RF energy modelling using machine learning for energy harvesting communications systems

Youjie Ye1 | Freeha Azmat2 | Idris Adenopo1 | Yunfei Chen1 | Rui Shi3

1School of Engineering, University of Warwick, Coventry, UK
2Warwick Manufacturing Group, University of Warwick, Coventry, UK
3Information & Telecommunication Branch, State Grid, Beijing, China

Correspondence
Youjie Ye, School of Engineering, University of Warwick, Coventry, U.K. CV4 7AL.
Email: Youjie.Ye@warwick.ac.uk

Summary
Machine learning (ML) theories and methods are mainly based on probability theory and statistics. It is a very powerful tool for data modelling. On the other hand, energy harvesting has been regarded as a viable solution to extending battery lifetime of wireless sensor network. Motivated by these, modelling of the radio frequency (RF) energy available to the wireless nodes is required for efficient operation of wireless networks. In this work, we will use different ML algorithms to model the RF energy data for efficient operation of energy harvesting communication systems. Four ML algorithms are studied and compared in terms of the accuracy for RF energy modelling using the energy data in the band between 1805 and 1880 MHz. The results show that linear regression (LR) has the highest accuracy and the most stable performance, while decision tree is the worst model. Also, in terms of the operation efficiency of the system, LR has the best performance, followed by support vector machine and random forest algorithm.

KEYWORDS
energy harvesting, machine learning, modelling, prediction algorithms, radio frequency

1 | INTRODUCTION

Wireless sensor network (WSN) consists of independent sensing nodes that are connected via wireless links over short distances. The sensing nodes are often small, multifunctional and of low power consumption. They are usually managed by a central controller to collect data in a specific area. In almost all wireless sensing nodes, the power unit is the most important unit of all the components. If the battery runs out, none of the other units in the sensor can operate. For this reason, we require perpetual operation of wireless nodes, and one such technique is to harvest energy from the ambient environment to extend the battery lifetime. Using the energy harvested, sensor nodes can be more independent of battery. Meanwhile, the burden of batteries can be reduced significantly, and the lifetime of nodes can be extended greatly.

Energy harvesting is a process in which energy is derived from external sources, captured and stored for small wireless devices, such as wearable electronics and wireless sensing nodes. Solar energy has wide availability and high energy density. Hence, it has been widely used in WSN to charge batteries. There have been works on the modelling of solar energy, but in general, solar energy is limited by the time and weather conditions. For example, on cloudy days, rainy days or at night, the amount of available solar energy is very limited. On the other hand, the ambient radio
frequency (RF) energy has many sources due to the recent development of wireless systems. The ambient RF energy is relatively stable and provides a cost-effective solution as substitutes for batteries. Hence, RF energy harvesting can be utilized in power units of wireless sensor nodes. Also, RF energy harvester is easier to be integrated into the sensor nodes. For these reasons, ambient RF energy harvesting has been widely used in WSNs with lower consumption. In Olgun et al., an ambient RF energy harvester was designed based on the requirements for wireless devices. By harvesting and converting the 2.45 GHz Wi-Fi signal energy for 20 min, the maximum current of 20 μA can be achieved to make the temperature and humidity LCD display work for 10 min continuously. In Vyas et al., a battery-free embedded sensor platform was designed, which used the ambient RF energy from the wireless digital TV as the power supply. It can successfully provide power for a 16-bit embedded sensor microcontroller and keep it working. In Shigeta et al., RF energy harvesting was optimized by using adaptive work cycle control technology. After improving the sensing rate and efficiency, the sensor was provided with an average voltage of 2.68 V at the height of the 11th floor, 6.3 km away from the Tokyo TV transmission tower. In addition to these works, many other RF sources can be harvested for low power wireless devices. Thus, ambient RF energy has been proved to be a reliable power in the application of WSNs.

For RF energy harvesting, the power between antenna and rectifier is maximized by the impedance matching circuit when it runs at a specific frequency, after which the RF energy is converted to DC through diodes in the rectifier circuit and the DC voltage is smoothed in the capacitor. Although the energy harvested from RF is reliable, it is still randomly fluctuating, due to the random channel and operational conditions. This means that, during the periods of low energy, the sensing node will not be able to receive or transmit data properly due to insufficient energy. Moreover, to extend the lifetime of the WSN, it is necessary to keep the power consumption of the energy harvester to a minimum level. For these reasons, we can set a threshold for the energy harvester as the sensitivity of the harvester. If the predicted energy is lower than the threshold, the energy harvester will sleep to save activation energy. Otherwise, it will start to harvest energy for data transmission. Therefore, it is valuable to predict the pattern of the ambient RF energy and use this prediction to take mitigating measures to ensure the efficient operations of the sensing nodes.

