Supervised Machine Learning Techniques for the Prediction of the State of Charge of Batteries in Photovoltaic Systems in the Mining Sector

ALEXA APAZA-PINTO1, JOSÉ ESQUICHA-TEJADA1, (Member, IEEE), PATRICIA LÓPEZ-CASAPERALTA2, AND JOSE SULLA-TORRES1, (Member, IEEE)

1Ingeniería de Sistemas, Universidad Católica de Santa María, Arequipa 04000, Peru
2Ingeniería de Minas, Universidad Católica de Santa María, Arequipa 04000, Peru
Corresponding author: Jose Sulla-Torres (jsullato@ucsm.edu.pe)

This work was supported by the Vicerrectorado de Investigación of the Universidad Católica de Santa María.

ABSTRACT  One of the critical aspects in the mining sector is energy, being of great importance for the operation since if it were to stop, one of the consequences would be the loss of large amounts of money. The research objective is to predict the State of Charge of Batteries of equipment powered by photovoltaic solar panels in the mining sector based on automatic supervised learning techniques. A monitoring system records each energy variable programmed in the photovoltaic system, for which an analysis of the data extracted from the monitoring system was carried out. The data were evaluated using automatic supervised learning techniques using the RapidMiner tool, whose prediction average was 90.12%. The technique of automatic supervised learning of artificial neural networks was chosen to predict the state of charge of batteries for photovoltaic systems. A software tool was built with the neural network. The analysis and discussion of the results of the training of the model were carried out, the contribution of this research being to determine the prediction of the state of charge of batteries in photovoltaic systems in the mining sector using techniques of supervised machine learning which was the neural network. Finally, with the model correctly trained, validation was carried out that allowed comparing the predictive data with the data in real-time, obtaining a good relationship and satisfactory results.

INDEX TERMS  Machine learning, photovoltaic systems, prediction, state of charge, supervised learning.

I. INTRODUCTION

The energy sector is essential for human beings in such a way that its absence would cause inconveniences such as the performance of different activities in daily life. However, they do not have this service in several parts of the world. Today one of the most viable solutions, technically and financially, to meet energy use needs is designated to those generated by photovoltaic (PV) energy systems.

Photovoltaic energies are technologies that, over time, have decreased maintenance costs, creating emissions that are harmful to the environment with great ease and linking to existing energy sources at the installation site. [1]

The interconnection of several photovoltaic cells is called a photovoltaic panel, which transforms solar radiation into electrical energy. A group of panels gives rise to a module, which, together with the batteries that reserve the electrical energy, the inverter, the switch, and the cables, make up the photovoltaic system.

Solar batteries store the electrical energy generated by photovoltaic solar panels so that they can be used at night and even on cloudy days; however, it is essential to accurately estimate the charge state in the batteries (SOC).

SOC estimation is a crucial link for the development of battery energy storage systems; In recently reviewed studies, no works like what is being proposed have been found, so the novelty of the proposal in this research is to use supervised machine learning techniques to evaluate the state of charge of batteries in PV systems in the mining sector.
The paper is made up of the following sections: In section I, it begins with the introduction to the research topic. In section II, a review of the literature related to the research is presented. Section III presents the methodology and materials used. In section IV, the results of the selected Supervised machine learning technique are shown, as well as a discussion about the results obtained and the import of the predictive model of the Supervised machine learning technique developed in a web computer module. Finally, conclusions and recommendations for future work are presented.

II. RELATED WORK

Machine Learning is one of the fastest growing areas of computing, with wide-ranging applications. It refers to the automated detection of significant patterns in data. Machine Learning tools are concerned with providing programs with the ability to learn and adapt [2]. The more data is supplied to a machine learning framework, the more it can be trained, and the consequences of the higher value of insights will be considered. Machine Learning is intelligent to discover and show the hidden patterns in the data [3].

Shanthamallu et al. [4] indicate that supervised learning techniques focus on using a labeled combination of data to train an algorithm that serves the purpose of the best function to describe the selection of input data. The supervised learning technique can be used to predict different aspects of human activity, such as sentiment analysis [5], classification for bone health [6], and in different emerging technologies [7], as well as in the monitoring of photovoltaic panels [8].

According to Rahimi-Eichi et al. [9], SoC estimation technology is essential for developing lithium-ion battery energy storage systems. An accurate SoC is the main element for the safe operation of battery packs. It is also a baseline for cell EQ control. The role of a cell balancer is to keep all cells equalized during charging and discharging to maximize pack capacity and protect cells from overcharging or discharging. It can cause permanent damage when a cell is overcharged or over-discharged. Therefore, improving the accuracy of the estimated SoC makes a difference in mitigating cell damage, prolonging the life cycle, and lowering the cost of maintenance [10].

This research will be focused on developing the prediction of the state of charge of batteries for photovoltaic systems using Supervised machine learning techniques and integrating it into a web computer module. The data collection of the energy variables for the equipment that provides radio coverage in the mine can be handy to estimate the prediction of the state of charge of the batteries using Supervised machine learning techniques having the voltage as input data of the batteries, temperature of the batteries, current consumption, current charge, and current panels.

According to Rodriguez [11], a Photovoltaic system is a set of equipment that uses solar energy for power generation and consists of the following components: a) Photovoltaic Panel, Formed by solar cells, which generally receive the sun’s rays and which uses the photoelectric effect, the direct conversion of energy from the sun to continuous electrical energy is obtained. b) Batteries, They operate as storers of the generated energy, which is distributed to their loads when the generation is low or when it does not exist in this case would be the absence of the sun. c) Regulator, an electronic device that works with the batteries to control their state of charge and ensure an optimal filling of energy. d) Inverter, this device allows the conversion of Direct Current (DC) into Alternating Current (AC), allowing loads that work with AC to work without any problem.

