Comparison of Machine Learning Algorithms for the Power Consumption Prediction

- Case Study of Tetouan city –

Khoi Nguyen Bui
Bachelor of Computer Science
Deakin University
khoinguyenbui2004@gmail.com

Abstract— This research will estimate the consumption of power electricity by the city of Tetouan, basing on ten-minute data and an hour that will go further in improving the operational efficiency and effectiveness of the utility systems. The study uses different types of machine learning models that are particularly robust in terms of prediction accuracy. The using models include a Feedforward Neural Network with Backpropagation algorithm, Random Forest, Decision Tree, and a Support Vector Machine for regression (SVR) model with a radial basis function kernel. All parameters of each model were fine-tuned using the Grid Search method to yield the best performance metrics.

The dataset used consisted of data on three different networks of power distribution in Tetouan, made throughout the year 2017, taken from Supervisory Control. Preliminary results from the models, judged by the value of the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), revealed the Random Forest model is always superior, as it scores lower prediction errors than other models. However, the Random Forest model significantly improved after performing hyperparameter tuning on both the training and testing sets. It particularly brought out the aspect that this model is the best in grasping the complexities associated with the dataset compared to the rest. This report, therefore, undertakes the task of a comparative study of the same and makes detailed insights into their efficacy and practical applicability in urban power management scenarios.

I. INTRODUCTION

As urban sizes continue to increase, the management of energy in a city becomes ever more complex. Efficient energy forecasting could be of great help to the urban planning in such a way that overproduction and undersupply of energy could finally belong to the past. The city of Tetouan, with different

temporal consumption patterns because of demographics and industrial activities, provides an excellent case for applying advanced machine learning techniques in predicting power usage. In this report, applications of different predictive models are tested and evaluated over their performance under optimized conditions.

II. OVERVIEW OF MACHINE LEARNING ALGORITHM

We will apply 5 models and then compare the performance of each models to find the best one that is suitable for this dataset.

A. Random Forest

The Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees. Particularly it is very effective for complex datasets with many variables, such as power consumption data. However, Random Forest helps relieve the overfitting problem through an average of the predictions taken from many trees, and this becomes robust to noise and outliers within the data. Such hyperparameters include the number of trees (n_estimators), maximum depth of trees (max_depth), and minimum samples required to split a node (min_samples_split).

B. Decision Tree

Decision Trees are very basic and powerful methods to model the nonlinear relationship of the features with the target. Basic principle: It works by successively partitioning the data into subsets according to input features' values and hence may represent a sort of tree structure of decision rules. On the other hand, Decision Trees are susceptible to overfitting problems, more so when the tree grows too deep, or the data required for training is not adequate. Some other hyperparameters like the maximum depth (max_depth), minimum samples required at a leaf node (min_samples_leaf), and minimum samples required to split an internal node (min_samples_split) are all helpful in controlling overfitting.

C. Support Vector Machine Regression(SVR)

Support Vector Machine Regression is a powerful algorithm and one of the best suited for this task, since it can handle both high-dimensional spaces. The way this model works is that it maps input data into a higher dimensional feature space, where a hyperplane is built to have a maximum margin among different classes or, in the case of SVR, try fitting the regression line. SVR performance mainly depends on the type of kernel function (linear, polynomial, radial basis function) and the value of its regularization parameter (C). SVR works well in the complex data set where the non-linear relationship exists, but hyperparameters should be properly tuned for better performance.

D. Linear Regression

Linear Regression is one of the simplest and most popular algorithms for predicting quantitative outcomes. It presupposes an adequate type of relation expectable in those cases between input features and the target variable. Linear Regression is computationally efficient and very human-explainable; however, it may not capture more complex patterns in the data. In power consumption prediction, Linear Regression can provide baseline performance and even explain some underlying linear relationship insight between the input features and power consumption.

E. Artificial Neural Networks

Multi-layer Perceptron Regressor and belongs to a neural network. It learns very complex patterns and relationships in the data through a number of layers of linked nodes (neurons) that process input data, holding a set of weighted connections along with nonlinear activation functions. MLPRegressor is excellent in capturing intricate patterns in data, although it requires a larger amount of data and computational resources for training, especially for deep architectures. It offers architecture flexibility, from the number of hidden layers and neurons per layer to activation functions, and it provides optimal algorithms, like Adam or Stochastic Gradient Descent.