Several measurement campaigns have been conducted to study the pattern of ambient RF energy. For example, In Kim et al., various ambient energy-harvesting technologies were reviewed and the applicability of ambient RF energy harvesting as an enabling technology for various self-sustaining wireless platforms was verified. In Chen et al., the average, the probability density function and the cumulative distribution function of harvested RF energy using linear and non-linear models for the energy harvester were derived to optimize the power transfer strategy. In Piñuelas et al., a survey of 270 underground stations in London was conducted to investigate the potential availability for ambient RF energy harvesting within urban and semi-urban environments. In Azmat et al., wireless data were analysed by setting a threshold to assume the occupancy of a particular band and comparing the classification accuracy of five machine learning (ML) algorithms (decision trees [DT], support vector machine [SVM], fire fly, hidden Markov model and naive Bayesian). In Long and Chen, two energy harvesting communications protocols, ‘harvest-store-use’ and ‘harvest-use’, were used to optimize the effective throughput of energy harvesting devices.

On the other hand, ML methods have been developed for many applications in communication recently. For example, it can be used to achieve human-computer interaction. In general, ML can be divided into two categories: supervised learning and unsupervised learning. Supervised learning builds an optimal model which will be used to predict future results from new data set based on the existing training data set. Unsupervised learning is used in applications where there are no labels for data, or no certain results. Very few works have considered the use of ML in the modelling of RF energy. In Abuzaib et al., an unsupervised Bayesian learning method was proposed to model the transmission power computed in each time slot at the hybrid access point of wireless RF energy harvesting networks to reduce the drop rate. In Zou et al., optimal sleeping and harvesting policies for RF energy harvesting devices were developed as a Bayesian adaptive Markov decision process based on knowledge of energy arrival from energy modelling. None of these works has considered the use of different ML algorithms as it is well known that different ML algorithms are suitable for different data sets. Azmat et al. studied the modelling of occupancy using ML. Occupancy is a binary quantization of the RF energy but not the RF energy itself, and hence, it can not be used to implement the optimal control of harvesting as in Zou et al. In Liang and Yuan, a kernel-density-based statistical model for the mobile service channels was proposed to predict RF energy. The sampling frequency was adjusted according to the channel power prediction. The accuracy of the model is higher than 80%.

Although works have been done to optimize RF energy harvesting, there has been no model for the prediction of time-series RF energy data. To achieve a balance between complexity and task performance, we need an effective model to make RF energy prediction. Generally, ML is a powerful tool to model or predict patterns from data. Motivated by the above observations, in this work, we will use different ML algorithms to develop predictive models for the amount
of ambient RF energy harvested. The ambient RF energy data are acquired from a measurement campaign performed at a university campus. Four supervised algorithms will be explored to build accurate predictive models, including linear regression (LR), SVM, random forest algorithm (RFA) and DT, to model the amount of available RF energy, not the occupancy. All of those algorithms are supervised learning. Based on the best performance of each model in the previous step, a suitable threshold will be set for energy harvester. The threshold will allow system designers to determine when the energy harvester should be activated to harvest energy and when it should go to sleep to save energy. The prediction accuracies of the models are compared, and recommendations are made on the most appropriate model to use.

In Section 2, the data set and pre-processing will be introduced. In Section 3, the selection of feature length (FL), number of observations, training split and learning algorithms will be explained. In Section 4, we will discuss the results on the prediction for the RF energy data. Section 5 is the conclusion.

2 DATA PRE-PROCESSING

2.1 Description of data

The RF energy data in this work were captured for a period of 4 months inside a research laboratory on the University of Warwick campus. The equipment is the Cambridge Radio Frequency Services (CRFS) node. CRFS is a cutting-edge RF designer, whose nodes focus on real-time 24/7 and cost-effective RF spectrum monitoring. The antenna is Rhode & Schwartz HF9070M Broadband omnidirectional antenna, covering 800 MHz to 26.5 GHz and is vertically polarized. All the measurements were saved in a two-dimensional matrix, whose row represents the time and whose column represents the frequency. For instance, band 1805–1880 MHz has 448 frequency bins as columns, where the bandwidth of each frequency bin is 0.167 MHz. The data were measured for 131 days (188 917 min) from February to June in 2013. Therefore, the data set of each frequency bin has 188 917 time instants as rows. Eight frequency bands are measured as 880–915 MHz, 925–960 MHz, 1710–1785 MHz, 1805–1880 MHz, 1900–1920 MHz, 1920–1980 MHz, 2110–2170 MHz and 2400–2500 MHz. They represent the U.K. 2G and 3G bands as well as the Wi-Fi band. Comparing these bands, their patterns are similar so that we use the 1805–1880 MHz band as an example. Therefore, in the following sections, the predictive models will be built for the 1805–1880 MHz band.