Gyu Gwang Kim et al. [12] presented a method to detect faults in a photovoltaic system based on the power ratio (PR), the voltage ratio (VR), and the current ratio (IR). Each ratio’s lower control limit (LCL) and upper control limit (UCL) were defined using data from a test site system under normal operating conditions.

Cook, Luo, and Weng [13] studied the constant resource-intensive re-evaluation of active residential PV locations, so they proposed to model solar sensing in a machine learning setup based on labeled data with supervised learning.

Karimi et al. [14] analyzed the photovoltaic cells from 5,400 cell images. From the data set, two unique degradation categories, “cracked” and “corroded,” were observed, while cells that were not degraded were classified as “good.” Cell images were classified into these three classes for supervised machine learning modeling, yielding 3,550 images. Using stratified sampling, a training and testing frame with a sampling ratio of 80:20 was generated. Three machine learning algorithms: support vector machine, random forest, and convolutional neural network.

Spyros Theocharides George et al. [15] evaluated the performance of different machine learning models to predict the power output of photovoltaic systems. Specifically, a variety of methods were explored, including artificial neural networks (ANNs), support vector regression (SVR), and regression trees (RT), with varied hyperparameters and features. Each model’s power output prediction performance was tested on real PV production datasets acquired over one year and compared to an existing persistence model (PM).

Úmit Agbulut et al. [16] designed and fabricated four layers of different sizes for a lumped PV system. The power outputs measured in the study were predicted with four machine learning algorithms, namely support vector machine, artificial neural network, kernel, nearest neighbor, and deep learning. To assess the success of these machine learning algorithms, coefficient of determination (R2), root mean square error (RMSE), mean bias error (MBE), t-statistics (t-stat), and mean absolute bias error (MABE) have been discussed in the document.

Kabilan R. et al. [17] presented a power prediction of a building-integrated photovoltaic system concerning various building orientations based on machine learning data science tools. The results showed that the application of linear regression coefficients to the forecast results of the developed PV power generation neural network improved the PV power.
A. Apaza-Pinto et al.: Supervised Machine Learning Techniques for the Prediction of the SoC of Batteries

generation forecast result. Jidong Wang et al. [18] used the Gradient Boost Decision Tree to make the prediction.

Dimd et al. [19] reviewed machine learning-based PV power output forecast models in the literature in the context of the Nordic climate. Ordoñez-Palacios et al. [20] predicted solar radiation in photovoltaic systems using Machine Learning techniques. Chen et al. [21] applied a random forest ensemble learning algorithm to detect and diagnose Photovoltaic array faults. Antunes Campos et al. [22] developed a machine learning application for temperature forecasting solar photovoltaic modules.

Kamarov and Susvoc [23] describe Autonomous Photovoltaic Systems with great applicability at different scales, from small photovoltaic systems that power lamps to networks capable of supplying electricity to entire populations. They usually have an energy storage system that allows them to function when there is no solar resource, at which time there is usually a greater demand for energy [24].

Ma, Hu, and Cheng [25] investigated a new data-driven method that can estimate SOC and SOE simultaneously based on a deep short-term memory (LSTM) neural network. The proposed algorithm is validated with two dynamic, driven cycles under various working conditions, such as different temperatures, different battery materials, and noise interference. Furthermore, the performance of the proposed method is compared with other popular algorithms, including support vector regression (SVR), random forest (RF), and simple recurrent neural network (Simple RNN). The results show that the proposed method obtains greater precision and robustness.

Tian et al. [26] indicate that recent development in deep learning provided an emerging solution for SOC estimation and proposed incorporating two types of domain knowledge in deep learning-based methods. First, the voltage and current sequences are decoupled into open circuit voltage (OCV), ohmic response, and bias voltage to augment the input of deep neural networks (DNN). Second, since conventional DNNs ignore the time dependency in SOC estimation results, they proposed a combination framework to adaptively merge DNN SOC estimation results from the DNN and short-term Ampere-hour predictions. The results show that the proposed method can drastically reduce the mean square error of the SOC estimate and the maximum absolute error.

In another work by Tian et al. [27], they propose a flexible method that uses only short chunks of load data to estimate maximum and remaining capacities to address state-of-health and state-of-charge estimation problems simultaneously. The proposed method is based on a convolutional neural network that only requires short-term load data to estimate two states. The results offer a flexible and easy-to-implement approach to accurate multi-state estimation over battery life.

III. METHODOLOGY

The methodology carried out for this research work is based on the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology [28], being a free methodology designed to provide standardization throughout the life cycle of a data analysis project. The CRISP-DM model covers the phases of a project, their respective tasks, and the relationships between these tasks. Fabio Porreca [26] used CRISP-DM for Decision analysis Based on Artificial Neural Networks to Power an Industrial Refrigeration System Using Photovoltaic Energy.

This methodology consists of six non-unidirectional phases, which indicates that the result of each phase determines which phase should be followed, each phase being handy for our research work.

Figure 2 shows the phases corresponding to the CRISP-DM methodology.

A. BUSINESS UNDERSTANDING

This initial phase will identify the problem that causes the perplexity in the organization, and a solution will be projected to solve the present problem.

Figure 3 shows a graph with the different activities of the Business Understanding phase, each of which will be described below.

1) SETTING OBJECTIVES

The main objective focuses on correctly monitoring the charge of the batteries in equipment powered by solar panels in the mining sector.