III. CASE STUDY

Located in the north of Morocco, Tetouan city is divided by the geographical and climatic diversity of its lands. These, in turn, set direct influences on the overall electricity use patterns pervading in the whole region. The city itself divides into three main zones of primary power distribution: Quads, Boussafou, and Smir. Due to the specifics described above, the activity in residential and industrial zones suggests different profiles of power consumption. Understanding and predicting the trends outlined above gives grounds for effective optimization of energy distribution and planning for future energy needs adequately.

This case study was based on the prediction of the power consumption time series of the city of Tetouan at fine granularity using machine learning techniques. This will not only aid in effective energy management but also help understand the dynamics of power use across the various zones of the city. This study aims to identify the model that provides the best prediction of electricity consumption with accurate values from various machine learning models used in the literature and to evaluate the effect of optimized model parameters on the accuracy of the prediction.

In order to see the distribution of each zone, we will use a pairplot to visualise a summary of the relationships and distributions between the total power consumption and the power consumption of each zone, which is very helpful to see the trend and distribution between them.

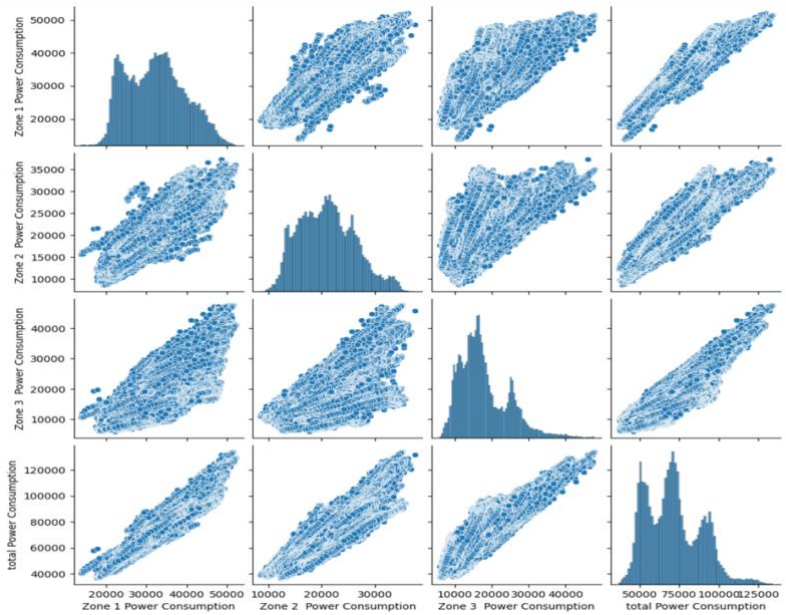

Fig 1. Relationships between total and zone power consumption.

Diagonal plots display the spread and central tendency in power consumption of all the variables both across different zones and overall. On the other hand, off-diagonal plots shows the relationship of pairs of different power consumptions. As expected, a great deal of positive association with correlation

was observed in the off-diagonal plots of power consumptions from different zones against the total power consumption.

High correlation coefficients close to 1 to total power consumption exist for power consumption in each zone. It, therefore, means changes in power usage by any single zone significantly influence the overall power consumption. The zones also depict very strong correlations with each other, meaning the driving factors in power usage are likely either mostly the same across all zones or that the zones are predominantly affected by the same external factors, such as time of day or weather conditions.

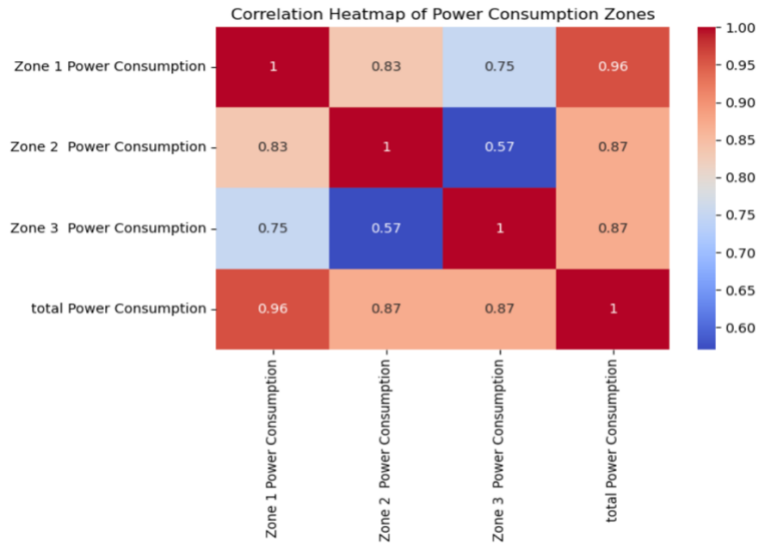

Fig 2. Correlation Heatmap of Power Consumption Zones

In order to detect outliers, we will use a histogram to visualise the value of each feature, which makes our model less likely to bias with a higher number.