2.2 Initial data pre-processing

The aim of the work is to predict the total power for a whole frequency band including all frequency bins. Therefore, it is important to arrange the data in a suitable data structure, as a time series. To calculate the total power for each time instant, the power values were converted from dBm measurements to mW, and added together, and then converted back into dBm. Thus, all the considered powers have a unit of dBm.

The time series representing the RF energy data has autocorrelation, which means that it is possible that future observation $y_t$ can be predicted as a function of past observations $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, where $p$ is the number of past observations, as in an autoregressive (AR) model. Thus, it is important to investigate the optimal parameter $p$, which will accurately illustrate the number of rows needed to be trained and tested to generate any ML model, to describe the random patterns of the harvested energy in the short term. In the remainder of this work, the parameter $p$ will be referred to as the feature length of the ML process, and the term ‘observations’ will be used to refer to the rows of measurements used for training and testing. To account for the the autocorrelation, the data in the time series are rearranged as in Figure 1.

The data set is then divided into training set and testing set. We use the normalized root mean square error (NRMSE) to represent the prediction accuracy as

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}},$$

$$NRMSE = \frac{RMSE}{\sum_{t=1}^{n}|y_t|},$$

where $\hat{y}_t$ is predicted value, $y_t$ is actual value and $n$ is the number of predictions.
3 | SELECTION OF IMPORTANT PARAMETERS

3.1 | Feature length

Feature length (FL) is the number of measurements in minutes before the \( n_{th} \) measurement that will be learned to make a prediction as discussed in the previous section. Theoretically, we can use autocorrelation function (ACF) to calculate the best feature size. The results of ACF for the data set are shown below in Figure 2.

In Figure 2, the autocorrelation function is tested for up to 50 lags. The degree of correlation decreases with the increase of lags. Although the overall ACF curve shows a downward trend, there are several values of lags that mean to be local maximums. From these lags, four different feature lengths as 1, 3, 10 and 15 are used for testing, and we will choose the best FL out of them.

In the following, prediction will be made by using the data structure described in Section 2.2 as multiple ‘chunks’ of data. For example, chunk 1 uses the first 100 data points in the data set to generate an ML model, chunk 2 uses the next 100 data points in the data set and so on. This is for all ML algorithms, and it is performed over 10 chunks to obtain 10 values of NRMSE for each ML model to reduce the randomness of error rates. The mean value of the NRMSE of 10 chunks is used as the final performance indicator.

From Figure 3, when using a feature length of 1, the lowest error is recorded in chunk 9 as 0.0348, while the highest error occurs in chunk 1 as 0.0553. The mean value of NRMSE is 0.0464, which gives a mean prediction accuracy of 95.36% for the feature length of 1. When using a feature length of 3, the lowest error is recorded in chunk 9 as 0.0357, while the highest error occurs in chunk 2 as 0.0564. The mean value of prediction NRMSE is 0.0471. Thus, a feature length of 3 has a mean prediction accuracy of 95.29%, lower than that for a feature length of
When using a feature length of 10, the lowest error is also recorded in chunk 9 as 0.0391, while the highest error occurs in chunk 2 as 0.0597. The mean value of prediction NRMSE is 0.0489, and the mean prediction accuracy is 95.11%. When a feature length of 15 is used, the lowest error is recorded in chunk 9 as 0.0408, while the highest error occurs in chunk 1 as 0.0574. The mean value of prediction NRMSE is 0.0501, and the mean prediction accuracy is 94.99%. These results are based on 120 observations and a training split of 80:20. The result shows that a feature length of 1 has the lowest mean error. Thus, it is best to use only the most recent observation in the prediction. We have done similar tests for other numbers of observations: 60, 120, 240 and 480. The tests of 60, 120 and 240 observations also show that FL = 1 is the optimal choice for accuracy. Hence, FL = 1 is chosen as the best feature length in later studies.