A grid-connected photovoltaic system (GCPS) is based on a dual technology, which has a photovoltaic-type generator at its disposal, in addition to being connected to a conventional power grid such as a backup system.

Figure 3 shows the scheme based on the battery bank in a photovoltaic system and made up of a solar panel that generates electricity from sunlight; this energy is stored in the
battery bank and constantly controlled by a regulator which regulates the charging intensity to extend the useful life of the batteries.

In Photovoltaic systems, two types of batteries are usually used: Lead Acid and Nickel-Cadmium. For cost reasons, the most common is lead acid batteries; however, Nickel-Cadmium ones are sometimes used in professional applications where cost is not a decisive parameter.

2) ASSESSMENT SITUATION
Mobile repeaters aim to provide coverage to different fixed and fleet infrastructure equipment in the mining sector in different open pit areas in the mine.

Being in different areas, the mobile repeaters work with solar energy through a photovoltaic system; now, this solar energy is accumulated in their batteries so that they supply energy and can also be used when there is no light, such as, for example, at night or cloudy days.

It was because the mobile repeaters that provided coverage were de-energized. However, the incidents caused by energy in the mobile repeaters have become a problem, causing the equipment of both the fixed infrastructure and the fleet in some areas of the pit not to communicate. The solution was to call the field personnel in charge of the mobile repeaters to carry out a corrective activity and get the batteries to recharge again. However, it took a few hours for the batteries of the repeaters to be recharged, and in the meantime, the mine operation stopped.

3) SETTING OBJECTIVES OF THE MACHINE LEARNING
At this stage, the determination will be made to apply a correct automatic learning technique to prevent the batteries of the mobile repeaters from running out of power. They can be checked previously, aiming to improve the service quality. Once the objectives and the problem are clear, we can move on to the next phase of the methodology.

Figure 4 shows a reference figure of the mobile repeater connected to a photovoltaic system.

B. DATA UNDERSTANDING

1) INITIAL DATA COLLECTION
The initial data was extracted through the Victron Energy System, which monitors the solar panels’ energy variables.

The taking of the energy variables is programmed to be recorded in a ratio of 15 to 25 minutes in the period of hours from 00:00 AM to 11:55 PM from Monday to Sunday, and each record is stored in the database. The Victron Energy System data were exported in an Excel Spreadsheet format.

2) DESCRIBE THE DATA
The data extracted from the Victron Energy System was exported in Excel Spreadsheet format with a total of 19062 corresponding records from April 10 to October 31, 2020.

All the fields have a Numeric and Text type field format since it seems that the exported data was obtained from different tables, which were put together in a single Spreadsheet.

3) EXPLORE THE DATA
When the data was explored, it was appreciated that within the fields that were exported, only three would contribute any value to the analysis of the State of Charge of the Batteries, which were Voltage, Consumption, and Current; thus, omitting the other fields exported from the Victron Energy System.

The three fields would correspond to voltage, consumption, and current.

C. DATA PREPARATION
In this phase, the final database that will feed the modeling tool will be built; the Spreadsheet was migrated to a Comma Delimited CSV File.

1) DATA SELECTION
Through the data section, it was possible to highlight those fields that would contribute any value to the analysis of the State of Charge of the Batteries, which were Voltage, Consumption, and Current; thus, omitting the other fields exported from the Victron Energy System.

However, it was considered to add one more field whose name would be SoC, to which we would add a formula to find the state of charge, refer to (1). This formula was applied through the values of the Voltage, Consumption, and Current.
Fields mentioned above.

\[
SoC(t) = SOC_0 + \frac{n}{C} \cdot \int_{t_0}^{t} I(t) \cdot dt
\]  

(1)

where: \(SOC_0\): the initial state of charge, \(n\): battery performance, \(I(t)\): battery current, \(t_0\): initial instant, \(SoC\): state of charge, \(C\): battery capacity.

The formula and the results obtained were validated by the Field Engineer, who would oversee the maintenance of the equipment’s solar panels.

2) DATA CLEANING

The data cleaning tasks to be executed are:

- Type conversion for the Voltage, Current, Consumption, and SoC fields from General to Numeric data type.
- Modify erroneous data, such as some special symbols for the alphabet.
- Elimination of blank spaces.

D. MODELING

When carrying out a detailed analysis of the data source to be used, it is concluded that the nominal data are numerical. Therefore, the decision was made that the most appropriate type of prediction would be regression.

For this reason, the comparison of some supervised machine learning techniques will be made for a regression-type model, among which we have:

- Decision Tree
- Random Forest
- Gradient-boosted
- Neural Network
- Support Vector Machine

To run the comparison of supervised machine learning techniques, we will carry it out using the RapidMiner Software.

RapidMiner is a data science platform that unifies data preparation, machine learning, and model operations, positioning itself within the “Magic Quadrant for Data Science and Machine Learning Platforms,” according to Gartner Inc in its evaluation of the year 2020.

Having a good preparation of our input data, these will be imported in a specific format; in this case, the type will be a CSV, whose format will allow our data to be represented in the form of a table where the columns are separated by commas and the rows by line breaks.

Figure 5 shows a plot of the data imported into the RapidMiner tool with a total of 19062 records and four attributes, including voltage, current, consumption, and SoC.

Figure 6 shows the modeling developed in Rapidminer for the machine learning neural network technique.

E. RESULTS OF SUPERVISED MACHINE LEARNING TECHNIQUES

For the RapidMiner platform in which different supervised machine learning techniques were modeled: Decision Tree, Random Forest, Gradient-boosted, Neural Network, and Support Vector Machine, data from the Victrom System Energy variables were used, which were a total of 19062 records. Of the data to model, 70% was used for training and 30% for testing.

