NON-INTRUSIVE LOAD MONITORING FOR LED LIGHT CLASSIFICATION: A DATA-DRIVEN MACHINE LEARNING APPROACH

Nguyen Thanh Cong[1], Nguyen Ngoc Son[1], Dao Ngoc Nam Hai[2], Nguyen Huy Tinh[1], Jonathan Andrew Ware[3], Nguyen Ngoc An[1*]
[1]VNU University of Engineering and Technology, [2]VNU Institute of Information Technology
[3]University of South Wales, United Kingdom

ARTICLE INFO

Received: 11/4/2024
Revised: 10/6/2024
Published: 10/6/2024

KEYWORDS

Non-intrusive load monitoring (NILM)
LED operational state classification
Discrete Fourier transform
Confident Learning
Data-centric machine learning
Machine Learning

ABSTRACT

Monitoring the operational status of LED lights is important to achieve energy efficiency and protect user health. Recent studies employed machine learning and several parameters, such as the LED's light output and electrical characteristics, to classify their operational status. However, under changing environmental conditions, these methods will no longer be effective, due to the compromise of the environmental noise to the input data of the models. In this study, we proposed a novel approach to identifying the operational status of household LED lights using non-intrusive load monitoring, machine learning models, confident learning, and the oscillation characteristic of the root-mean-square (RMS) current. By using the oscillation characteristics of the RMS current, we significantly reduced the number of inputs to the models and their computational hardware requirements compared to models using the RMS current. With the introduction of confident learning, we improved the prediction accuracy of the models by 2% on average. The models achieved prediction accuracy ranging from 94% to 97.5%. The proposed method shows potential in applying to different kinds of electrical devices.

PHÂN LOẠI TRẠNG THÁI ÁNH SÁNG CỦA ĐÈN LED SỬ DỤNG GIÁM SÁT TẢI KHÔNG XÂM LẤN VÀ HỌC MÁY HƯỚNG DỮ LIỆU

Nguyễn Thành Công[1], Nguyễn Ngọc Sơn[1], Đào Ngọc Nam Hải[2], Nguyễn Huy Tình[1], Jonathan Andrew Ware[3], Nguyễn Ngọc An[1*]
[1]Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội, [2]Viện Công nghệ Thông tin - Đại học Quốc gia Hà Nội
[3]Đại học South Wales – Vương quốc Anh

THÔNG TIN BÀI BÁO

Ngày nhận bài: 11/4/2024
Ngày hoàn thiện: 10/6/2024
Ngày đăng: 10/6/2024

TỪ KHÓA

Giám sát tải không xâm nhập
Phân loại trạng thái hoạt động của đèn LED
Biến đổi Fourier rời rạc
Học tự tin
Học máy hướng dữ liệu
Học máy

TÓM TẮT

Việc theo dõi trạng thái hoạt động của đèn LED có vai trò quan trọng trong việc sử dụng năng lượng hiệu quả và bảo vệ sức khỏe người dùng. Một số nghiên cứu gần đây sử dụng học máy kết hợp với một số tham số, như công suất phát sáng và đặc tính điện, nhằm phân loại trạng thái hoạt động của đèn LED. Tuy nhiên, trong điều kiện môi trường thay đổi, các phương pháp này sẽ không còn hiệu quả do ảnh hưởng của nhiễu môi trường đến dữ liệu đầu vào của mô hình. Trong nghiên cứu này, chúng tôi đề xuất một phương pháp mới để xác định trạng thái hoạt động của đèn LED gia dụng bằng cách sử dụng giám sát tải không xâm nhập, kết hợp cùng với học máy và học tự tin. Bằng cách sử dụng các đặc tính dao động của dòng RMS, chúng tôi đã giảm đáng kể số lượng đầu vào cho các mô hình học máy và yêu cầu phần cứng của chúng nhằm thực hiện tính toán so với các mô hình sử dụng dòng RMS. Với việc bổ sung thêm phương pháp học tập tự tin, độ chính xác dự đoán của các mô hình được cải thiện thêm trung bình 2%. Các mô hình học máy đạt độ chính xác trong việc dự đoán dao động từ 94% đến 97,5%. Phương pháp đề xuất cho thấy tiềm năng áp dụng cho các loại thiết bị điện khác nhau.

DOI: https://doi.org/10.34238/tnu-jst.10115

* Corresponding author. *Email: ngocan@vnu.edu.vn*

1. Introduction

LED lights are a preferable option in residential and industrial lighting [1], [2]. Compared to traditional lighting systems, LED lights offer several outstanding advantages, such as a longer lifespan, higher luminous efficiency, lower energy consumption, high color rendering index, and suitability for human physiology [3] – [10]. However, they also encounter issues such as a gradual decline in lifespan and the susceptibility of luminous efficiency to various operating factors. When the luminous efficiency decreases, the light quality gradually deteriorates [11]. However, the degradation is not easy to detect with the naked eye and potentially affects the visual health of users. The degradation also leads to significant electricity wastage, especially when lighting systems account for around 20% of global electricity consumption [12]. Therefore, monitoring the operational status of LED lights is essential to optimize energy usage and ensure user health.

