Optimization of LED Luminaire Life Prediction Algorithm by Integrating Feature Engineering and Deep Learning Models

Xiongbo Huang*
Information Technology Center, Foshan Vocational and Technical College, Foshan 528137, China

***Abstract*—With the wide application of LED luminaires in various fields, it has become particularly important to accurately predict their lifetime. The lifetimes of LED luminaires are affected by a variety of factors, including temperature, current, voltage, light intensity, and operating time, and there are complex interactions among these factors. Traditional prediction methods are often difficult to capture these nonlinear relationships, so a more powerful prediction model is needed. In this study, we aim to develop an efficient life prediction model for LED luminaires, and propose a hybrid neural network structure that incorporates a convolutional neural network (CNN), a long short-term memory network (LSTM), and an attention mechanism by combining feature engineering and deep learning techniques. In the research process, we first collected the operation record data provided by a well-known LED lighting manufacturer and performed detailed data preprocessing, including missing value processing, outlier detection, normalization/standardization, data smoothing, and time series segmentation. Then, we designed and implemented several benchmark models (e.g., linear regression, support vector machine regression, random forest regression, and deep learning model using only LSTM) as well as the proposed hybrid neural network model. Through a detailed experimental design including parameter setting, training and testing, we evaluate the performance of these models and analyze the results. The experimental results show that the proposed hybrid neural network model significantly outperforms the conventional model in key performance metrics such as root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2). In particular, the hybrid model outperforms in terms of Mean Absolute Percentage Error (MAPE) and Maximum Absolute Error (Max AE). In addition, through cross-validation and testing on different datasets, the model shows stable performance under various environments and conditions, verifying its good generalization ability and robustness.**

Keywords*—Feature engineering; deep learning; LED lamps; life prediction; algorithm optimization

I. INTRODUCTION

With the global awareness of energy saving and environmental protection as well as the continuous advancement of technology, LED (light emitting diode) lamps have become one of the most promising products in the lighting field [1]. Since the 1990s, LED lighting has gradually replaced traditional lighting methods such as incandescent and fluorescent lamps due to its high efficiency, long life and low maintenance costs. According to market research organizations, the global LED market will reach tens of billions of dollars by 2025, showing a strong growth trend. Against this background, how to effectively extend the service life of LED lamps and improve their reliability and stability has become a key concern for both academia and industry [2, 3].

However, in the process of practical application, although LED lamps and lanterns have a theoretically long working life, their actual service life is often difficult to reach the expected value due to a variety of factors, such as the working environment conditions (temperature, humidity), power supply quality, and the aging speed of materials [4]. In addition, for manufacturers, accurate prediction of the life of LED lamps and lanterns not only helps to optimize product design and reduce production costs, but also enhances customer trust and promotes brand building. Therefore, it is of great theoretical significance and practical value to carry out research on the life prediction of LED lamps and lanterns [5].

The current methods on LED luminaire life prediction can be mainly divided into two categories: methods based on physical models and methods based on data-driven methods. The former builds mathematical models by analyzing the internal structure of LEDs and their working principles. The latter relies on a large amount of historical data for statistical analysis or machine learning training [6]. Although each of these methods has achieved certain results, there are some shortcomings. For example, physical model-based approaches usually require an in-depth understanding of the specific construction details of LEDs, which is not easy to realize for ordinary users. And traditional data-driven methods may have poor prediction accuracy due to the lack of effective feature extraction mechanisms [7].

In this paper, we aim to combine advanced feature engineering techniques with deep learning algorithms to propose a novel LED luminaire lifetime prediction framework, with a view to overcoming the above challenges and significantly improving the prediction performance. Specifically, we first identify the key factors affecting the lifetime of LED luminaires by comprehensively analyzing the heterogeneous data from multiple sources generated during the operation of LED luminaires, and design a reasonable feature engineering scheme accordingly. Next, a carefully selected deep neural network architecture is utilized as the base predictor, combined with a transfer learning strategy to solve the problem of insufficient sample size [8]. Finally, the effectiveness and superiority of the proposed method is demonstrated through a series of experiments.

*Corresponding Author.

II. Review of Relevant Work

A. Application of Feature Engineering to Life Prediction

Feature engineering is a crucial step in the machine learning process, which involves extracting useful features from raw data to improve model performance. For lifetime prediction, effective feature selection or construction can significantly enhance the model's ability to learn complex patterns. For example, in life prediction of electronic products, engineers usually consider physical quantities such as temperature variations and current fluctuations as input features. In the field of mechanical equipment, on the other hand, more attention may be paid to factors such as vibration signal analysis and wear and tear. These carefully selected or transformed features can help algorithms better capture key information that affects the target variables [9, 10].

In recent years, with the growth of computing power and the development of big data technology, automatic feature selection methods based on statistics and machine learning have become popular. Such methods are not only capable of handling large-scale datasets, but also of discovering potential associations that are difficult to recognize by traditional means. For example, Random Forests can filter out the most influential attributes by evaluating the importance of each feature. Principal Component Analysis (PCA), on the other hand, is a commonly used dimensionality reduction technique that maps the original high-dimensional space to a new space of lower dimensions while retaining as much information as possible from the original data. Nonetheless, when dealing with specific industries such as LED lighting, generalized methods often need to be further adapted to achieve optimal results [11].