Next, a table of the results obtained from the modeling that was executed on the Rapidminer platform for the different techniques that were previously chosen will be shown, using the CSV file as input data. Made up of the Energy variables.

Table 1 shows the Comparison of Results of the selected Supervised Machine Learning Techniques.

| Performance Vector | Decision Tree | Random Forest | Gradient Boosted | Scikit-Learn | SVM |
|--------------------|---------------|---------------|------------------|--------------|-----|
| RMSE               | 0.265         | 0.11          | 6.7              | 0.37         | 3.23|
| Squared Correlation| 0.99          | 1.00          | 1.00             | 0.99         | 0.98|
| Prediction Average | 90.12         | 90.12         | 90.13            | 90.12        | 90.12|
| Execution Time (sec) | 0.4          | 0.4           | 4                | 0.2          | 18  |

Table 1 shows that for the chosen techniques, the ones that gave the best results concerning the RMSE (Mean Square Error) were: Decision Tree, Random Forest, and Neural Network, whose values were the lowest, indicating a better fit to the estimated prediction. For the Squared Correlation and Prediction Average metrics, the algorithms gave similar values, some varying by tenths. However, for the execution time, a significant difference was observed between the algorithms, seeing that the Neural Network had the shortest time with 0.2 seconds of execution. Support Vector Machine was the algorithm that took the longest time in execution, with a total of 18 seconds.
Considering the results obtained and identifying that the nominal data used in the modeling is a Numerical type, we proceed with the choice of the Neural Network algorithm, which is the one that best suits the type of prediction that will be executed in this research, this with the support that neural networks create their interpretation of their information inside and are more robust to fault tolerance and flexible when the input data may present changes that are not so significant, such as noise, ANN can handle these changes properly [29] which, unlike algorithms such as Decision Tree and Random Forest, are decision trees that can present instability before any minimal change in the input data can lead to a totally different decision tree, decision trees decision although they are more specialized in categorical data types than in numerical data types and their ability predictive in a single tree is much lower than that achieved with other models, due to its tendency to overfitting and high variance [30].

F. SOFTWARE DEVELOPMENT

This final phase allows the development of the supervised machine learning technique chosen, the Neural Network, whose technique met the objectives set at the beginning.

For the development of the chosen technique, we will use the Python programming language and the Tensorflow library, and we will use the development of the Neural Network through the Google Collaboratory tool.

Google Collaboratory is a widely used tool in the machine learning community; with Colab, it can import an image dataset, train an image classifier on that dataset, and evaluate the model with just a few lines of code. Colab notebooks run code on Google’s cloud servers, allowing one to harness the power of Google hardware, including GPUs and TPUs, no matter how powerful the computer is. All needed is a browser.

1) DATA READING
The data will be loaded in a specific format; in this case, the type will be a CSV, where the parameters are Voltage, Consumption, Current, and SoC.

The CSV file will be imported through the Pandas read_csv() function and stored in the “dataframe” variable; a total of 19062 records will be imported.

2) DATA CLEANING
Then it is done with a data cleaning where the Pandas dataframe.dropna() function will be used, allowing different ways to analyze and eliminate rows and columns with null values.

3) DATA DIVISION
A new variable will be created with the name “train-dataset” The random fraction of the data will be stored in 80% for training and 20% for tests through the Pandas dataframe.sample() function.

traindataset = dataframe.sample(frac=0.8, random_state=0)

4) DATA INSPECTION
To visualize how the data is being related, the Seaborn library will be imported, a data visualization library for Python, and we will call the function pairplot. This function creates a grid of axes, executed as the following form sns.pairplot().

Figure 7 shows different graphs where the relationship of the Data Set between each variable is displayed: voltage, consumption, current, and SoC.

5) SEPARATION OF THE OBJECTIVE VALUE
The objective value or label will be separated from the characteristics.

The different target value or label is “SoC”, and its characteristics are “voltage”, “current” and “consumption”.

““SoC,”” this label is the value that the model will train to predict.

To carry out this separation, the pop() function was used, and it will be stored in a new variable named “train_stats”

Once the target value or label is separated from the dataset, its general statistics will be verified, which will be displayed in the following table and were obtained thanks to the describe() function.

In Figure 8, the statistical information must be known for good data analysis. First, the information of the objective value is shown where the number of samples, the average value, the standard deviation, the minimum, the maximum,
FIGURE 8. Statistical information of the target value using the `describe()` function.

FIGURE 9. Statistical information of the characteristics using the `describe()` function.

FIGURE 10. Normalized data.

and the median is displayed. And the values under the 25%, 50, and 75% percentiles.

With the statistical information of the target value, it is also required to review the general statistics of the characteristics (Voltage, Current, and Consumption). In Figure 9 now, the information of the characteristics are the voltage, the current, and the consumption will be shown where the number of samples, the average value, the standard deviation, the minimum, the maximum, the median, and values below the 25%, 50 and 75% percentiles.

6) NORMALIZATION
After reviewing the statistical information, it was identified how different the ranges of each characteristic are.

With the use of normalization, we will reduce the range of the data to have a single length and thus be able to facilitate the training of the neural network.

Figure 10 shows some random rows of normalized data for each characteristic (voltage, current, and consumption).

7) CONSTRUCTION OF THE NEURAL NETWORK
For the neural network construction, a sequential type model will be used with an input layer, an output layer, and an output layer that returns a single continuous value.