Currently, several methods exist to monitor and predict the LED light operational states based on some measurable parameters. For example, those methods include the measurement and analysis of the output electrical parameters of the LED light source, such as voltage and current, the optical indices of the LED light, such as flicker index or luminous flux, the LED chip temperature, the combined information from the optic, thermal, and electrical parameters of the LED light [13] – [16]. However, optical and temperature measurement methods are often susceptible to environmental influences. Meanwhile, methods to measure the LED power output parameters typically require hardware intervention to the LED light components, disrupting the system and sometimes inconveniently necessitating the placement of measurement devices on the LED light. It is, therefore, inconvenient and lacks readily available measurement devices in the market to meet these requirements.

In addition, many studies have chosen the non-intrusive load monitoring (NILM) approach and used machine learning (ML) techniques to deal with this problem [13], [15], [17]. Y. Shang et al. used the Supported Vector Machine (SVM) algorithm to monitor the FSL LEDs with an accuracy of 100% and the OSRAM LEDs with an accuracy of 89.3% under ideal conditions [15]. However, the method will no longer be effective when encountering optical interference or changing the lighting system. Meanwhile, H. Jiang et al. have also used the SVM algorithm to classify LED lamp failures with an accuracy of 65.4% on the test set [17]. However, the method's performance is also compromised by environmental factors. The dataset includes parameters such as average illuminance, lumen maintenance level, and color rendering index, which are affected by environmental conditions.

In a previous study, we proposed using only the RMS current obtained by the NILM method to classify the operational states of the LED lights [18]. The constructed ML models have stable performance while suffering negligible effects from optical noise. However, there is still the need to increase the prediction accuracy and decrease the computational complexity.

In this study, we propose a novel approach using the oscillation characteristics of the RMS current as the input to machine learning models, combined with the confident learning technique. Using the oscillation characteristics obtained by taking a discrete Fourier transform (DFT) of the RMS current as model input, we aim to reduce the computational requirements of the machine learning models. Furthermore, the confident learning technique will increase the models' prediction accuracy. In the meantime, the advantages of the NILM method are maintained.

2. Data and methods

2.1. Measurement system and LED light operational states

2.1.1. The NILM mesurement system

We propose a NILM system to measure the RMS current as shown in Figure 1. The SCT013-100A current transformer and BL0940 IC are selected to collect the RMS current data for

monitoring and storage. Compared to devices with similar functionality, the SCT013-100A current transformer offers higher accuracy and a sampling frequency of up to 1 kHz. At the same time, the BL0940 IC also features a high sampling rate with exceptional precision with no calibration. Furthermore, it possesses robust noise-handling capabilities with data transmission speeds up to 900 kHz. The NILM measurement system is implemented by clamping the current sensor onto the power supply wire of the LED lights without necessitating alterations to the device's original design. With a compact size, the proposed measurement system can be easily installed in various locations, making it suitable for widespread applications in household settings.

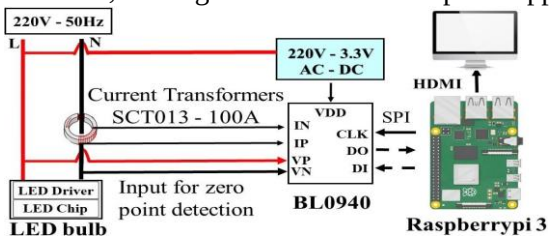


Figure 1. *The NILM measurement system diagram*

Table 1. *Technical specifications of LED Bulb A55N4/5W.H, RANG DONG*

Specification	Value
Power	5W
Voltage	150-220V AC
Luminous flux	475 lm (6500K)
Luminous efficiency	95lm/W (6500K)
Operating temperature	-10 to 40°C
Lifetime	20000 hours (L70)

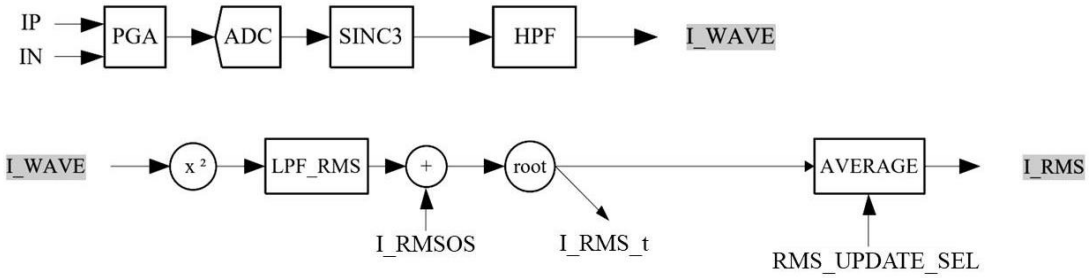


Figure 2. *The measurement of the RMS current*

The measurement of the RMS current is depicted in Figure 2. Firstly, the alternating current intensity is captured by the current sensor. After that, the signal undergoes calibration and processing via the hardware of the BL0940 module with the following procedures: The AC sensor signal is amplified by a programmable gain amplifier (PGA) and then sampled by a high-frequency ADC. After super-discretization, the SINC3 filter is applied to the data to remove high-frequency components. DC components are also removed from the signal. The filtered data is self-multiplied. The output squared value goes through a low-pass filter to eliminate high-frequency components. The signal is then added to the calibrated value I_RMSOS and square-rooted to calculate the RMS value of the signal. Subsequently, the RMS value is averaged over many samples to enhance accuracy. Finally, the Raspberry Pi embedded system retrieves data from the BL0940 module via the SPI protocol.