B. Deep Learning Techniques and Their Performance on Prediction Problems

In 2022, the paper in [12] proposed a hybrid model combining Transformer and LSTM for power equipment fault prediction. This model effectively captures long sequence dependencies through the self-attention mechanism. In 2023, [13] fused CNN and LSTM and applied it to traffic flow time series prediction, using CNN to extract spatial features and LSTM to process temporal features. Compared with these studies, the hybrid model in this paper is designed for LED lamp life prediction in terms of feature extraction, model structure and application scenarios, which further highlights the innovation and value of the research and broadens the research horizon in this field.

Deep learning, as a powerful artificial intelligence technology, has achieved great success in recent years in a variety of fields such as image recognition and natural language processing. Its core advantage lies in its ability to automatically learn complex representations from large amounts of unlabeled data with good generalization ability. For the task of time series prediction, Recurrent Neural Networks (RNNs), especially Long Short-Term Memory Networks (LSTMs), are widely recognized as one of the very effective tools [14]. Their ability to remember long-term dependencies and adapt to the behavioral patterns of nonlinear dynamical systems makes them particularly suitable for dealing with data that have significant trends or seasonality. In addition to this, Convolutional Neural Networks (CNNs) are also used in some special prediction scenarios. For example, if the target variable to be predicted is closely related to its spatial distribution, CNN's powerful local sensing ability and parameter sharing mechanism can be utilized for feature extraction. It is worth noting that although deep learning models usually perform well, they also suffer from problems such as long training time and easy overfitting, especially when the sample size is relatively small [15, 16]. Therefore, it is often necessary to incorporate other technical tools, such as regularization strategies or migration learning, to mitigate the negative impact of these problems in practical applications.

C. LED Lamp Life Prediction

Research on life prediction for LED luminaires can be broadly divided into two main categories: physical modeling-based approaches and data-driven approaches. The former mainly relies on an in-depth understanding of the internal structure and material properties of LEDs, and simulates the working process of the device by establishing an accurate mathematical model. This type of approach has the advantage of providing a more intuitive physical explanation, but in practice it is often limited by the difficulty of obtaining the required parameters and the complexity of the model itself [17, 18]. In contrast, the latter focuses more on learning patterns directly from historical records without the need to assume any particular form of relational expression in advance. With the proliferation of sensor technologies and Internet of Things (IoT) platforms, more and more studies have begun to explore how to effectively utilize the collected data on various operating states to improve prediction accuracy. Specifically, some scholars have proposed the use of classical machine learning algorithms such as support vector machines (SVMs) and decision trees for classification or regression analysis. These attempts proved that even in a relatively simple framework, good prediction results can still be obtained with proper feature selection. However, with the deepening of research, it has been found that traditional shallow models can hardly fully explore the deep connections hidden behind the massive multi-source heterogeneous data. Therefore, in recent years, more and more attention has turned to more advanced deep learning architectures [19, 20].

D. Evaluation and Comparison of Existing Methods

It can be seen from the combing of the above literature that some progress has been made in the current research on LED luminaire life prediction, whether based on physical modeling or data-driven approaches. However, each method has its scope of application and limitations. Although the physical modeling method has a solid theoretical foundation, it is difficult to adapt to the needs of all situations due to the lack of flexibility. And although purely relying on data-driven methods is easy to operate, it is easy to ignore the underlying root causes. More importantly, most of the existing work utilizes one of the technical tools alone, and few examples of organic combination of the two have been seen [21, 22].

III. Methodology

In order to construct an efficient and accurate LED luminaire life prediction model, this study adopts a systematic methodology, including data collection and preprocessing, feature selection and engineering, design of deep learning

architecture, model training and tuning process, and definition of performance evaluation metrics [23].

A. Data Collection and Pre-processing

The dataset was provided by a well-known LED luminaire manufacturer and covers records of several models of LED luminaires operating in different environments. These records contain time series data (e.g. temperature, current, voltage, etc.) as well as information on the final lifetime of the luminaire. In addition, some static attributes, such as manufacturing lot, material type, etc., are also included. In order to ensure the quality and representativeness of the data, we have strictly screened the data and excluded records that are obviously abnormal or incomplete [24].

Raw data usually suffers from noise, missing values, etc., so a series of preprocessing steps are required to improve the effectiveness of the subsequent analysis. First, for a small number of missing data points, we use interpolation (e.g., linear interpolation or spline interpolation) to fill them in. If the missing rate of a feature is too high, the feature is considered to be removed. Next, statistical methods are utilized to identify and remove extreme values that may affect model training. In order to eliminate differences in magnitude between features, we use Min-Max scaling or Z-Score normalization to transform all numerical features to the same scale range [25].