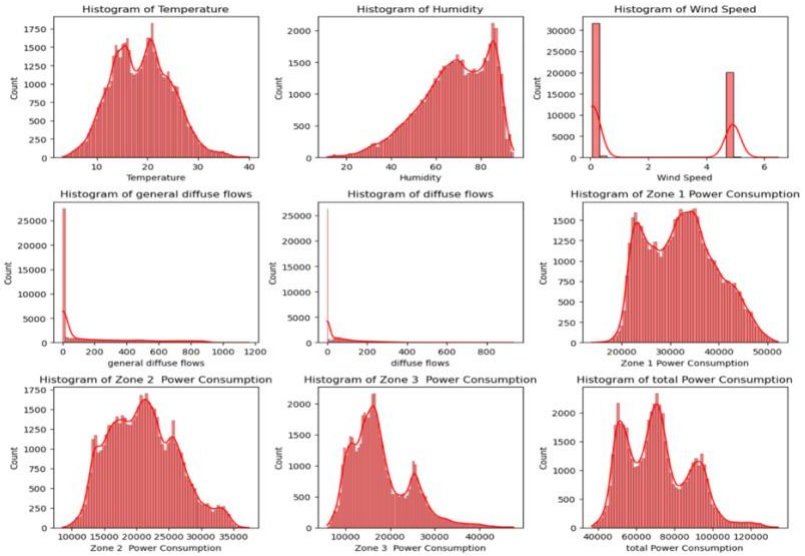

Fig 3. Evaluate the Distribution value of features

Both temperature and humidity will tend to be a little normally distributed, while wind speed will lean substantially toward low values. General Diffuse Flow and Diffuse Flow show strong skewness, which could be indicative of any particular environmental condition where higher values seem rare. Power Consumption in Zone 1, Zone 2, Zone 3, and Total show that all power consumption variables are highly skewed toward the higher values, which would imply raised demand energy levels.

The paper data uses historical data retrieved from the SCADA (Supervisory Control and Data Acquisition) system of the power distribution networks in Tetouan. It includes several records measured by every ten minutes over the period of one

year (2017). The dataset consists of several features: power consumption for each zone, related environmental data—like temperature, humidity, and wind speed, which are recognized as part of the elements that affect power distribution.

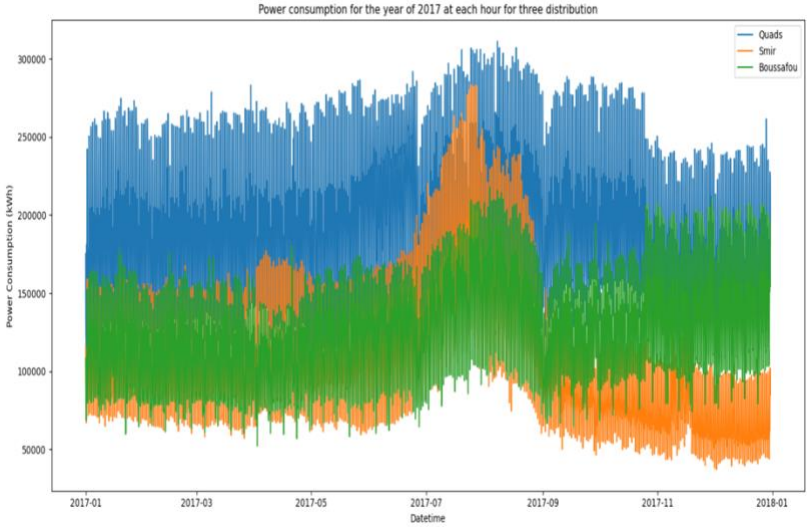

Fig 4. Power consumption for the year of 2017 at each hour for three distribution

Seasonal variation: All year energy consumption increases are experienced from June to August in the zones, most probably due to higher temperatures and perhaps increased tourist activity, considering it is the holiday period. On the other hand, While Quads and Smir both show a drop in the end consumption of these months, Boussafou has an unusual hike, likely indicating different patterns or outside factors taking place in that area.

The prediction models used are month, hour, day of month, day of year, week of year, and so on. In this, these characteristics act as independent variables in the prediction model, and their correlated relation with the dependent variable which is power consumption has been studied to know their impact on the prediction accuracy.