2.1.2. LED operational status

In this study, we have surveyed the operations of 300 LED bulbs. The error tests on general LED bulbs suggest that lifespan deterioration is susceptible to operating conditions and environmental temperature. Accordingly, we categorized the operational states of the LED bulb into the following groups, namely normal functioning (normal, 30 sets), current surpassing rated values (overcurrent, 60 sets), being affected by high temperatures (overheating, 30 sets), complete failure (broken, 110 sets), and insufficient current for guaranteed luminous efficiency (error, 100 sets). In this study, we used LED bulbs with the technical specifications shown in Table 1.

2.2. Data curation and cleaning

2.2.1. Data curation

We measured the RMS current of these 300 LED bulbs, gathering 12,000 data segments over 30 hours. Each data segment had 9 seconds and was recorded every minute using an electricity energy meter. Each data label has an equal number of samples to mitigate the impact of data imbalance. The distribution of the data labels is shown in Figure 3. The collected RMS current data was divided into three datasets, including the training set (72%), the validation set (10%), and the test set (18%), as shown in Table 2. Discrete Fourier transform (DFT) was applied to the datasets to collect the oscillation characteristics of the RMS current to use as model input. The data collection and splitting process is illustrated in Figure 4.

Table 2. *Compositions of training set, validation set, and test set*

Datasets Labels	Training set	Validation set	Test set	Quantity	Percent (%)
Error	1718	237	445	2400	20%
Normal	1734	247	419	2400	20%
Overcurrent	1709	237	454	2400	20%
Overheating	1766	223	411	2400	20%
Broken	1722	247	431	2400	20%
Total	8649	1191	2160	12000	100%
Proportion (%)	72%	10%	18%	100%	

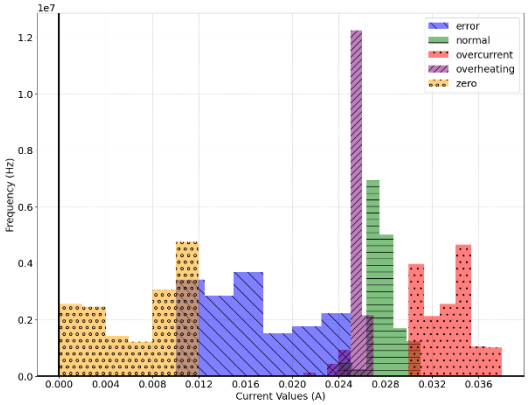


Figure 3. *Measured data distribution on measured data*

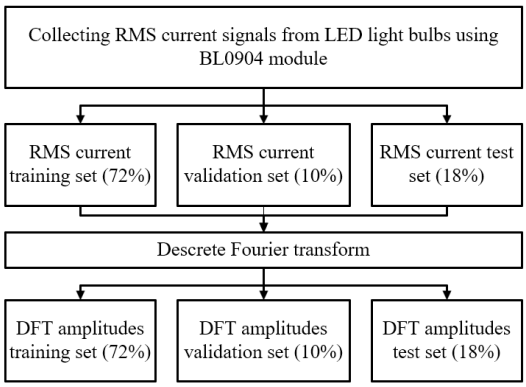


Figure 4. *Data collection and splitting process*

2.2.2. Data cleaning

Northcutt et al. have mentioned that unwanted label errors can critically affect the performance of machine learning models [19]. They also discussed confident learning (CL) as an effective method to find and prune label errors from the datasets. In this case, the CL process began with training the XGBoost model on each of the datasets in a manner called cross-validation. Here, the cross-validation splitting strategy was a 5-fold cross-validation, implemented by the cross-validation API from the scikit-learn library [20]. During each round of training, four-fifths of each dataset was for model input. After training, the output model was used to calculate the prediction probability of each data point in the remaining one-fifth of the dataset. A label is considered an error label if its predicted probability is lower than the threshold corresponding to its class. Meanwhile, if its predicted probability is larger than the class threshold, it is considered a correct label. Using the number of correct and error labels, a statistical data matrix to group and count error labels was constructed and called the confident joint [19]. While the diagonal entries of the matrix show the number of correct labels, the off-diagonals represent label error counts. Then, the error labels were pruned to create clean datasets.

The compositions of the datasets before and after cleaning by CL are presented in Table 3. The CL process was executed before applying the DFT to the datasets.

Table 3. *Compositions of the training set, validation set, and test set before and after cleaning by CL*

Labels	Training set			Validation set			Test set		
	Collected	No. of label errors	After CL	Collected	No. of noisy labels	After CL	Collected	No. of label errors	After CL
Error	1718	96	1622	237	6	231	445	31	414
Normal	1734	122	1612	247	18	229	419	25	394
Overcurrent	1709	56	1653	237	17	220	454	33	421
Overheating	1766	78	1688	223	23	200	411	37	374
Broken	1722	109	1613	247	4	243	431	22	409
Total	8649	461	8188	1191	68	1123	2160	148	2012

2.3. Model training

2.3.1. Model selection and parameters

The classifiers were trained using two kinds of data: the RMS current and the oscillation characteristics of the RMS current obtained from a DFT transformation. Three supervised machine learning algorithms, namely Support Vector Machines (SVM) [21], Random Forest (RF) [22], and XGBoost [23], were employed to classify and predict the labeled datasets. The first two models were trained using the scikit-learn library (version 1.3.0), while XGBoost with the XGBoost library (version 1.7.6)[1]. Maintaining a data-centric approach, we tried to minimize the hyperparameter optimization process. Only a few essential hyperparameters were manually chosen, while others were left as default. With SVM models, the essential hyperparameter is the kernel method (*kernel*), which was set as the Radial Basis Function. For the RF models, two essential hyperparameters were chosen: the number of decision trees (*n_estimators*) and the evaluation method (*criterion*), which were set to 100 and "entropy," respectively. Finally, for the XGBoost model, the *objective* parameter was set to "binary: logistic," while the *tree_methods* parameter used is "gpu_hist" to utilize the GPU's fast computational. The parameters of these models are kept the same with both data types.

