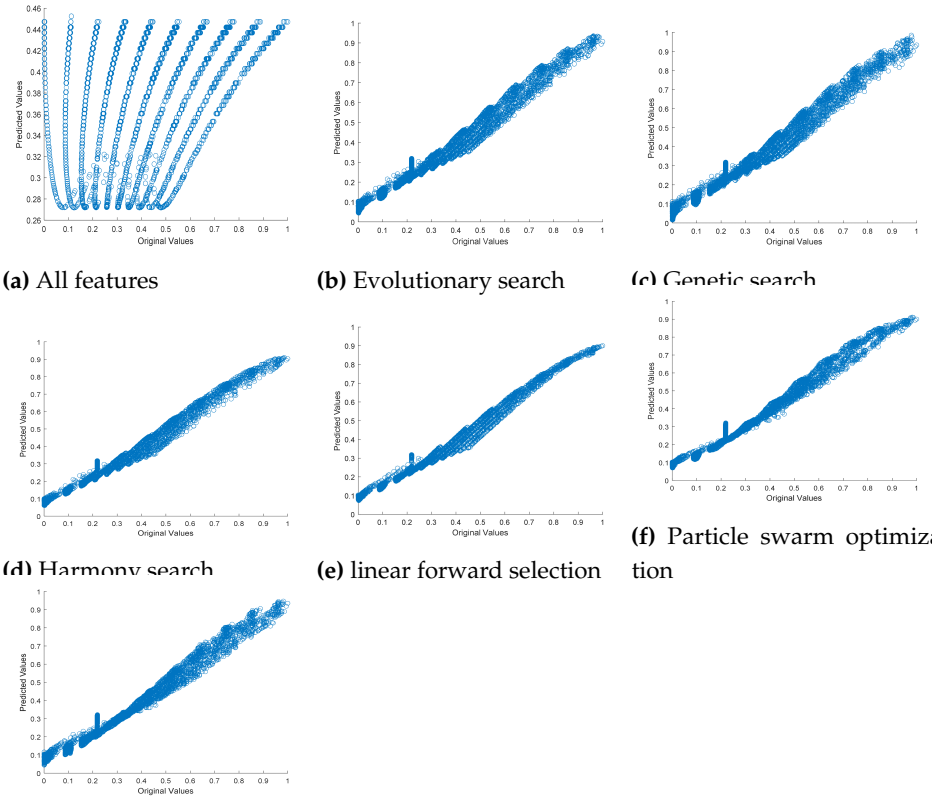
| Technique | RMSE score |
|---|---|
| Multi-Layer Perception | 6.3000 |
| Linear Support Vector Regression | 2.4090 |
| K-means clustering | 0.7170 |
| Support Vector Regression integrated with Harris Hawks Optimization | 0.1961 |
| Neural Network | 0.4240 |
| Proposed Algorithm | 0.0525 |

The performance comparison of the proposed methodology in terms of and RMSE is shown in TABLE 3. The proposed algorithm has the least value of RMSE i.e., 0.0525 as compared to other state-of-the-art methods when feature selection algorithms is utilized.

Fig. 4 shows the scatter plots of the proposed methodology using all features and with feature selection algorithms. The horizontal axis of each scatter plot represents the original values of battery autonomy computed using MILP and the vertical axis represents the predicted values of battery autonomy. The ideal case i.e., the best result would be if the data points of the scatter plot are aligned along the positive diagonal. Fig. 4 (a) shows the scatter plot of the original vs predicted values when all features are used. It can be observed that the data points are not aligned along the diagonal, hence the performance of the system can be improved. Fig. 4 (b) shows the scatter plot of the original vs predicted values when an evolutionary search feature selection algorithm is used. It can be observed that the data points are better aligned along the diagonal in comparison to when all features are used. It is also validated by the higher SROCC score of 0.9639. Fig. 4 (c) shows the scatter plot of the original vs predicted values when the genetic search is used to select features. It can be observed that the data points are better aligned along the diagonal in comparison to when all features are used. It is also validated by the higher SROCC score of 0.9640. Fig. 4 (d)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0379.v1






**(a)** All features **(b)** Evolutionary search **(c)** Genetic search

**(d)** Harmony search **(e)** linear forward selection **(f)** Particle swarm optimization

**(g)** Ranker search

**Figure 4.** Scatter plots for the original vs predicted battery autonomy scores (a)All features, (b)evolutionary search, (c)genetic search, (d)harmony search, (d)linear forward seection (e) particle swarm optimization (f) ranker search.