For the construction of the neural network, where KERAS was implemented, a high-level API to compile and train machine learning models. KERAS has two ways of building neural networks, and in this case, the sequential model (Kera.Sequential) was used. Once the type of model is defined, it will be determined how many layers will make up the neural network and what its activation functions will be for each layer, defining that there will be an input layer with four neurons and its “sigmoid” activation function, a hidden layer of 4 neurons with its “sigmoid” activation function, and an output layer with one output neuron with its “linear” activation function.

With the defined neural network, it must be completed by adding a cost function, an optimizer, and performance metrics, using the RMSPROP algorithm as an optimizer with a learning rate of 0.001 and the “mae, mse and accuracy” metrics, which measure the average magnitude of the errors in a set of predictions.

8) NEURAL NETWORK TRAINING
It will start with the training of the model, where we will define that the number of learning iterations will be 1000 (epochs) of training, and the training and validation accuracy will be recorded in the “History” variable.

For the training of the neural network where the `fit()` function was used, which receives the normalized data set (normenttrain), the set of expected results (temp_train), and the number of 1000 learning iterations (epochs) of training.

9) NEURAL NETWORK PREDICTION
We will take some of the normalized test data and proceed to predict the State of Charge (SoC).

Figure 11 shows the predict() function that allows the prediction of the labels of the data values based on the trained neural network model.

10) IMPORT THE MODEL IN TENSORFLOW.JS
To implement the model with its weights in a web computing module, the Tensorflow.JS library must be installed. When the library is imported, the model and its weights will be converted based on the Tensorflow.JS library and saved to a new location in a file with the JS extension.

With the Saved Model, it will continue to be imported into the JS code of the web computing module that will be implemented for this research work.

Within the code of the web computing module to be implemented, the model will be loaded along with its weights.

11) WEB COMPUTER MODULE WITH MODEL IMPORT IN TENSORFLOW.JS
Next, the Web Computing module will be shown to predict the State of Charge of the Batteries (SoC) that was
proposed in part to the specific objectives of this research work.

It is worth mentioning that only a new web computer module was integrated since the company where the research is being carried out already had its Web Platform. Only one more module was chosen to be integrated.

The Web Computing Module will have three Textbox to enter the data such as “Voltage,” “Current,” and “Consumption” which are fundamental for the prediction of the main objective of this research work; with the three entered data, a button is also observed with the name “Execute,” the mentioned button will oversee the analysis of the State of Charge prediction. Figure 12 shows the State of Charge (SoC) Web Computing Module interface.

The code link is at https://github.com/alexaxiomara22/State-of-Charge-Prediction.git

IV. RESULTS

In this section, the results obtained from the chosen supervised machine learning technique will be presented.

In addition to these results, the developed tests of the web computing module will be shown with the import of the model of the selected learning technique in Tensorflow JS.

A. TRAINING RESULTS OF THE NEURAL NETWORK MODEL

The TensorFlow library was used to find the best result for the parameters.

The number of learning iterations used in model training was 1000 (epochs); the results obtained are shown in the following image:

Figure 13 shows that iteration 282 achieves 100% hits for both the training and validation data. However, with the values of the Metrics “MAE” and “MSE” For each iteration of the neural network model training, the error decreases, thus giving a good response for the relationship between the input data and the results.

Figure 14 shows a graph with the error rate for the training data set (Train Error) and the error level for the validation data set (Val Error), executed based on the MAE metric for each iteration of the neural network model.

With good results in the training of the Neural Network model, then we will proceed with the evaluation of the model but with the test data set.

Figure 15 shows the predicted results using the test data set.

B. MODEL TRAINING RESULTS

In the training of the neural network model, the metrics that monitored the learning of the model were the “MAE” (Mean Absolute Error) and the “MSE” (“Mean Square Error”) where it was observed that the training error for each “epoch” it was lower and lower. In this way a good answer was obtained for the relationship between the input data and the results.

The neural network model must be evaluated, and if the training data is not evaluated correctly, it is possible that the results will not lead to misleading or incorrect conclusions. However, the “MAE” regression metric was previously chosen as it is the most robust and less sensitive to outliers, such as the mean square error. Therefore, the evaluation of the model was considered in four categories underfitting, overfitting, good fit, and unknown setting.

However, it was previously shown that the training and validation error approach each other as learning continues.
However, if it is observed that the training begins to be lower than the validation error, the model will begin to be overfitted.

So, to avoid this overfitting in the model, the “EarlyStopping” function was used, which will stop the training when the validation score does not improve, meaning that if a certain number of “Epochs” elapse without showing improvement, then the training stops automatically. As observed in the model’s training, 1000 “Epochs” were initially declared. However, the results only show up to 300 “Epochs,” concluding that after 300 “Epochs,” there will be no more improvements or degradations concerning the “MAE” metric. Finally, a final validation of the predictive data and the actual data of the target value “SoC” was made, where a good relationship between both data is shown, as well as the demonstration of a part of the data that was predicted and the comparison with the actual data the difference was minimal. The attributes that were used for this investigation were very relevant to arriving at good accuracy results because high precision was obtained for both the training data and the actual test data.

C. TESTING THE WEB COMPUTER MODULE WITH MODEL IMPORT IN TENSORFLOWJS

Actual data extracted from the Victron Energy System were used for the tests carried out in the web computer module, and the head of the ISM area validated the results. Next, the following tests carried out will be presented:

First test: with the following input data from a running computer:

– Voltage: 53.03
– Current: 31.3
– Consumption: 0

Figure 16 shows the prediction result that our equipment was 100% loaded.

Second test: with the following input data from a running computer:

– Voltage: 49.03
– Current: −12.3
– Consumption: −126.3

Figure 17 shows the prediction result that our team is loaded to 84%.