2.3.2. Performance metrics

Regarding the performance metrics, we consider accuracy, macro precision, macro recall, and macro F1-Score [24] for the model evaluation. Equations (1) to (4) show the expressions of the performance metrics. When considering each class, we present the confusion matrix for each of them as $M_i = \begin{bmatrix} TP_i & FN_i \\ FP_i & TN_i \end{bmatrix}$, where i is a class in the classification. True positive (TP) represents the number of positive data points corresponding to the label being evaluated that the model correctly predicts. In contrast, true negative (TN) represents the number of negative data points corresponding to the evaluated label that the model correctly predicts. False positives (FP) are data points predicted to belong to the positive class but actually belong to another class. Conversely, false negative (FN) refers to instances that are incorrectly classified as not belonging to the positive class, despite their true class being the positive class.

$$Accuracy = \frac{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} TN_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} TN_i + \sum_{i=1}^{n} FP_i + \sum_{i=1}^{n} FN_i} \quad (1)$$

$$Macro\ precision = \frac{\sum_{i=1}^{n} Precision_i}{n} \text{ where } Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (2)$$

[1] Code and material are available in: *https://github.com/Lelvels/mylab-nilm-led-operation-detection.git*

$$Macro\ recall = \frac{\sum_{i=1}^{n} Recall_i}{n} \text{ where } Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$Macro\ F1\ Score = \frac{\sum_{i=1}^{n} F1\ Score_i}{n} \text{ where } F1\ Score_i = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \quad (4)$$

where i presents the i^{th}class in the classification and n is the number of the classes.

2.3.3. Input features selection for models using the DFT amplitude inputs

Table 4. *A comparison between DFT models' accuracy with different frequency inputs.*

Accuracy / Models	With only the DC component	Frequency range with the highest prediction accuracy	Max. frequency range 0-500Hz
SVM	0.929	0.934 (0-126.66Hz)	0.932
Random Forest	0.966	0.967 (0-1.11Hz)	0.75
XGBoost	0.966	0.974 (0-28.88Hz)	0.967

The RMS current was recorded at the sampling rate of 1 kHz. Therefore, when processing each RMS current data sample segment, the model will intake a large amount of data, hence a longer training time. Observations showed that the RMS current of the LED bulbs oscillates slowly and differently among different operational states. Therefore, the oscillation characteristics of the RMS current can be used to detect the LED bulb's operational states. Additionally, since the RMS current fluctuates slowly, its oscillation characteristics can be represented by a small number of frequencies and amplitudes obtained from DFT transformation. In this manner, the number of inputs to the models can be reduced significantly, and the training time can be reduced. Here, we investigated the variation in the prediction accuracy of the models when increasing the range of input frequencies from 0 to 500Hz.

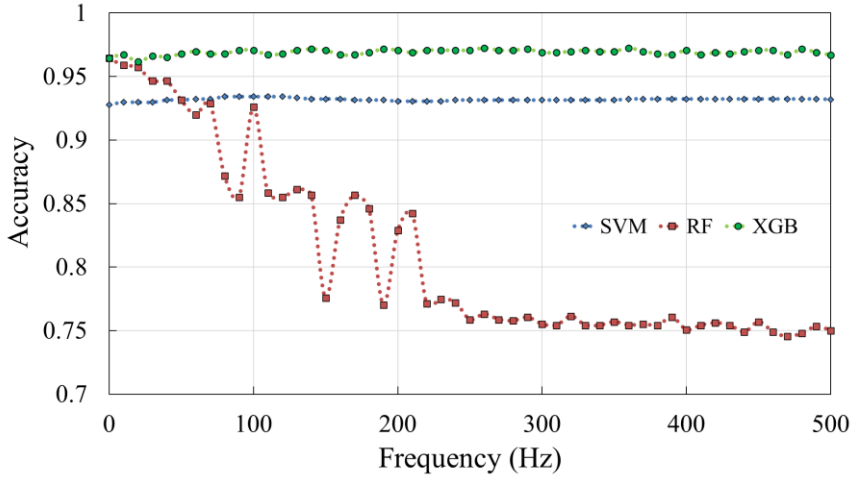


Figure 5. *Variation of the model's accuracy on the validation set corresponding to the input frequency ranges (The x-axis shows the maximum frequencies of the input frequency ranges)*

The RMS current was recorded at the sampling rate of 1 kHz. Therefore, when processing each RMS current data sample segment, the model will intake a large amount of data, hence a longer training time. Observations showed that the RMS current of the LED bulbs oscillates slowly and differently among different operational states. Therefore, the oscillation characteristics of the RMS current can be used to detect the LED bulb's operational states. Additionally, since the RMS current fluctuates slowly, its oscillation characteristics can be represented by a small number of frequencies and amplitudes obtained from DFT transformation. In this manner, the number of inputs to the models can be reduced significantly, and the training time can be reduced. Here, we investigated the variation in the prediction accuracy of the models when increasing the range of input frequencies from 0 to 500Hz.