shows the scatter plot of the original vs predicted values when harmony search is used to select features. It can be observed that the data points are better aligned along the diagonal in comparison to when all features are used. It is also validated by the higher SROCC score of 0.9528. Fig. 4 (e) shows the scatter plot of the original vs predicted values when linear forward selection is used. It can be observed that the data points are better aligned along the diagonal in comparison to when all features are used. It is also validated by the higher SROCC score of 0.9443. Fig. 4 (f) shows the scatter plot of the original vs predicted values when particle swarm optimization is used to select features. It can be observed that the data points are better aligned along the diagonal in comparison to when all features are used. It is also validated by the higher SROCC score of 0.9645. Fig. 4 (g) shows the scatter plot of the original vs predicted values when ranker search is used for feature selection. It can be observed that the data points here are best aligned along the diagonal in comparison to all others. It is also validated by the highest SROCC score of 0.9756.

## 5. Conclusion

Microgrids are becoming more popular with each passing day but microgrids require bulk storage capacity to provide the stored energy in times of emergency or peak loads. Mixed-integer linear programming (MILP) is an established technique for the integration and optimization of different energy sources and parameters for optimal battery sizing. A new MILP based dataset is introduced in this work. Furthermore, a machine learning-based approach using regression is analyzed in this work for optimal battery sizing. It can be observed that the performance of the regression model when all the features of the MILP formation are used require improvement. Hence, feature selection algorithms have been utilized that help in selecting the most relevant features that have a high impact on battery sizing. The performance of six top-performing feature selection algorithms is

analyzed here. The ranker feature selection algorithm shows the best performance, particle swarm optimization is ranked second, genetic search is ranked third, evolutionary search is ranked fourth, harmony search is ranked fifth and linear forward selection is ranked sixth. The performance analysis shows that feature selection algorithms help in improving the performance of the proposed methodology in predicting the optimal battery size.

**Author Contributions:** "Conceptualization, I.F.N. and A.W.; methodology, I.F.N. and H.K.; software, H.K.; validation, A.W. and S.M.Q.; formal analysis, I.F.N. and A.W.; investigation, S.M.Q. and M.K.; resources, A.W. and S.M.Q.; writing—original draft preparation, H.K., I.F.N., and A.W.; writing—review and editing, S.M.Q., A.W., and M.K.; visualization, H.K. and I.F.N.; supervision, I.F.N.; project administration, I.F.N., A.W. and S.M.Q.; funding acquisition, S.M.Q.; All authors have read and agreed to the published version of the manuscript."

**Funding:** "This research work is supported by the Effat University under the grant number UC#9/2June2021/7.21(3)5, Effat University, Jeddah, Saudi Arabia.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** This article does not contain any studies with human participants or animals performed by any of the authors.

**Data Availability Statement:** Not Applicable

**Acknowledgments:** The research is technically supported by Bahria University. The authors acknowledge the financial support from the Effat University under the grant number UC#9/2June2021/7.2-21(3)5 , Effat University, Jeddah, Saudi Arabia.

**Conflicts of Interest:** "The authors declare no conflict of interest"

**Sample Availability:** Samples of the compounds ... are available from the authors.

## Abbreviations

The following abbreviations are used in this manuscript:

## Abbreviations

The following abbreviations are used in this manuscript:

 doi:10.20944/preprints202205.0379.v1



| | |
|---|---|
| MILP | Mixed integer linear programming |
| SVM | Support vector machine |
| DGENs | distributed generation market demand models |
| RE | renewable energy |
| DERs | distributed energy resources |
| BESS | Battery Energy Storage System |
| GAMS | Generic Algebraic Modelling System |
| SDO | Simulink Design Optimization |
| MDP | Markov Decision Processes |
| ELM | Extreme Learning Machine |
| SLFNN | Single-Layer-Feed-Forward Neural Network |
| DNN | Deep Neural Networks |
| GA | Genetic Algorithm |
| ELF | Equivalent Loss Factor |
| NPC | Net Present Cost |
| FFNN | Feedforward neural network |
| LSVR | Linear Support Vector Regression |
| TNPC | Total net present cost |
| TACC | Total annual capital cost |
| TARC | Total annual replacement cost |
| TOMC | Total operations & maintenance cost |
| TFC | Total fuel cost |
| TAC | Total annual cost |
| COE | Cost of energy |
| PV | Photovoltaic |
| dGENP | Distributed generation production |
| GP | Grid purchases |
| GNP | Grid net purchases |
| TEP | Total electrical production |
| AC-PLS | AC primary Load Served |
| DLS | Deferrable load served |
| RF | Renewable fraction |
| CS | Capacity shortage |
| UL | Unmet load |
| ULF | Unmet load fraction |
| EE | Excess electricity |
| CO2 | Carbon dioxide |
| CO | Carbon mono oxide |
| UHC | Unburned Hydrocarbons |
| PM | Particulate matter |
| SO2 | Sulfur dioxide |
| NOx | Nitrogen oxide |
| HS | Harmony search |
| ES | Evolutionary search |
| GS | Genetic Search |
| PSO | Particle Swarm Optimization |
| SVR | Support Vector Regression |
| SROCC | Spearman Ranked Correlation Coefficient |
| KCC | Kendal Correlation Constant |
| LCC | Linear Correlation Constant |