Similarly, we have studied the effect of FL on the accuracies of other ML algorithms: for SVM, in the tests of 60, 120 and 240 observations, FL = 1 records an NRMSE of 0.0510, 0.0475 and 0.0504, respectively, and has the lowest error rate. Therefore, FL = 1 is regarded as the best FL for SVM. For RFA, in the tests of 60, 240 and 480 observations, FL = 15 records 0.0500, 0.0500 and 0.0518, and has the lowest error rate of. In the test of 120 observations, the error rate of FL = 15 is 0.0476 and is the second best. Considering both accuracy and generality, FL = 15 is regarded as the best FL for RFA. For DT, in all tests of four observations, FL = 1 records 0.0585, 0.0559, 0.0607 and 0.0650 for 60, 120, 240 and 480 observations, respectively, and it has the lowest error rate. Therefore, FL = 1 is regarded as the best FL of DT. In summary, the results reveal that when FL is 1, the algorithms of LR, SVM and DT have the best performances in terms of NRMSE, and when FL is 15, the algorithm of RFA has the best performance. Therefore, FL = 1 will be used in LR, SVM and DT tests, while FL = 15 will be used for RFA in the following. These figures are not shown here to make the paper compact but are available upon request.

**FIGURE 3** NRMSE of different feature lengths of 1, 3, 10 and 15 for LR, with 120 observations and a split of 80:20
3.2 Number of observations

In this subsection, we will study the effect and the choice of the number of observations. To do this, four different numbers of observations will be tested to determine the best choice of the number of observations as 60, 120, 240 and 480. Larger numbers of observation, with up to 15 000 observations for 10 chunks, which gives a total of up to 150 000 observations, have also been tested in the research. The results have showed little improvement comparing with the result of less than 480 observations.

In Figure 4, we set FL = 1, split = 80:20, for the LR algorithm. One can see that, when using a number of 60 observations, the lowest error is recorded in chunk 8 as 0.0443, while the highest error occurs in chunk 1 as 0.0614. The mean value of the prediction NRMSE is 0.0500, which gives a mean prediction accuracy of 95.00%. When using a number of 120 observations, the lowest error is recorded in chunk 9 as 0.0348, while the highest error that occurs in chunk 1 as 0.0553. The mean value of the prediction NRMSE is 0.0464. Thus, a number of 120 observations has a mean prediction accuracy of 95.36%, higher than the number of 60 observations. When using a number of 240 observations, the lowest error is recorded in chunk 5 as 0.0404, while the highest error that occurred in chunk 9 as 0.0592. The mean value of the prediction NRMSE is 0.0495, which gives a mean prediction accuracy of 95.05%. When a number of 480 observations is used, the lowest error is recorded in chunk 8 as 0.0458, while the highest error that occurred in chunk 10 as 0.0619. The mean value of the prediction NRMSE is 0.0534, which gives a mean prediction accuracy of 94.66%. The results illustrate that, the performances of both the mean prediction accuracy and the lowest error are the best when we use 120 observations. Even when a higher number of observations is tested, such as 15 000, the results are not better than 120 observations. Thus, 120 is the best choice for the number of observations for LR algorithm. Similarly, we can obtain the best number of observations for other ML algorithms. The results are listed in Table 1 below.

**FIGURE 4** NRMSE of different observations of 60, 120, 240 and 480 when FL = 1, split = 80:20 for LR algorithm
The results in Table 1 show that, when the number of observations is 120, the NRMSEs of LR, SVM, RFA and DT are 0.0464, 0.0475, 0.476 and 0.0559, respectively. All the algorithms reach their lowest error rates when 120 observations are used. Therefore, we will use 120 as the number of observations to compare these ML algorithms later.

| ML algorithm          | Number of observations |       |       |       |
|-----------------------|------------------------|-------|-------|-------|
|                       | 60                     | 120   | 240   | 480   |
| Linear regression     | 0.0500                 | 0.0464| 0.0495| 0.0534|
| Support vector machine| 0.0510                 | 0.0475| 0.0504| 0.0547|
| Random forest         | 0.0496                 | 0.0476| 0.0493| 0.0508|
| Decision tree         | 0.0585                 | 0.0559| 0.0607| 0.0650|

The results in Table 1 show that, when the number of observations is 120, the NRMSEs of LR, SVM, RFA and DT are 0.0464, 0.0475, 0.476 and 0.0559, respectively. All the algorithms reach their lowest error rates when 120 observations are used. Therefore, we will use 120 as the number of observations to compare these ML algorithms later.