Finally, Figure 18 shows the relationship between the actual data of the target value “SoC” and the predictive data of the target value “SoC,” where it can be seen that the model predicts reasonably well, where the X axis is the positions, and the Y axis is the values.

V. DISCUSSION AND CONCLUSION

Our results show that the state of charge of batteries in photovoltaic systems in the mining sector could be predicted using supervised learning techniques in an acceptable manner, which was validated with the construction of the software to make the indicated prediction through the passing results in the tests carried out.

These results coincide with those found by Henri and Lu [31] when they state that a machine learning approach can predict and program the operation mode in real time of an operation interval for residential battery/photovoltaic systems. This is related to what was expressed by Henri, Lu, and Carreio [32], who state that a machine learning approach for predicting the optimal operating mode of the battery in real-time in residential photovoltaic applications. Rehman et al. [33] show that machine learning can investigate monitoring data and simulate and optimize case studies.

This study represents a significant contribution to the prediction of the state of charge of batteries; however, a limitation found in the study was not having had sufficient resources to have achieved a greater coverage of companies dedicated to this area, and although it was not the purpose of the study, it could have generated a baseline on the load status prediction of photovoltaic batteries.

The collection of all the fundamental data was obtained through the Victron Energy System, which has the function of monitoring the energy variables of the solar panels;
preparation of the data was carried out where it was appreciated that within the fields that were exported, only three would provide no value for the analysis regarding the state of charge, this observation is corroborated by the field engineer in charge of maintaining the equipment’s solar panels. However, it was considered to add one more field whose name would be state-of-charge [34], to which we would add a formula to be able to find the State of Charge, such as studies for the recovery of the energy storage system adaptive with the regulator ANFIS [35].

It was established that the most appropriate supervised machine learning technique would be Neural Networks, which were used in works such as [36] and other cases with Deep Neural Networks [37]. In the present study, an average prediction of 90.12% was found in an execution time of 0.02 seconds; it was also the one that best suited the type of prediction that was executed in this investigation since neural networks create their interpretation of their information inside and are more robust to fault tolerance and flexible when the input data may present changes that are not so significant. Despite this, some works use Non-supervised machine learning techniques for the electronic discovery of novel technology that presents detailed reviews and analyzes of solar energy technology [38].

The effectiveness of the training of the neural network model for the prediction of the state of charge of the Batteries was verified, observing that from iteration 282, 100% of successes are achieved for both the trained data and the validation data; however, the Metrics chosen in the Model Training were also evaluated, which in this case was the (“MAE”) for being the most robust and less sensitive to outliers such as the mean square error.

It was shown that the training and validation error approach each other as learning continues; To avoid overfitting, the model, the “EarlyStopping” function was used, which will stop training when the validation score does not improve, meaning that if a certain number of “Epochs” elapses without showing improvement, then the training stops automatically.

Thanks to the chosen automatic supervised learning technique; it was possible to predict the state of charge of the batteries by comparing the predictive data and the data in real-time, obtaining a good relationship and satisfactory results.

In addition, it was possible to import the model together with its weights from the neural network to the web computer module proposed in part for the specific objectives of the research, consisting of a form where the user can enter the data of the monitored energy variables and predict the state of charge that dumps your batteries.

In future work, it is recommended to expand the determining attributes, such as the temperatures of the batteries, among others, in order not only to be able to find the state of charge of the batteries but also to be able to find the life cycle of the batteries. As well as carry out the construction of the application of other regression algorithms to compare the results and check the efficiency, as well as Deep Learning. To identify if the model is presenting problems such as underfitting or overfitting, it is recommended to divide the input data set into two subsets with the division of training and evaluation and evaluation percentages varied.

Based on the results and the software tool built, it can be applied in companies dedicated to determining the state of charge of photovoltaic batteries and personnel interested in knowing their state.