In particular, the DFT, shown in Equation (5), was performed on every RMS current sample to collect these oscillation features. Here, N is denoted as the number of data points in a single sample, X_k is the k^{th} frequency component of the signal, with $0 \leq k \leq N-1$, while x_n as the n^{th} sample of the signal. Equation (6) shows the computation to extract the amplitudes A_k of the sinusoidal wave with frequency k/N. To reduce the computational cost, we normalize the frequency component by dividing it by the number of data points.

$$X_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N} \quad (5)$$

$$A_k = |X_k| \quad (6)$$

After the DFT, we obtained the amplitudes of 4501 points of frequencies ranging from 0 to 500Hz, which is half of the sampling rate, following the Nyquist-Shannon sampling theorem [25]. These amplitudes were the input to the models (DFT models). The range of frequencies whose amplitudes were input to the models affects the prediction accuracy and the computational requirements. The results are shown in Table 4 and Figure 5. One may notice that the DC component can provide adequate prediction accuracy while significantly reducing the number of inputs. Therefore, we used the DC components obtained from DFT as the input to the machine learning models for simplicity.

3. Results and discussions

3.1. Performance metrics and training time of the models

Table 5. *Comparing performance metrics of the models using RMS current features on the test set without CL* [18] *and with CL applied*

Metrics	SVM		RF		XGBoost	
	Without CL [18]	CL applied	Without CL [18]	CL applied	Without CL [18]	CL applied
Precision	0.915	0.941	0.96	0.982	0.944	0.97
Recall	0.909	0.941	0.96	0.982	0.944	0.968
ACC	0.91	0.94	0.96	0.981	0.944	0.968
F1-score	0.909	0.94	0.96	0.982	0.944	0.968

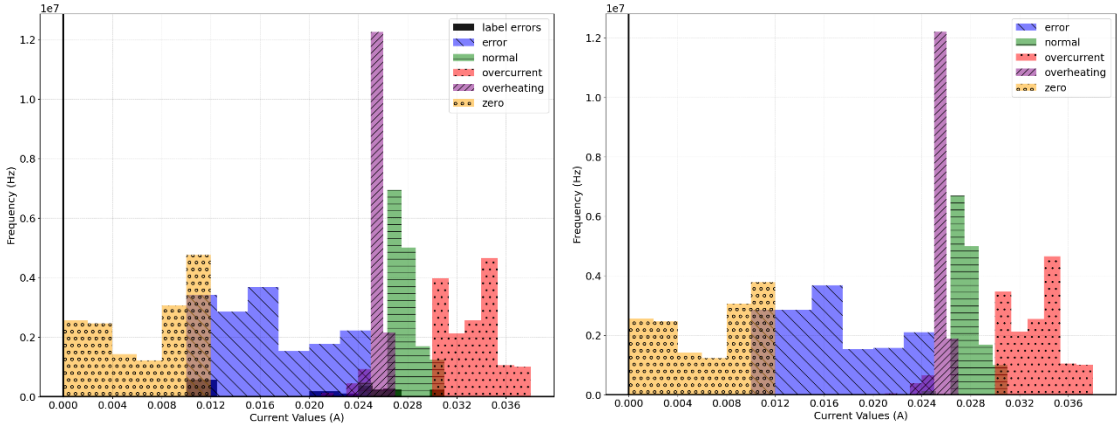
Table 6. *Performance metrics on the test set of the DFT models using the oscillation characteristics of the RMS current (DC component only)*

Metrics	SVM with DC component		RF with DC component		XGBoost with DC component	
	Without CL	CL applied	Without CL	CL applied	Without CL	CL applied
Precision	0.899	0.941	0.951	0.975	0.944	0.976
Recall	0.896	0.941	0.952	0.975	0.943	0.975
ACC	0.896	0.94	0.951	0.975	0.943	0.975
F1-Score	0.896	0.94	0.951	0.975	0.944	0.975

Table 5 presents a metric comparison between the RMS current models trained and tested on datasets without CL and on the cleaned datasets with CL applied. In both cases, the parameters and conditions for model training are identical. The results show that the cleaned datasets help improve the models' performance metrics by about 2~3%. The metric increase was observed in all the SVM, RF, and XgBoost models suggesting an algorithm-independent improvement via data cleaning. Meanwhile, Table 6 shows the performance metrics of DFT models on test sets. Here, one can also observe a similar trend of improvement where the Confident Learning process helps improve the models' performance metrics by about 2~3%. The performance metrics improved because confident learning helped remove the data points that are likely to be

mislabeled and located in the overlapping region between classes, as illustrated by the black "Label errors" in Figure 6a.

Using only the DC component of the DFT transform, the DFT models have a significantly smaller number of inputs. Subsequently, the DFT models require much less computational time while still being robust in classifying the operational states of the LED bulbs. Table 7 shows a summary of the comparisons. Regarding the computer specifications, we used an Intel Core i5-12400F CPU, an NVIDIA RTX 3060 graphics card, and 32 GB of RAM, working on the Ubuntu 22.04 LTS operating system. This comparison result also emphasized the importance of selecting suitable data features as input for the models to enhance computational efficiency and prediction accuracy.