## References

1. M. A. Houran, X. Yang and W. Chen, "Energy Management of Microgrid in Smart," *Energy management of microgrid in smart building considering air temperature impact*, pp. 2398-2404, 2015.
2. C. Klansupar and S. Chaitusaney, "Optimal Sizing of Utility-scaled Battery with Consideration of Battery Installation Cost and System Power Generation Cost," *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology (ECTI-CON)*, pp. 498-501, 2020.

3. J. Sobon and B. Stephen, "Model-Free Non-Invasive Health Assessment for Battery Energy Storage Assets," *IEEE Access*, vol. 9, pp. 54579-54590, 2021.
4. X. Peng, C. Zhang, Y. Yu, and Y. Zhou, "Battery remaining useful life prediction algorithm based on support vector regression and unscented particle filter," *2016 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 1-6, 2016.
5. S. Jayashree and K. Malarvizhi, "Methodologies for Optimal Sizing of Battery Energy Storage in Microgrids: A Comprehensive Review", *International Conference on Computer Communication and Informatics (ICCCI)*, 2020. Available: 10.1109/iccci48352.2020.9104131 [Accessed 15 November 2021].
6. A. Cano, P. Arévalo and F. Jurado, "A comparison of sizing methods for a long-term renewable hybrid system. Case study: Galapagos Islands 2021", *Sustainable Energy & Fuels*, vol. 5, no. 5, pp. 1548-1566, 2021. Available: 10.1039/d1se00078k [Accessed 14 November 2021].
7. J. Kumar, C. Parthasarathy, M. Västi, H. Laaksonen, M. Shafie-Khah and K. Kauhaniemi, "Sizing and Allocation of Battery Energy Storage Systems in Åland Islands for Large-Scale Integration of Renewables and Electric Ferry Charging Stations", *Energies*, vol. 13, no. 2, pp. 317, 2020. Available: 10.3390/en13020317 [Accessed 15 November 2021].
8. M. Hannan, M. Faisal, P. Jern Ker, R. Begum, Z. Dong and C. Zhang, "Review of optimal methods and algorithms for sizing energy storage systems to achieve decarbonization in microgrid applications", *Renewable and Sustainable Energy Reviews*, vol. 131, pp. 110022, 2020. Available: 10.1016/j.rser.2020.110022 [Accessed 15 November 2021].
9. K. El-Bidairi, H. Nguyen, T. Mahmoud, S. Jayasinghe and J. Guerrero, "Optimal sizing of Battery Energy Storage Systems for dynamic frequency control in an islanded microgrid: A case study of Flinders Island, Australia", *Energy*, vol. 195, pp. 117059, 2020. Available: 10.1016/j.energy.2020.117059 [Accessed 15 November 2021].
10. Z. Yang, L. Xia, and X. Guan, "Fluctuation Reduction of Wind Power and Sizing of Battery Energy Storage Systems in Microgrids", *IEEE Transactions on Automation Science and Engineering*, pp. 1-13, 2020. Available: 10.1109/tase.2020.2977944 [Accessed 16 November 2021].
11. T. Gao and W. Lu, "Machine learning toward advanced energy storage devices and systems", *iScience*, vol. 24, no. 1, pp. 101936, 2021. Available: 10.1016/j.isci.2020.101936 [Accessed 16 November 2021].
12. P. Boonluk, A. Siritaratiwat, P. Fuangfoo and S. Khunkitti, "Optimal Siting and Sizing of Battery Energy Storage Systems for Distribution Network of Distribution System Operators", *Batteries*, vol. 6, no. 4, pp. 56, 2020. Available: 10.3390/batteries6040056 [Accessed 16 November 2021].