REFERENCES

[1] S. M. S. BItar and F. B. Chamas, “Estudio de factibilidad para la implementación de sistemas fotovoltaicos como fuente de energía en El sector industrial de Colombia,” J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2013.
[2] S. Shalvey-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, vol. 9781107057135. Cambridge, U.K.: Cambridge Univ. Press, 2013.
[3] K. R. Pradeep and N. C. Naveen, “A collective study of machine learning (ML) algorithms with big data analytics (BDA) for healthcare analytics (HaA),” Int. J. Comput. Trends Technol., vol. 47, no. 3, pp. 149–155, May 2017.
[4] U. S. Shanthamallu, A. Spanias, C. Tepedelenliglu, and M. Stanley, “A brief survey of machine learning methods and their sensor and IoT applications,” in Proc. 5th Int. Conf. Inf., Intell., Syst. Appl. (HSA), Aug. 2017, pp. 1–8, doi: 10.1109/HISA.2017.8316459.
[5] A. A. Aziz and A. Starkey, “Predictive supervisor machine learning performances for sentiment analysis using contextual-based approaches,” IEEE Access, vol. 8, pp. 17722–17733, 2020, doi: 10.1109/ACCESS.2019.2958702.
[6] J. Sulla-Torres, C. I. Nina, M. R. Portugal, M. Cossio-Bolanos, and R. G. Campos, “Comparison of classification algorithms for the detection of bone weakness in students using anthropometric data,” in Proc. Int. Conf. Inclusive Technol. Educ. (CONTIE), Oct. 2019, pp. 56–566, doi: 10.1109/CONTIE49246.2019.00020.
[7] C. Möhring and M. Grotke, “Artificial intelligence in innovation: How to spot emerging trends and technologies,” IEEE Trans. Eng. Manag., vol. 69, no. 2, pp. 493–510, Apr. 2022, doi: 10.1109/TEM.2020.2989214.
[8] S. Samara and E. Natsheh, “Intelligent real-time photovoltaic panel monitoring system using artificial neural networks,” IEEE Access, vol. 7, pp. 50287–50299, 2019, doi: 10.1109/ACCESS.2019.2911250.
[9] H. Rahimi-Eichi, U. Ojha, F. Baronti, and M.-Y. Chow, “Battery management system: An overview of its application in the smart grid and electric vehicles,” IEEE Ind. Electron. Mag., vol. 7, no. 2, pp. 4–16, Jun. 2013, doi: 10.1109/MEI.2013.2250351.
[10] Z. Wang and D. Yang, “State-of-charge estimation of lithium iron phosphate battery using extreme learning machine,” in Proc. 6th Int. Conf. Power Electron. Syst. Appl. (PESA), Dec. 2015, pp. 1–5, doi: 10.1109/PESA.2015.7398906.
[11] L. Rodríguez, “Componentes de un sistema fotovoltaico,” España Univ. Grup. Idea. Energía Sol. Y Automática, Tech. Rep., 2004. [Online]. Available: http://www.uaemex.es/investiga/solar07/cursosolarat/.
[12] G. G. Kim, W. Lee, B. G. Bhang, J. H. Choi, and H.-K. Ahn, “Fault detection for photovoltaic systems using multivariate analysis with electrical and environmental variables,” IEEE J. Photovolt., vol. 11, no. 1, pp. 202–212, Jan. 2021, doi: 10.1109/JPHOTOV.2020.3032974.
[13] E. Cook, S. Luo, and Y. Weng, “Solar panel identification via deep semi-supervised learning and deep one-class classification,” IEEE Trans. Power Syst., vol. 37, no. 4, pp. 2516–2526, Jul. 2022, doi: 10.1109/TPWRS.2021.3125613.
[14] A. M. Karimi, J. S. Fada, M. A. Hossain, S. Yang, T. J. Peshek, J. L. Braid, and R. H. French, “Automated pipeline for photovoltaic module electromunecesence image processing and degradation feature classification,” IEEE J. Photovolt., vol. 9, no. 5, pp. 1324–1335, Sep. 2019, doi: 10.1109/JPHOTOV.2019.2920732.
[15] S. Theocharides, G. Makrides, G. E. Georghiou, and A. Kyprionou, “Machine learning algorithms for photovoltaic system power output prediction,” in Proc. IEEE Int. Energy Conf. (ENERGYCON), Jun. 2018, pp. 1–6, doi: 10.1109/ENERGYCON.2018.8398737.
[16] U. Ağbulut, A. E. Gürel, A. Ergün, and I. Ceylan, “Performance assessment of a V-trough photovoltaic system and prediction of power output with different machine learning algorithms,” J. Cleaner Prod., vol. 268, Sep. 2020, Art. no. 122269, doi: 10.1016/j.jclepro.2020.122269.
A. Apaza-Pinto et al.: Supervised Machine Learning Techniques for the Prediction of the SoC of Batteries

[17] R. Kabilan, V. Chandran, J. Yogapriya, A. Karthick, P. P. Gandhi, V. Mohanavel, R. Rahim, and S. Manoharan, “Short-term power prediction of building integrated photovoltaic (BIPV) system based on machine learning algorithms,” *Int. J. Photoenergy*, vol. 2021, pp. 1–11, Apr. 2021, doi: 10.1155/2021/5882418.

[18] J. Wang, P. Li, R. Ran, Y. Che, and Y. Zhou, “A short-term photovoltaic power prediction model based on the gradient boost decision tree,” *Appl. Sci.*, vol. 8, no. 5, p. 689, Apr. 2018, doi: 10.3390/app8050689.

[19] B. D. Dimd, S. Voller, U. Cali, and O.-M. Midgurd, “A review of machine learning-based photovoltaic output power forecasting: Nordic context,” *IEEE Access*, vol. 10, pp. 26404–26425, 2022, doi: 10.1109/ACCESS.2022.3156942.

[20] L. E. Ordoñez-Palacios, D. A. León-Vargas, V. A. Bucheli-Guerrero, and H. A. Ordoñez-Erase, “Predicción de radiación solar en sistemas fotovoltaicos utilizando técnicas de aprendizaje automático,” *Revista Facultad de Ingeniería*, vol. 29, no. 54, Sep. 2020, Art. no. e11751, doi: 10.19053/01211129.v29.n54.2020.11751.

[21] Z. Chen, F. Han, L. Wu, J. Yu, S. Cheng, P. Lin, and H. Chen, “Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents,” *Energy Convers. Manage.*, vol. 178, pp. 250–264, Dec. 2018, doi: 10.1016/j.enconman.2018.10.100.

[22] R. A. Campos, G. L. Martins, and R. Rüther, “Aplicación de máquina learning para una previsión de temperatura de módulos solares fotovoltaicos,” in *Proc. VIII Congresso Brasileiro de Energia Solar*, 2020.

[23] D. N. Karamov and K. V. Suslov, “Structural optimization of autonomous photovoltaic systems with storage battery replacements,” *Energy Rep.*, vol. 7, pp. 349–358, Apr. 2021, doi: 10.1016/j.egyr.2021.01.059.

[24] M. A. Abella and F. C. Romero, “Dimensionado de sistemas fotovoltaicos,” *Dinest. Sist. Fotovoltaicos*, vol. 1, no. 1, pp. 1–73, 2005.