### 3.3 Training split

In this subsection, we will study the effect and the choice of the number of training split. To do this, four different splits between training set and testing set will be tested to determine the best choice of the training splits as 80:20, 70:30 and 60:40 and 50:50.
In Figure 5, we set FL = 1, the number of observations = 120 for LR algorithm, as LR shows the best prediction accuracy in the previous subsections. One can see that, when using a training split of 80:20, the lowest error is recorded in chunk 9 as 0.0348, while the highest error occurs in chunk 1 as 0.0553. The mean value of the prediction NRMSE is 0.0464, which gives a mean prediction accuracy of 95.36%. When using a training split of 70:30, the lowest error is recorded in chunk 9 as 0.0361, while the highest error that occurs in chunk 1 as 0.0560. The mean value of the prediction NRMSE is 0.0467, Thus, a training rate of 70:30 has a mean prediction accuracy of 95.33%, lower than 80:20. When using a training split of 60:40, the lowest error is recorded in chunk 9 as 0.0363, while the highest error that occurred in chunk 9 as 0.0565. The mean value of the prediction NRMSE is 0.0470, which gives a mean prediction accuracy of 95.30%. When a training split of 50:50 is used, the lowest error is recorded in chunk 8 as 0.0458, while the highest error that occurred in chunk 10 as 0.0574. The mean value of the prediction NRMSE is 0.0477, which gives a mean prediction accuracy of 95.23%. The results illustrate that the performances of both the mean prediction accuracy and the lowest error are the best when we use 80% training split. Thus, 80:20 is the best choice of the proportion between training set and testing set for the ambient RF energy modelling.

3.4 | Algorithms

As mentioned above, in the research, four different algorithms are tested. They are LR, SVM, RFA and DT. To explain the principle of algorithms more clearly, the processing of these algorithms is illustrated by the flow graph in Figure 6. LR uses the gradient descent method to optimize the loss function and fit data linearly. SVM realizes it by calculating hyper-plane and determining the kernel function. DT extracts features of data by classifying feature points as different branches and calculating their information entropy. Compared with DT, RF alleviates the effect of the overfitting problem by generating many separate event trees randomly.

4 | NUMERICAL RESULTS AND DISCUSSION

For all the following results, the FL for RFA is set as 15 and the FL for other algorithms is set as 1, and the number of observations is set as 120 with 80/20 split between training sets and testing sets. These choices are based on the tests in Section 3. Each prediction is also made using 10 chunks. There are 120 consecutive observations in each chunk, with a total of 1200 records used. It has been discussed that using larger records does not improve the accuracy significantly. The 10 data fragments in each chunk used in this section are obtained randomly from the original data set by excluding data used in Section 3.
4.1 Minutes

In Figure 7, the average prediction errors for LR, SVM, RFA and DT are 0.0464, 0.0475, 0.0476 and 0.0559, respectively, over all chunks, and the mean value of these errors is 0.0494. Overall, the algorithm with the best performance is LR with a mean error of 0.0464, which gives a mean prediction accuracy of 95.36%. The lowest error of LR is recorded in chunk 9 as 0.0348, while the highest error occurs of LR in chunk 1 as 0.0553. The range of errors for LR is 0.0205. Next, SVM has a mean error of 0.0475, which gives a mean prediction accuracy of 95.25%. The lowest error of SVM is recorded in chunk 9 as 0.0363, while the highest error of SVM occurs in chunk 1 as 0.0552. The range of errors for SVM is 0.0189. Next is RFA, which has a mean error of 0.0476, which gives a mean prediction accuracy of 95.24%. The lowest error of RFA is recorded in chunk 9 as 0.0359, while the highest error of RFA occurs in chunk 1 as 0.0558. The range of errors for RFA is 0.0199. The algorithm with the worst performance is DT with a mean error of 0.0559, which gives a mean prediction accuracy of 94.41%. The lowest error of DT is recorded in chunk 9 as 0.0409, while the highest error of DT occurs in chunk 1 as 0.0706. The range of errors for DT is 0.0297. The results show that LR outperforms all other ML algorithms considered in terms of the errors of the 10 chunks. Although LR has the best performance in mean error, the range of errors of SVM is the smallest among all, and RFA is the second best, which means that they are the most stable model due to less variation of error.

To improve the accuracy of modelling, the data set is then labelled by days before prediction. We assume that data from the same day of different weeks will resemble each other. For this reason, models are built for each day to improve accuracy.