(a) *(b)*

Figure 6. *Data distribution (a) Before applying confident learning; (b) After applying confident learning*

Table 7. *Survey on resources consumed for training concerning the model's prediction accuracy*

Models	The RMS current models (With CL applied)			The DFT models (With CL applied)		
Model name	SVM	RF	XGBoost	SVM	RF	XGBoost
Number of features	9000	9000	9000	1	1	1
Estimated training time (seconds)	105.1	4.1	8.9	0.4	0.3	0.3
Memory consumption (MBs)	0.55	0.125	125.86	0.5	0.125	139.03515
Number of function calls	1246	251677	11977	1246	251677	11977
Accuracy on the test set	0.941	0.982	0.97	0.94	0.975	0.975

3.2. Overfitting

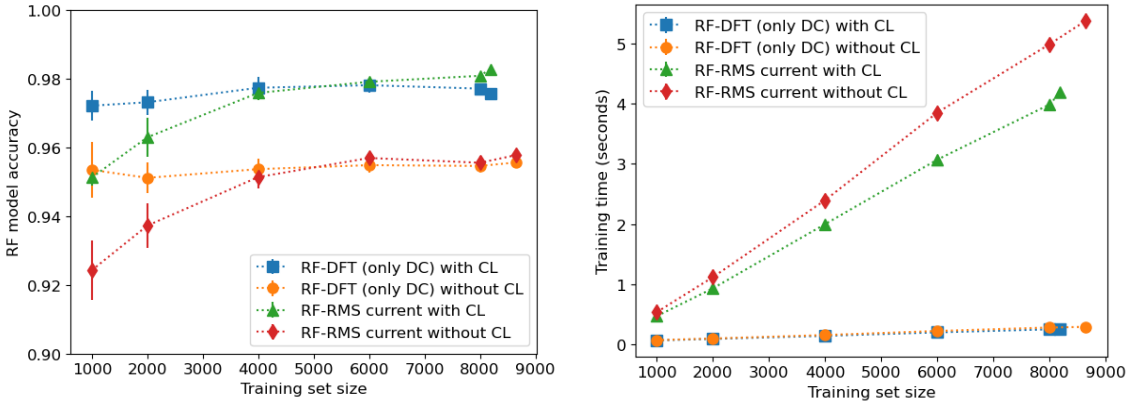
Table 8. *Train, validation, and test set error of models (CL applied) with different feature types*

Models	The RMS Current models (CL applied)			The DFT models (CL applied)		
	Training set error	Validation set error	Test set error	Training set error	Validation set error	Test set error
SVM	0.058	0.072	0.059	0.058	0.072	0.059
RF	1.2e-4	0.032	0.018	4e-3	0.035	0.024
XGBoost	1.2e-4	0.045	0.031	0.021	0.035	0.024

The overfitting status of the models is investigated by comparing the prediction errors that the model made on the train, validation, and test sets. Since all the models under test are classification models, the prediction errors are estimated by subtracting the prediction accuracy from 1, as summarized in Table 8. If a model has a significantly higher error rate on the test and validation sets than on the training set, then the model is more likely overfitting [26]. Table 8 shows that there is not much difference between the test error and the training error of the model,

and there is relative stability between the validation error and the test error. Therefore, one can assume that the models fit the data well and provide reliable prediction results.

3.3. Effects from the size of the training set on prediction accuracy and training time



(a) (b)

Figure 7. *Effects from the size of the training set on 10-time averaged prediction accuracy (a) and training time (b)*

Table 9. *Standard deviation of the 10-time averaged prediction accuracy*

Models \ Size	Standard deviation at different training set sizes				
	1000	**2000**	**4000**	**6000**	**8000**
RF-RMS with CL	0.0023	0.0057	0.0022	0.0012	0.0010
RF-DFT with CL	0.0042	0.0037	0.0031	0.0023	0.0008
RF-RMS without CL	0.0088	0.0064	0.0034	0.0010	0.0020
RF-DFT without CL	0.0082	0.0044	0.0029	0.0020	0.0020

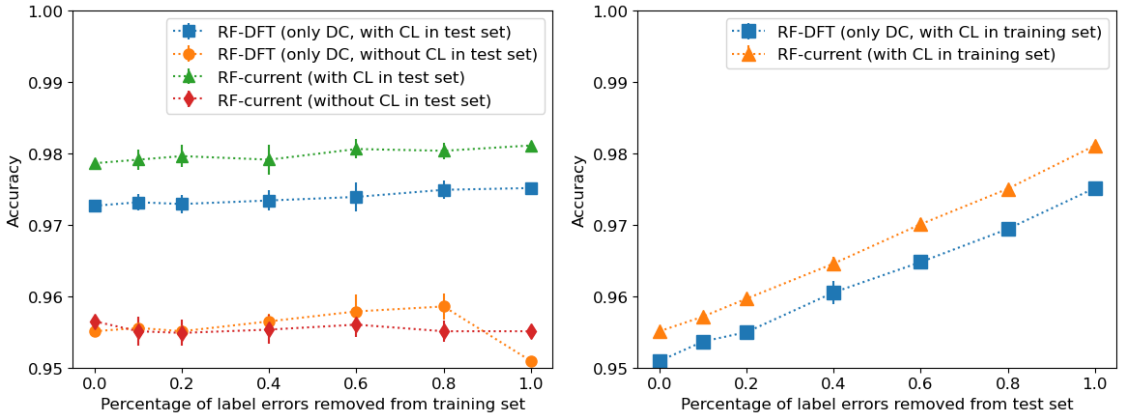
For further analysis, we varied the size of the training set and investigated the fluctuation of the RF model's prediction accuracy on the test set and the model's training time. The RF models were chosen because they have the highest prediction accuracy and the shortest training time. In this investigation, the training set sizes were increased from 1000 to over 8000 samples. With each size, samples were randomly selected and used to train the RF models. The process was repeated ten times. Then, the average values of the model's prediction accuracy were taken. The averaged results are plotted in Figure 7, and the standard deviations are shown in Table 9.