[25] L. Ma, C. Hu, and F. Cheng, “State of charge and state of energy estimation for lithium-ion batteries based on a long short-term memory neural network,” *J. Energy Storage*, vol. 37, May 2021, Art. no. 102440, doi: 10.1016/j.est.2021.102440.

[26] J. Tian, R. Xiong, J. Lu, C. Chen, and W. Shen, “Battery state-of-charge estimation amid dynamic usage with physics-informed deep learning,” *Energy Storage Mater.*, vol. 50, pp. 718–729, Sep. 2022, doi: 10.1016/j.enconman.2022.06.007.

[27] J. Tian, R. Xiong, W. Shen, J. Lu, and F. Sun, “Flexible battery state of health and state of charge estimation using partial charging data and deep learning,” *Energy Storage Mater.*, vol. 51, pp. 372–381, Oct. 2022, doi: 10.1016/j.enconman.2022.06.053.

[28] P. Chapman, “CRISP-DM-cross-industry standard process for data mining-1.0 step-by-step data mining guide,” *Cris. Consort.*, 2000, [Online]. Available: https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-9685f1bd1a20

[29] J. A. Rodrigo, (2020). *Arboles de Decisión, Random Forest, Gradient Boosting Y C5.0*. [Online]. Available: https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting

[30] G. Henri and N. Lu, “A supervised machine learning approach to control energy storage devices,” *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 5910–5919, Nov. 2019, doi: 10.1109/TSG.2019.2892586.

[31] G. Henri, N. Lu, and C. Carrejo, “A machine learning approach for real-time battery optimal operation mode prediction and control,” in *Proc. IEEE/PES Transmiss. Distrib. Conf. Expo. (T&D)*, Apr. 2018, pp. 1–9, doi: 10.1109/TDC.2018.8440141.

[32] H. U. Rehman, T. Kordba, R. Abdurafikov, T. Laakko, A. Hasan, and F. Reda, “Data analysis of a monitored building using machine learning and optimization of integrated photovoltaic panel, battery and electric vehicles in a central European climatic condition,” *Energy Convers. Manage.*, vol. 221, Oct. 2020, Art. no. 113206, doi: 10.1016/j.enconman.2020.113206.

[33] U. Datta, A. Kalam, and J. Shi, “Battery energy storage system control for mitigating PV penetration impact on primary frequency control and state-of-charge recovery,” *IEEE Trans. Sustain. Energy*, vol. 11, no. 2, pp. 746–757, Apr. 2020, doi: 10.1109/TSTE.2019.2904722.

[34] H. Jagadeesh and K. V. Kumar, “Mitigating PV penetration effect on primary frequency control and state-of-charge recovery by utilizing ANFIS controller,” in *Proc. Symm. Power Electron. Renew. Energy Systems Control*, vol. 616, 2021, pp. 239–248, doi: 10.1007/978-981-16-1978-6_21.

[35] S. Rao, G. Muniraju, C. Tepedelenlioglu, D. Srinivasan, G. Tamizhmami, and A. Spanias, “Dropout and pruned neural networks for fault classification in photovoltaic arrays,” *IEEE Access*, vol. 9, pp. 120034–120042, 2021, doi: 10.1109/ACCESS.2021.3108684.

[36] F. Succetti, A. Rosato, R. Araneo, and M. Panella, “Deep neural networks for multivariate prediction of photovoltaic power time series,” *IEEE Access*, vol. 8, pp. 211490–211505, 2020, doi: 10.1109/ACCESS.2020.3039733.

[37] A. J. C. Trappey, P. P. J. Chen, C. V. Trappey, and L. Ma, “A machine learning approach for solar power technology review and patent evaluation analysis,” *Appl. Sci.*, vol. 9, no. 7, p. 1478, Apr. 2019, doi: 10.3390/app9071478.

ALEXA APAZA-PINTO was born in Arequipa, Peru. She is a System Engineering with a specialization in information system with the Universidad Catolica de Santa Maria (UCSM), in 2021. She managed to have public congratulations from the Universidad Catolica de Santa Maria for supporting her thesis “Prediction of the state of battery charge for photovoltaic systems in the mining sector using supervised machine learning techniques.” She is certified in Scrum Master Professional (Eng.) SMPC® and a member of the College of Engineers of Peru.

JOSE SULLA-TEJADA (Member, IEEE) was born in Arequipa, Peru. He received the M.S. degree in strategic telecommunications management from the Miguel de Cervantess European University, in 2013, the degree in systems auditing and information security from the Universidad Catolica de Santa Maria (UCSM), in 2019, and the Ph.D. degree in environmental sciences and renewable energies from the Universidad Nacional de San Agustin de Arequipa, in 2022. He is currently a Systems Engineer with UCSM, in 2008, where he is currently an Assistant Professor at the Faculty of Physical and Formal Sciences and Engineering and a Seedbed Coordinator. His research interests include the Internet of Things (IoT) and educational technology.

PATRICIA LÓPEZ-CASAPERALTA, photograph and biography not available at the time of publication.

JOSE SULLA-TEJADA (Member, IEEE) was born in Arequipa, Peru. He received the B.S. and M.S. degrees in system engineering from the Universidad Catolica de Santa Maria, Arequipa, in 2004, and the Ph.D. degree in computer science from the Universidad Nacional de San Agustin, Arequipa, in 2018, with second specialty in software engineering. Since 2018, he has been a Renacyt-Concytec Researcher in Peru. He leads basic and applied research projects with lines of research in information technology and applied artificial intelligence in health and sport sciences. He is Full Research Professor with the Universidad Catolica de Santa Maria, Peru.