In Figure 8, the performances of different ML algorithms on Tuesday and Wednesday are presented. On Tuesday, the prediction error for LR, SVM, RFA and DT are 0.0436, 0.0442, 0.0448, 0.0539, respectively. Overall, the algorithm with the best performance is LR with a mean error of 0.0436, which gives a mean prediction accuracy of 95.64%. The lowest error of LR occurs in chunk 6 as 0.0384, while the highest error of LR occurs in chunk 5 as 0.0550. The range of errors for LR is 0.0166. The algorithm with the worst performance is DT with a mean error of 0.0539, which gives a mean prediction accuracy of 94.61%. The lowest error of DT is recorded in chunk 6 was 0.0449, while the highest error of DT occurs in chunk 5 is 0.0702. The range of errors for DT is 0.0253. The results show that LR outperforms all other ML algorithms when comparing the error rate. However, the range of errors of RFA is the smallest and LR records the second best, which means that RFA and LR are the most stable models on Tuesday. Finally, LR also has the lowest error in chunks.

On Wednesday, the prediction errors for LR, SVM, RFA and DT are 0.0456, 0.0464, 0.0463 and 0.0558, respectively. Overall, the algorithm with the best performance is LR with a mean error of 0.0456, which gives a mean prediction accuracy of 95.44%. The lowest error of LR is recorded in chunk 3 as 0.0357, while the highest error occurs of LR in chunk 10 as 0.0560. The range of errors for LR is 0.0203. The algorithm with the worst performance is DT with a mean
error of 0.0262, which gives a mean prediction accuracy of 97.38%. However, though LR has the best average accuracy, the range of error of SVM is the smallest and LR records the second best, which means that SVM and LR are the most stable models on Wednesday. Moreover, both the lowest error and the highest error in chunks of LR are lower than other models.

By labelling data with days, there is a slight improvement in the accuracy of predictions. On Tuesday, the mean error decreases by 5.67% from 0.0494 to 0.0466, and on Wednesday, the mean error decreases by 1.82% from 0.0494 to 0.0485, and the lowest errors on both days also decrease, compared with predictions using minutes in Figure 7.

4.2 Hours

We can also combine the measurements for different minutes into aggregate measurements for different hours, by taking an arithmetic mean for the measurement at each minute within the hour, to reduce randomness and therefore to increase prediction accuracy.

In Figure 9, the prediction error for LR, SVM, RFA and DT are 0.0352, 0.0353, 0.0387 and 0.0424, respectively. Overall, using the hourly data, the algorithm with the best performance is LR with a mean error of 0.0352, which gives a mean prediction accuracy of 96.48%. The lowest error of LR is recorded in chunk 8 as 0.0301, while the highest error of LR occurs in chunk 4 as 0.0440. The range of errors for LR is 0.0139. The algorithm with the worst performance is DT.
with a mean error of 0.0250, which gives a mean prediction accuracy of 97.52%. The lowest error of DT is recorded in chunk 8 as 0.0365, while the highest error of DT occurs in chunk 4 as 0.0547. The range of errors for DT is 0.0182. The results show that LR outperforms all other ML algorithms in terms of average accuracy. However, SVM has the smallest range of error, and LR is the second best, which means that SVM is the most stable models due to less error variation.

4.3 Harvester operation efficiency

The energy models built in the previous subsections can be used to optimize the control of harvester operation. If the predicted energy is smaller than the predetermined threshold, the harvester will go to sleep to save energy. Only when the predicted energy is above the threshold, the harvester will be activated to harvest energy. In this part, a threshold is set for the energy harvester to determine turn-ons and turn-offs of the energy harvester. If the actual energy falls below it while the predicted energy is above it, or if the harvested energy is above it, while the predicted energy is below it, the harvester will make a mistake. Thus, the number of false operations is recorded, and a penalty is given. The penalty using each hour will be compared.

Figure 10 uses the LR model with its best parameter settings. The results use the 8th chunk of the uncategorized data that has the lowest error among all chunks. The threshold is set as $-17.2$ dBm. According to Figure 10, the mean number of overestimates is 4.55, while the mean number of underestimates is 9.35. The mean error rate of estimation is 0.1390. In Figure 10, the prediction follows the trend of the actual data with similar patterns. However, the prediction lags the actual value slightly which may lead to misoperation in time-critical applications.