Figure 7a shows that the RF-DFT models have relatively stable prediction accuracy of around 97.5% (with CL) and 95.5% (without CL) when the training set size increases from 1000 to over 8000 samples. Meanwhile, the RF-RMS current models need a larger training set (>4000 samples) to reach similar stable prediction accuracy. When the size of the training set is small (<4000 samples), the DFT models perform better than the RMS current models. One can notice that the improvement in model performance due to the application of CL is around 2-3% at the investigated training set sizes. This improvement shows little dependency on the size of the training set. The RF-DFT model with CL outperforms the RMS current model without CL, especially when the training set is small (<4000 samples).

Figure 7b shows that the DFT models require shorter training time than the RMS current models do in the investigated range of training set sizes. For instance, the training time for the RF-RMS current model with CL at 8000 training set size is 3.98 seconds, while that of the DFT model with CL/without CL is only 0.25 seconds, which is about 16 times faster. When the training set size increases, the training time of all models also increases. However, the rate of change in cases of the DFT models is not as significant as in the case of the RMS current

counterparts. Figure 7b also suggests that CL helps reduce the training time of the RMS current model when the training set size increases. Meanwhile, CL shows almost no effect on the training time of the DFT model when applied. In general, the DFT models are preferable over the RMS current models when dealing with big data because they help reduce the training time while maintaining adequate prediction accuracy.

3.4. Effects from confident learning on the prediction accuracy improvement



(a) *(b)*

Figure 8. *Effects of removing label errors from the training set (a) and test set (b) on 10-time averaged prediction accuracy*

The essence of confident learning is to improve the model's performance by removing label errors from the datasets. Initially, we enforced the removal of all label errors from the datasets. In this section, we investigated the effect of the number of label errors removed on the RF model's performance gain. The experiments examined the model's prediction accuracy when the number of label errors removed increased from 0% (without CL) to 100% (remove all label errors). The percentage of label errors to be removed was taken equally from each label class randomly. The results are summarized in Figure 8. In Figure 8a, we investigated the prediction accuracy of the RF models, with different levels of label error removal from the training set, on the test sets with CL (all label errors removed), and without CL. In Figure 8b, we used RF models trained with a cleaned training set (all label errors removed) to investigate the variation of prediction accuracy on the test set with different levels of label error removal. One may notice that the prediction accuracy at 0 on the x-axes of Figure 8 represents the prediction accuracy without CL applied.

Figure 8a shows that the application of CL in the test set has improved the prediction accuracy of the models by ~1.5% for the RF-DFT models and ~2% for the RF-RMS Current models. Furthermore, when the number of label errors removed from the training set increases, the prediction accuracy on the test set with CL also shows a slowly rising tendency. Meanwhile, a similar but stronger improvement trend in prediction accuracy is observed in Figure 8b when more label errors were removed from the test set. These improvements in both cases may be due to the gradual settlement of the cleaned test set into the applicability domain of the models as more label errors were removed.

3.5. Discussion

In this study, we proposed using confident learning techniques and the oscillation characteristics of the RMS current to improve the prediction efficiency of LED operational state classification models and reduce the computational requirements. The proposal has achieved performance improvements and reduced the number of model inputs. Compared to preceding research in LED light operational-status classification [15], [17], [18], this research used only the

DC component of the DFT transform of the RMS current as the input. As a result, the effects of environmental noises and the stringent requirements in hardware setup are reduced. While the training time is preferably lower, the models still provide a desirable prediction accuracy. More comparison details are shown in Table 10.

Table 10. *Comparison to preceding studies*

Year	Feature type	Algorithm (if any)	Results	Advantages/ Disadvantages	Ref
2024	RMS current and RMS current oscillation	k-NN, SVM, RF, and XGBoost	Best accuracy with RMS current (Random Forest): 98.1% Best accuracy with and RMS current oscillation (XGBoost): 97.5%	Low susceptibility to environmental noise Short training time Only one model input (DFT model)	This study
2023	RMS current	k-NN, SVM, RF, and XGBoost	Best accuracy with RMS current feature (Random Forest): 96.1%	Low susceptibility to environmental noise Many model inputs	[18] (Our previous study)
2023	LED light output time-frequency	SVM	SVM-Quadric accuracy with OSRAM LED data: 89.3% SVM-Quadric accuracy with FSL LED data: 100%	High susceptibility to environmental noise Many model inputs	[15]
2017	The intensity of illumination, lumens maintaining degree, color temperature, color rendering index	Artificial Neural Network, SVM	Best accuracy (SVM): 82.4%	High susceptibility to environmental noise Many model inputs	[17]

Although the model's performance looks promising, one should notice that the dataset included only one specific type of LED bulb among several other types of LED bulbs available on the market. To develop more general and useful models, one may need to include more types of LED bulbs. Furthermore, one could include more features in the input data, such as active power, power factor, reactive power and the frequency characteristics of the LED bulbs.

4. Conclusions

In this paper, we proposed a novel approach to identifying the operational status of household LED lights using non-intrusive load monitoring, machine learning models, confident learning, and the oscillation characteristic of the RMS current. By identifying the oscillation characteristics of the RMS current, we significantly reduced the number of inputs to the models and their computational hardware requirements compared to models using the RMS current. With the introduction of confident learning, we improved the prediction accuracy of the models by 2% on average. The models achieved prediction accuracy ranging from 94% to 97.5%. The proposed method shows potential for applying to different kinds of electrical devices.