Fig 5. Evaluate the Time Distribution value of features

The dataset contains aggregate power consumption data without specifics of the types of buildings. The way power consumption is affected presents a weekday and weekend pattern but with different kinds of influences, seemingly due to changes in people's behavior. From statistics, power consumption from households shows minimal or no increase on weekdays, and gradually, the trend in factories, commercial, and public establishments is increasing. On the contrary, during weekends, household consumption increases, while the usage in other sectors reduces. This trend of power consumption, as denoted by picture below shows that the consumption of power on Sundays is slightly lower than other days. It is this variation that makes it important for day-of-week effects to be accounted for while accurate prediction of power consumption.

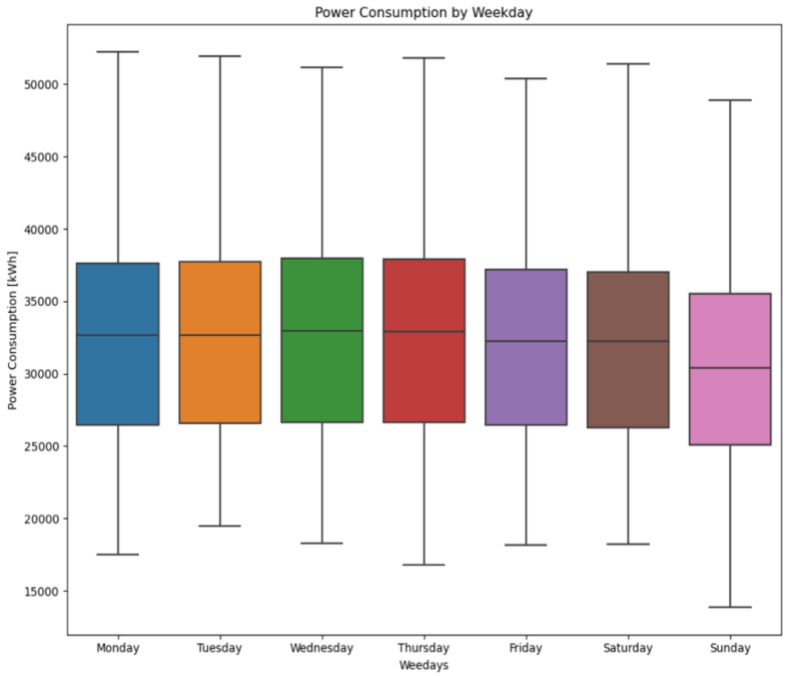

Fig 6. Box plot comparison of electricity consumption among week days

After checking human behaviour on the day of the week, which contributes to the power of onsumption, We are checking how weather affects humans using electricity. The weather in this case includes temperature, humidity, wind speed, diffuse flows, and global diffuse. The picture below show the correlation between the input variables and the output variable.

TABLE 1. WEATHER PROPERTIES AND COEFFICIENT OF CORRELATION BETWEEN THE INPUT VARIABLES AND THE OUTPUT VARIABLE

	Count	Mean	STD*	Min	Max	Correlation
Quads	52416	32344.9	7130.5	13895.6	52204.3	1
Temperature	52416	18.81	5.81	3.24	40.01	0.440221
Humidity	52416	68.25	15.55	11.34	94.8	-0.287421
Wind Speed	52416	1.95	2.34	0.05	6.48	0.167444
Diffuse flows	52416	75.02	124.2	0.011	936.00	0.080274
Global Diffuse	52416	182.69	284.4	0.004	1163.00	0.187965

*STD: Standard derivation

from the picture above, we can see that temperature has a strong positive correlation with power consumption in the Quads area, which can reflect more use of heating or cooling systems. In contrast, humidity has an inverse correlation, which means there could be less need for heating or cooling systems when there is higher humidity. Both flows of general and diffuse solar show a positive moderate correlation, possibly suggesting that slight influences due to solar conditions are found on power needs.