Figure 11 uses the SVM model with its best parameter settings. The results use the 8th chunk of uncategorized data, which has the lowest error for SVM model among all chunks. The threshold is set as $-17.2$ dBm. According to Figure 11, the mean number of overestimates is 2.45, while the mean number of underestimates is 11.65. The mean error rate of estimation is 0.1410, higher than the error rate of the LR model. In Figure 11, the prediction follows the trend of the actual data with similar patterns. However, the prediction of SVM model also lags the actual value slightly. In the hours which record high energy, SVM model shows the worse performance of fitting actual value than the LR model, which may lead to more underestimation in the harvester.

Figure 12 uses the RFA model with its best parameter settings. The results use the 9th chunk of uncategorized data, which has the lowest error for RF model among all chunks. The threshold is set as $-17.2$ dBm. According to Figure 12, the mean number of overestimates is 0.35, while the mean number of underestimates is 29.7. The mean error rate of estimation is 0.3005. In Figure 12, although the prediction of RFA can also follow the trend of the actual data, compared with LR and SVM models, the prediction of RFA lags the actual value more than other two models, which causes a higher mean error rate than LR's and SVM's. The accuracy of the RFA prediction is not acceptable.
Upon comparison, LR model shows the best performance in prediction accuracy. Figure 13 shows the LR model obtained. The selected LR model formula is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{15} X_{15} + \epsilon,$$

and the coefficients obtained are shown in Table 2 below. In Table 2, Estimate means coefficients of each terms which are estimated by the model. SE is standard error of the coefficients. tStat is $t$ statistic to test null hypothesis for each coefficient, with $tStat = \frac{Estimate}{SE}$. $p$ value for the $t$ statistic is used to test whether the corresponding coefficient is zero or not at the determined significant level. The number of observation is $n = 96$, because 80% observations of 120 are split as training set to generate model. The degree of freedom for the error is $n - p = 80$, in which $p = 16$ is the number of coefficients in the model. As a standard of fitness, $R^2$ is 0.603, while adjusted $R^2$ is 0.529.

**FIGURE 11** Prediction efficiency test for hourly data using SVM with a threshold of $-17.2$ dBm

**FIGURE 12** Prediction efficiency test for hourly data using RFA with a threshold of $-17.2$ dBm
5 | CONCLUSIONS

This work has studied the use of reliable ML models for RF energy data. Four different ML algorithms (LR, SVM, RFA and DT) have been discussed and their performances in RF energy harvesting works have been compared using the data set in the 1805–1880 MHz band. The results have shown that, in terms of average accuracy, LR has the highest and most stable accuracy, followed by SVM and RFA, with DT being the worst model. For the harvester operation efficiency, LR has the highest accuracy, followed by SVM, and RFA has given an unacceptable error rate in the energy harvesting efficiency. The advancement of knowledge includes the following. First, to our best knowledge, this is the first time that ML is used to predict energy for RF harvesting. Second, our proposed predictive models have very high accuracies. This allows system designers to operate the energy harvester efficiently. These contributions justify the work. In the future, we intend to explore better ways to set threshold, called ‘dynamic threshold’, to improve the energy harvesting efficiency. More experiments will be carried out based on algorithm computation time, to ensure that the harvester can make prediction. Finally, we will integrate the prediction with energy management in wireless sensor networks to design new protocols that extend network lifetime.

### TABLE 2  Estimated coefficients of selected LR model

| Estimate | SE  | tStat | p value |
|----------|-----|-------|---------|
| Intercep | -4.7224 | 2.0971 | -2.2519 | 0.027069 |
| x1       | -0.029871 | 0.10847 | -0.27538 | 0.78373 |
| x2       | -0.027799 | 0.11851 | -0.23457 | 0.81514 |
| x3       | 0.13541 | 0.12299 | 1.101 | 0.27421 |
| x4       | -0.021097 | 0.12335 | -0.17103 | 0.86463 |
| x5       | 0.095555 | 0.12267 | 0.77895 | 0.43831 |
| x6       | 0.050455 | 0.12479 | 0.40431 | 0.68706 |
| x7       | -0.246 | 0.12188 | -2.0184 | 0.046898 |
| x8       | 0.062442 | 0.12486 | 0.50011 | 0.61387 |
| x9       | -0.2374 | 0.1222 | -1.9428 | 0.055561 |
| x10      | -0.06328 | 0.12729 | -0.49712 | 0.62047 |
| x11      | 0.24891 | 0.12425 | 2.0032 | 0.048538 |
| x12      | -0.050529 | 0.12542 | -0.40289 | 0.6881 |
| x13      | 0.048124 | 0.12498 | 0.38504 | 0.70123 |
| x14      | 0.35397 | 0.11902 | 2.974 | 0.0038838 |
| x15      | 0.41164 | 0.11188 | 3.6792 | 0.00042267 |
DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID
Youjie Ye https://orcid.org/0000-0002-1962-2980
Yunfei Chen https://orcid.org/0000-0001-8083-1805