13. O. Talent and H. Du, "Optimal sizing and energy scheduling of photovoltaic-battery systems under different tariff structures", *Renewable Energy*, vol. 129, pp. 513-526, 2018. Available: 10.1016/j.renene.2018.06.016 [Accessed 16 November 2021].
14. P. Prabpal, Y. Kongjeen and K. Bhumkittipich, "Optimal Battery Energy Storage System Based on VAR Control Strategies Using Particle Swarm Optimization for Power Distribution System", *Symmetry*, vol. 13, no. 9, pp. 1692, 2021. Available: 10.3390/sym13091692 [Accessed 16 November 2021].
15. Y. Gupta, R. Vaidya, H. Kumar Nunna, S. Kamalasadan and S. Doolla, "Optimal PV – Battery Sizing for Residential and Commercial Loads Considering Grid Outages", *IEEE International Conference on Power Electronics, Smart Grid and Renewable Energy (PESGRE2020)*, 2020. Available: 10.1109/pesgre45664.2020.9070371 [Accessed 16 November 2021].
16. N. Lazaar, E. Fakhri, M. Barakat, J. Sabor and H. Gualous, "A Genetic Algorithm Based Optimal Sizing Strategy for PV/Battery/Hydrogen Hybrid System", *Artificial Intelligence and Industrial Applications*, pp. 247-259, 2020.
17. Xie Shaobo, Zhang Qiankun, Hu Xiaosong, Liu Yonggang, Lin Xianke, "Battery sizing for plug-in hybrid electric buses considering variable route lengths", *Energy*, vol. 226, 2021.
18. Mirhoseini, Peiman, and Navid Ghaffarzadeh. "Economic battery sizing and power dispatch in a grid-connected charging station using convex method". *Journal of Energy Storage*, pp. 101651, vol. 31, 2020.
19. Sampietro, José Luis, Vicenç Puig, and Ramon Costa-Castelló. "Optimal sizing of storage elements for a vehicle based on fuel cells, supercapacitors, and batteries", *Energies*, vol. 12, no. 5, pp. 925, 2019.
20. Ali, Asfand Yar, Abdul Basit, Tanvir Ahmad, Affaq Qamar, and Javed Iqbal. "Optimizing coordinated control of distributed energy storage system in microgrid to improve battery life", *Computers & Electrical Engineering*, vol. 86, pp. 106741, 2020.
21. Liu, Ye, Xiaogang Wu, Jiuyu Du, Ziyou Song, and Guoliang Wu, "Optimal sizing of a wind-energy storage system considering battery life", *Renewable Energy*, vol. 147, pp. 2470- 2483, 2020.
22. Liu, W., Yan, L., Zhang, X., Gao, D., Chen, B., Yang, Y. and Peng, J., "A Denoising SVR-MLP Method for Remaining Useful Life Prediction of Lithium-ion Battery", *In 2019 IEEE Energy Conversion Congress and Exposition (ECCE)*, pp. 545-550, 2019.
23. Kanwal, S., Khan, B., & Ali, S. M., "Machine learning based weighted scheduling scheme for active power control of hybrid microgrid", *International Journal of Electrical Power & Energy Systems*, vol. 125, pp. 106461, 2021.
24. Tang, R., Yildiz, B., Leong, P. H., Vassallo, A., & Dore, J., "Residential battery sizing model using net meter energy data clustering", *Applied Energy*, vol. 251, pp. 113324, 2019.
25. Abba, S. I., Rotimi, A., Musa, B., Yimen, N., Kawu, S. J., Lawan, S. M., & Dagbasi, M., "Emerging Harris Hawks Optimization based load demand forecasting and optimal sizing of stand-alone hybrid renewable energy systems–A case study of Kano and Abuja, Nigeria", *Results in Engineering*, vol. 12, pp. 100260, 2021.
26. Elnozahy, A., Ramadan, H. S., & Abo-Elyousr, F. K., "Efficient metaheuristic Utopia-based multi-objective solutions of optimal battery-mix storage for microgrids", *Journal of Cleaner Production*, vol. 303, pp. 127038, 2021.