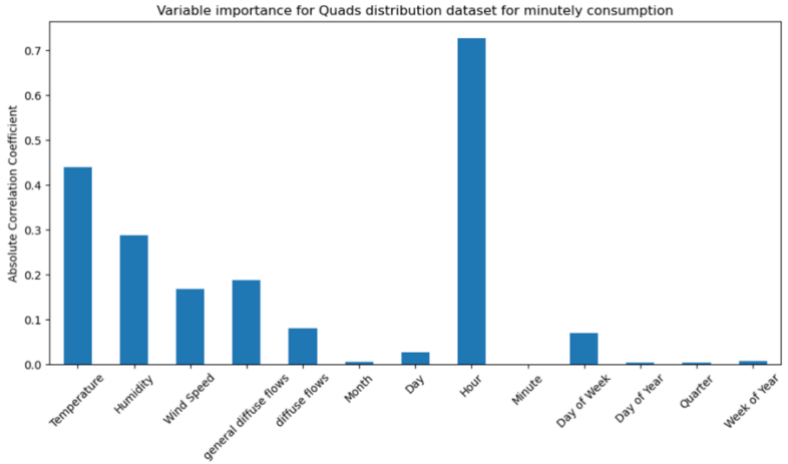

Fig 7. Variable importance for Quads distribution dataset for minutely consumption

We presented the correlation between power consumption and various calendar and weather attributes. Additionally, we conducted feature selection to determine the importance of predictive variables and eliminate irrelevant features. Temperature has the most correlation, and it shows that it has a pretty effective impact on the power consumption of the Quads area. Clear from the output is the indication that the relativity between the fluctuation of temperature and changes in energy consumption is close and most likely occasioned by heating or cooling demands. General diffuse flows and humidity also show notable correlations. In general, the diffuse solar flows may indicate the impact of solar heating or cooling loads, and the impact due to humidity is negative and could be associated with comfort levels or the effectiveness of the heating and cooling systems. Wind Speed and Diffuse Flows both have smaller but still significant correlations, thus contributing to the variation in power consumption but being far less critical than temperature or humidity.

However, for day of the year, months, quarters, and weeks of the year, So, we will remove it in order to save memory and enhance execution time. The minute feature is less important but will be kept as we will be using it to analyse data in the future.

IV. METHODOLOGY

A. Best Performing Model Pre-Hyperpameter Tuning

This table will evaluate several machine learning algorithms applied on Tetouan City power distribution data to pursue effective energy management and predictive accuracy. Comparative analysis is carried out for all models based on the performance of RMSE and MAE without hyperparameter tuning.

Algorithm	**Quads Distribution**				**Smir Distribution**				**Boussafou Distribution**				**Aggregated Distribution**			
	RSME		**MAE**		**RSME**		**MAE**		**RSME**		**MAE**		**RSME**		**MAE**	
	Train	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
RF	459.88	**1222.5**	296.03	**798.5**	493.94	**1350.4**	311.3	**847.8**	417.1	**1128.2**	270.2	**734.8**	972.8	**2622.4**	605.01	**1634.3**
DT	3.17e-14	1786.3	2.77e-16	1004.9	0.0	2023.33	0.0	950.7	0.0	1745.9	0.0	933.6	0.0	3895.9	0.0	2074.6
SVR	6054.09	6071.4	4913.2	4936.0	6262.5	6307.9	4572.03	4631.1	4400.2	4412.9	3497.4	3500.3	16120.8	16202.8	12770.1	12868.7
FFNN	5218.7	5236.1	4127.3	4143.4	5024.5	5038.8	4080.6	4110.2	3803.7	3827.5	2987.04	3013.2	15894.1	15956.1	12557.6	12635.8
LR	4332.25	4321.2	3433.6	3419.3	5024.5	5038.8	4080.6	4110.2	3552.0	3562.4	2806.1	2812.5	10624.8	10640.6	8346.6	8394.1

- *Decision Tree*

A common indicative, pointing across the board to overfitting, is observed to be RMSE and MAE close to zero for the training data but orders of magnitude higher for the test dataset. The Decision Tree algorithm, however, simple and easy for explaination, showed the performance problem. This was due to a great tendency to memorize the training data, hence

having poor performance when making predictions for unseen data.

- *Support vector machine for Regression(SVR)*

Since the RMSE and MAE values remain pretty high and distributed, it has to be said that the kernel and hyperparameters defaults were not apt for this given dataset. The problem with SVR models, especially with the radial basis function kernel, lies in the need for an exact and correct tuning of the hyperparameters to effectively cope with the complex nonlinear patterns present in complicated datasets, like those of power consumption data.

- *Artificial Neural Networks*

The FFNN generally does bad with a high RMSE and MAE values due to the initially selected architecture and learning rate hyperparameter fit for this problem. With such a large variability in performance due to the complexity of the FFNN and its training process, it is certainly clear that the network could not have reached the correct balance without failure. This could mean that the network was not complex enough to model the data with exact precision or it overfitted the data, which again points to the fine-tuning done with the very best judgment for results.