REFERENCES

[1] M. M. A. S. Mahmoud, "Economic Applications for LED Lights in Industrial Sectors," in *Light-Emitting Diodes and Photodetectors*, M. Casalino and J. Thirumalai, Eds., Rijeka: IntechOpen, 2021, doi: 10.5772/intechopen.95412.

[2] L. Ciabattoni *et al.*, "A smart lighting system for industrial and domestic use," in *2013 IEEE International Conference on Mechatronics, ICM 2013*, Jan. 2013, pp. 126–131.

[3] N. Narendran and Y. Gu, "Life of LED-based white light sources," *Journal of Display Technology*, vol. 1, pp. 167–171, Jan. 2005.
[4] K. Shahzad *et al.*, "Comparative Life Cycle Analysis of Different Lighting Devices," *Chem Eng Trans*, vol. 45, Jan. 2015, doi: 10.3303/CET1545106.
[5] M. Y. Soh, T. H. Teo, W. X. Ng, and K. S. Yeo, "Review of high efficiency integrated LED lighting," in *2017 IEEE 12th International Conference on Power Electronics and Drive Systems (PEDS)*, 2017, pp. 93–97.
[6] M. M. Aman, G. B. Jasmon, H. Mokhlis, and A. H. A. Bakar, "Analysis of the performance of domestic lighting lamps," *Energy Policy*, vol. 52, pp. 482–500, 2013.
[7] P. Mukherjee, "An Overview of Energy Efficient Lighting System Design for Indoor Applications of an Office Building," *Key Eng Mater*, vol. 692, pp. 45–53, Jan. 2016.
[8] A. Jhunjhunwala *et al.*, "Energy efficiency in lighting: AC vs DC LED lights," *2016 First International Conference on Sustainable Green Buildings and Communities (SGBC)*, 2016, pp. 1–4.
[9] X. Zhan, W. Wang, and H. S. Chung, "A Novel Color Control Method for Multicolor LED Systems to Achieve High Color Rendering Indexes," *IEEE Trans Power Electron*, vol. 33, pp. 8246–8258, 2018.
[10] C.-Y. Chen, M.-D. Ke, J.-H. Wu, and P.-J. Wu, "The Investigation on Physiological Influences and the Working Efficiencies of Several Lighting in Market," in *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2013, pp. 96–99.
[11] T. Clark *et al.*, "LED Luminaire Reliability: Impact of Color Shift," Apr. 2017. [Online]. Available: https://www.energy.gov/sites/prod/files/2017/04/f34/lsrc_colorshift_apr2017.pdf [Accessed Apr. 09, 2024.]
[12] S. Uddin, H. Shareef, A. Mohamed, M. A. Hannan, and K. Mohamed, "LEDs as energy efficient lighting systems: A detail review," in *2011 IEEE Student Conference on Research and Development*, 2011, pp. 468–472.
[13] V. Gupta, B. Basak, and B. Roy, "A Fault-Detecting and Motion-Sensing Wireless Light Controller for LED Lighting System," in *2020 IEEE Calcutta Conference (CALCON)*, 2020, pp. 462–466.
[14] C. Ding and T. Zhang, "Research on health monitoring of LED lighting system," in *2016 Prognostics and System Health Management Conference (PHM-Chengdu)*, 2016, pp. 1–5.
[15] Y. Shang, F. Sun, Q. Fang, B. Chen, and J. Xie, "A novel fault diagnosis strategy for LED lamps via light output time-frequency characteristics analysis and machine learning," *Heliyon*, vol. 9, no. 9, 2023, Art. no. e19737.
[16] A. Freddi, G. Ippoliti, M. Marcantonio, D. Marchei, A. Monteriu, and M. Pirro, "A Fault Diagnosis and prognosis LED lighting system for increasing reliability in energy efficient buildings," in *IET Conference on Control and Automation 2013: Uniting Problems and Solutions*, 2013, pp. 1–6.
[17] H. Jiang, Q. Ma, F. Yang, and M. Shen, "LED device fault diagnosis base on neural network and SVM model analysis," in *2017 14th China International Forum on Solid State Lighting: International Forum on Wide Bandgap Semiconductors China (SSLChina: IFWS)*, 2017, pp. 45–47.
[18] N. S. Nguyen, T. C. Nguyen, H. T. Nguyen, and N. A. Nguyen, "Determining the Operating Status of LEDs Using Non-Intrusive Load Monitoring and Machine Learning," (in Vietnamese) in *National Conference on Electronics, Communications Communications and Information Technology XXVI, REV-ECIT 2023*, 2023, pp. 364–368.
[19] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident Learning: Estimating Uncertainty in Dataset Labels," *Journal of Artificial Intelligence Research (JAIR) (2021)*, vol. 70, pp. 1373–1411, Oct. 2019.
[20] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[21] Y. Li, Y. Kong, M. Zhang, A. Yan, and Z. Liu, "Using Support Vector Machine (SVM) for Classification of Selectivity of H1N1 Neuraminidase Inhibitors," *Mol Inform*, vol. 35, no. 3–4, pp. 116–124, Jan. 2016.
[22] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.
[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. ACM, Aug. 2016, pp. 785–794.
[24] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal Thresholding of Classifiers to Maximize F1 Measure," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2014, pp. 225–239.
[25] C. E. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.
[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.