27. A. F. Seyed and V. Ardalan, "Mixed-Integer Linear Programming for Optimal Scheduling of Autonomous Vehicle Intersection Crossing", *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 287, 2018.
28. Adriana, C. L., Nelson, L. D., Moises, G., Juan, C. V., Josep, M. G. "Mixed-integer- linear-programming-based energy management system for hybrid PV-wind-battery microgrids: modeling, design, and experimental verification", *IEEE transactions on power electronics*, vol. 32, pp. 2769-2783, 2017.
29. Hossein, S., Majid, M., S. Hamid, F., Gevork, B. G., "Optimal sizing and energy management of a grid connected microgrid using homer software", *Smart Grids Conference 2016*, pp. 20-21 Dec 2016., Kerman, Iran.
30. S. Kanwal, B. Khan, and S. M. Ali, "Machine learning based weighted scheduling scheme for active power", *Electrical Power and Energy Systems*, pp. 2-14, 2010.
31. Bo Lu and M. Shahidehpour, "Short-term scheduling of battery in a grid-connected PV/battery system", *in IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 1053-1061, May 2005.
32. R. M. Elavarasan, Shafiullah, G. M., Padmanaban, S., Kumar, N. M., Annam, A., Vetrichelvan, A. M., Holm-Nielsen, J. B "A Comprehensive Review on Renewable Energy Development, Challenges, and Policies of Leading Indian States with an International Perspective", *IEEE Access*, vol. 8, pp. 74432-74457, 2020.
33. D. C. Momete, "Analysis of the Potential of Clean Energy Deployment in the European Union", *IEEE Access*, vol. 6, pp. 54811-54822, 2018.
34. D. Papadaskalopoulos, D. Pudjianto, and G. Strbac, "Decentralized Coordination of Microgrids with Flexible Demand and Energy Storage"', *IEEE Transactions on Sustainable Energy*, vol. 5, no. 4, pp. 1406-1414, Oct. 2014.
35. S. Ganesan, U. Subramaniam, A. A. Ghodke, R. M. Elavarasan, K. Raju and M. S. Bhaskar, "Investigation on Sizing of Voltage Source for a Battery Energy Storage System in Microgrid with Renewable Energy Sources", *IEEE Access*, vol. 8, pp. 188861-188874, 2020.
36. C. Klansupar and S. Chaitusaney, "Optimal Sizing of Utility-scaled Battery with Consideration of Battery Installation Cost and System Power Generation Cost", *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology (ECTI-CON)*, pp. 498-501, 2020.
37. J. Sobon and B. Stephen, "Model-Free Non-Invasive Health Assessment for Battery Energy Storage Assets", *IEEE Access*, vol. 9, pp. 54579-54590, 2021.
38. Y. Wang, Y. Li, L. Jiang, Y. Huang, and Y. Cao, "PSO-based optimization for constant- current charging pattern for li-ion battery", *Chinese Journal of Electrical Engineering*, vol. 5, no. 2, pp. 72-78, June 2019.
39. X. Peng, C. Zhang, Y. Yu, and Y. Zhou, "Battery remaining useful life prediction algorithm based on support vector regression and unscented particle filter", *2016 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 1-6, 2016.
40. R. Kolluri and J. de Hoog, "Adaptive Control Using Machine Learning for Distributed Storage in Microgrids", *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, 2020.
41. K. Shivam, J. Tzou and S. Wu, "A multi-objective predictive energy management strategy for residential grid-connected PV-battery hybrid systems based on machine learning technique", *Energy Conversion and Management*, vol. 237, pp. 114103, 2021. Available: 10.1016/j.enconman.2021.114103 [Accessed 13 November 2021].
42. F. Pilati, G. Lelli, A. Regattieri and M. Gamberi, "Intelligent management of hybrid energy systems for techno-economic performances maximisation", *Energy Conversion and Management*, vol. 224, pp. 113329, 2020. Available: 10.1016/j.enconman.2020.113329 [Accessed 13 November 2021].
43. M. Mehrtash, F. Capitanescu and P. Heiselberg, "An Efficient Mixed-Integer Linear Programming Model for Optimal Sizing of Battery Energy Storage in Smart Sustainable Buildings", *IEEE Texas Power and Energy Conference (TPEC)*, pp. 1-6, 2020. Available: 10.1109/tpec48276.2020.9042498 [Accessed 13 November 2021].
44. M. Bagheri-Sanjareh, M. Nazari and G. Gharehpetian, "A Novel and Optimal Battery Sizing Procedure Based on MG Frequency Security Criterion Using Coordinated Application of BESS, LED Lighting Loads, and Photovoltaic Systems", *IEEE Access*, vol. 8, pp. 95345-95359, 2020. Available: 10.1109/access.2020.2995461 [Accessed 13 November 2021].
45. H. Khorramdel, J. Aghaei, B. Khorramdel, and P. Siano, "Optimal battery sizing in microgrids using probabilistic unit commitment," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 834–843, 2015.