- *Linear Regression*

Similar to Decision Tree, Support Vector Machine, and Feedforward Neural Network. Linear Regression also have bad perforamnce metric in both RMSE and MAE because the model is very basic, but out features is very complicated. Overall, the power consumption pattern are too complex for a simple linear model to capture.

- *RandomForest*

The Quads distribution had a really good performance, characterized by the lowest error margins of test RMSE and MAE. Other good performances are from the Smir and Boussafou distributions, displaying regular consistency in generalization ability of different metrics. The Aggregated Distribution keep the lowest error metrics, reaffirming its robustness in handling complex, aggregated data. Overall, the Random Forest algorithm performs the best with all tested distributions. On the other hand, the ensemble approach of aggregating the results of many decision trees does very well in balancing bias with variance, and hence it makes more accurate predictions.

B. Best Performing Model Hyperpameter Tuning

The value of power consumption and parameters would be changed, and the optimization should be done for the comparative algorithms again through the same optimizer method (grid search) by the same sets of parameters. The following table represents optimized parameters for each comparative model and, of course, the difference from one distribution to the other.

TABLE 2. OPTIMIZING COMPARITIVE MODEL PARAMETERS FOR EVERY MINUTE POWER CONSUMPTION BY USING GRID SEARCH

Model	**Quads Distribution Parameter**	**Smir Distribution Parameter**	**Boussafou Distribution Parameter**	**Aggregated Distribution Parameter**
RF	max_depth=None, min_samples_split=2, n_estimators=150	max_depth=None, min_samples_split=2, n_estimators=150	max_depth=None, min_samples_split=2, n_estimators=150	max_depth=None, min_samples_split=2, n_estimators=150
DT	max_depth=None, min_samples_leaf=5, min_samples_split=2	max_depth=None, min_samples_leaf=1, min_samples_split=5	max_depth=None, min_samples_leaf=5, min_samples_split=2	max_depth=20, min_samples_leaf=1, min_samples_split=10
SVR	C=10, gamma=1	C=10, gamma=1	C=10, gamma=1	C=10, gamma=1
FFNN	Hidden_layer_sizes=150, learning_rate_init=0.01, max_iter = 300	Hidden_layer_sizes=100, learning_rate_init=0.1, max_iter = 100	Hidden_layer_sizes=150, learning_rate_init=0.01, max_iter = 300	Hidden_layer_sizes=150, learning_rate_init=0.01, max_iter = 300

TABLE 3. OPTIMIZING COMPARITIVE MODEL PARAMETERS FOR EVERY MINUTE POWER CONSUMPTION BY USING GRID SEARCH

Algorithm	**Quads Distribution**				**Smir Distribution**				**Boussafou Distribution**				**Aggregated Distribution**			
	RSME		**MAE**		**RSME**		**MAE**		**RSME**		**MAE**		**RSME**		**MAE**	
	Train	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
RF	454.09	**1219.67**	293.3	**795.5**	490.01	**1351.4**	309.3	**848.2**	411.9	**1125.6**	267.6	**733.3**	960.8	**2623.9**	600.4	**1635.1**
DT	958.6	1756.2	596.1	1090.5	422.1	2005.05	172.74	968.3	926.3	1708.2	565.9	1040.5	1772.7	3939.2	1050.7	2243.8
SVR	4377.3	4375.3	3396.1	3402.0	5265.4	5301.9	3953.4	3996.3	3488.04	3496.4	2722.6	2726.1	12324.4	12388.3	9329.4	9425.2
FFNN	4332.8	4322.1	3431.9	3418.3	4036.2	4047.3	3177.2	3186.5	3552.07	3562.2	2806.4	2812.6	10625.3	10642.6	8342.4	8391.2
LR	4332.25	4321.2	3433.6	3419.3	5024.5	5038.8	4080.6	4110.2	3552.0	3562.4	2806.1	2812.5	10624.8	10640.6	8346.6	8394.1

The table above gives the results of the models under each distribution, followed by the aggregation of the three distributions after the application of hyperparameters. Although all models having been decreasing in their performance, RandomForest still the best models in Power Consumption Dataset.

CONCLUSION

The results of this study revealed that the Random Forest algorithm is the best performing machine learning model to predict power consumption within Tetouan city with time and temperature were found to be the most significant. Further confirmation of its superiority in forecasting accuracy was found to be true with optimization through Grid-search. The future directions include an extension of this approach to wider Moroccan energy sectors and economic impact evaluation from precise energy forecasting, which will result in sustainability and efficiency in power management systems.