REFERENCES
1. Dhanoriya S, Pandey M. A survey on wireless sensor networks: faults, misbehaviour and protection against them. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2017; Delhi:1-7.
2. Akyildiz IF, Weilian S, Sankarasubramaniam Y, Cayirci E. A survey on sensor networks. IEEE Commun Mag. 2002;40(8):102-114.
3. Lu X, Wang P, Niyato D, Kim DI, Han Z. Wireless networks with RF energy harvesting: a contemporary survey. IEEE Commun Surv Tut. 2015;17(2):757-789. Secondquarter.
4. Guler U, Sendi MSE, Ghoovanloo M. A dual-mode passive rectifier for wide-range input power flow. IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS); 2017:1376-1379.
5. Liu Y, Li W, Jia B. Design of ZigBee-based energy harvesting wireless sensor network and modeling of solar energy. Sec Priv New Comput. 2019;284:576–584.
6. Chen Y. Energy Harvesting Communications: Principle and Theories. Hoboken, New Jersey: Wileys-IEEE Press; 2019:8-9.
7. Olgun U, Chen C, Volakis JL. Design of an efficient ambient WiFi energy harvesting system. IET Microwaves Anten Propag. 2012;6(11): 1200-1206.
8. Vyas RJ, Cook BB, Kawahara Y, Tentzeris MM. E-WEHP: a batteryless embedded sensor-platform wirelessly powered from ambient digital-TV signals. IEEE Trans Microw Theory Techniq. 2013;61(6):2491-2505.
9. Shigeta R, et al. Ambient RF energy harvesting sensor device with capacitor-leakage-aware duty cycle control. IEEE Sens J. 2013;13(8):2973-2983.
10. Dhivya V, Kalaiyarasi P, Shamini GI. Survey on spectrum occupancy by using different techniques. In: International Conference on Computation of Power, Energy Information and Communication (ICCPEC), Vol. 2017; 2017; Melmaruvathur:316-320.
11. Kim S, Vyas R, Bito J, et al. Ambient RF energy-harvesting technologies for self-sustainable standalone wireless sensor platforms. Proc IEEE. 2014;102(11):1649-1666.
12. Chen Y, Zhao N, Alouini M. Wireless energy harvesting using signals from multiple fading channels. IEEE Trans Commun. 2017;65(11): 5027-5039.
13. Piñuela M, Mitcheson PD, Lucyszyn S. Ambient RF energy harvesting in urban and semi-urban environments. IEEE Trans Microw Theory Techniq. 2013;61(7):2715-2726.
14. Azmat F, Chen Y, Stocks N. Predictive modelling of RF energy for wireless powered communications. IEEE Commun Lett. 2016;20(1):173-176.
15. Long M, Chen Y. Performance analysis of energy harvesting communications using multiple time slots. IET Commun. 2019;13(3):289-296.
16. Flach P. Machine Learning: The Art and Science of Algorithms That Make Sense of Data. Cambridge, U.K.: Cambridge Univ Press; 2012.
17. Weiss SM, Kulikowski CA. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. San Francisco, CA, United States: Morgan Kaufman; 1991.
18. Le Q, Ranzato M, Monga R, et al. Building high-level features using large scale unsupervised learning. ICML. 2012;2:3.
19. Abuzaibain N, Saad W, Maham B. Robust Bayesian learning for wireless RF energy harvesting networks. In: 2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt); 2017; Paris:1-8.
20. Zou Z, Gidmark A, Charalambous T, Johansson M. IEEE J Select Areas Commun. 2016;34(12):3528-3539.
21. Liang Z, Yuan J. Modelling and prediction of mobile service channel power density for RF energy harvesting. IEEE Wirel Commun Lett. 2020;9(5):741-744.
22. Nau R. Introduction to ARIMA: Non Seasonal Models. Durham, North Carolina, United States: Fuqua School of Business Duke University. https://people.duke.edu/rnau/411arim.htm

How to cite this article: Ye Y, Azmat F, Adenopo I, Chen Y, Shi R. RF energy modelling using machine learning for energy harvesting communications systems. Int J Commun Syst. 2021;34:e4688. https://doi.org/10.1002/dac.4688