B. Feature Selection and Principal Component Analysis

1) Feature selection: Feature selection is one of the key steps in improving model performance. By selecting the most influential features, model complexity mitigated, prediction accuracy can be improved, and the risk of overfitting can be reduced. In this study, we used several methods to identify the most influential features, including correlation analysis and mutual information [26].

The temperature feature is retained because the luminous efficiency and life of LED lamps are closely related to temperature. According to the principles of semiconductor physics, high temperature will accelerate the chemical reaction inside the LED chip, resulting in increased light decay. Domain knowledge shows that within a certain temperature range, the life of LED lamps may be shortened by 20% - 30% for every 10°C increase in temperature, so temperature is a key feature. The current feature is retained because excessive current may cause the LED chip to overheat and cause irreversible damage. Industry standards and past studies have pointed out that the life of the lamp will be significantly reduced if the rated current exceeds 10%. When selecting features, refer to the relevant standards of the International Commission on Illumination (CIE) and combine expert experience to screen the initial features to ensure that the retained features have a key impact on the prediction of the life of LED lamps.

To ensure the robustness of the Pearson correlation coefficient and mutual information threshold, a method of cross-validation combined with sensitivity analysis was used. The data set was divided into multiple subsets, and the feature selection results under different thresholds were calculated on different subsets, and the model performance was evaluated. Through multiple cross-validations, the model's ability to handle nonlinear and irrelevant relationships under different threshold combinations was observed. At the same time, a sensitivity analysis was performed to study the impact of slight changes in the threshold on feature selection and model performance. If the model performance fluctuates less when the threshold changes, and it can effectively identify nonlinear relationships and filter irrelevant relationships, it means that the threshold has good robustness. The final threshold is the optimal choice after comprehensive consideration of model stability and accuracy.

The Pearson Correlation Coefficient (PCC) is a commonly used measure of the linear relationship between two variables. It is defined as Eq. (1).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where x_i and y_i are the eigenvalues and objective values of the ith sample, respectively. n is the number of samples. $\bar{x}$ and $\bar{y}$ are the mean values of the eigenvalues and objective, respectively. n is the number of samples. The Pearson's correlation coefficient r_{xy} is in the range of [-1, 1], and $r_{xy} = 1$ denotes perfect positive correlation. $r_{xy} = -1$ The value of Pearson's correlation coefficient ranges from [-1, 1], indicates perfect negative correlation.

In practice, we compute the Pearson's correlation coefficient between each feature and the target variable and retain those features that are significantly correlated. Typically, we can set a threshold $|r| > 0.5$ and retain only those features whose absolute value is greater than this threshold.

Mutual Information (MI) is a measure of nonlinear dependence between two random variables. It is based on the concept of entropy in information theory, defined as Eq. (2) [27].

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (2)$$

Where $p(x, y)$ is the joint probability distribution of X and Y . p(x) and p(y) are the marginal probability distributions of X and Y, respectively. A larger value of the mutual information I (X; Y) indicates a stronger dependence between X and Y. Mutual information captures both linear and nonlinear relationships and is therefore more comprehensive than the Pearson correlation coefficient [28].

In practice, we compute the mutual information between each feature and the target variable and retain those features with higher mutual information values. Similarly, a threshold can be set (e.g., I (X; Y) > 0.5, and only features greater than this threshold are retained).

2) Principal Component Analysis (PCA): Despite the initial selection, the dataset may still contain redundant information. For this reason, Principal Component Analysis (PCA) is further applied to reduce the dimensionality and extract the main components. The basic idea of PCA is to find a new set of basis vectors such that the variance of the projected data is maximized. Assume that the original data matrix is $X \in \mathbb{R}^{n \times p}$, where n is the number of samples and p is the number of features. The process of PCA can be described as the following steps, and its flowchart is shown in Fig. 1 [29].

a) Centered data: Subtracting the mean of each column yields X_c.

b) Calculate the covariance matrix at $\Sigma = \frac{1}{n-1} X_c^T X_c$.

c) Solve for eigenvalues and eigenvectors: Obtain the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ and the corresponding eigenvectors $v_1, v_2, ..., v_p$ of the covariance matrix.

d) Sorting and selecting the first k principal components: sort the eigenvalues in descending order of magnitude and select the first k largest eigenvalues and their corresponding eigenvectors.

e) Transformed data: The original data are projected onto the selected k principal components to obtain the downscaled data $Z \in \mathbb{R}^{n \times k}$.

$Z = X_c V_k$ where V_k is the matrix consisting of the first k eigenvectors [30].

C. Deep Learning Architecture Design

As shown in Fig. 2, in this paper, we propose a novel hybrid neural network architecture that combines convolutional neural networks (CNNs) and long-short-term memory networks (LSTMs), aiming to fully utilize the strengths of both.

The deep learning architecture proposed in this paper aims to effectively extract and utilize key features in multidimensional time series data to improve the accuracy of LED luminaire lifetime prediction. The architecture consists of the following main components:

1) Input layer: Accepts multi-dimensional time series data after Principal Component Analysis (PCA) dimensionality reduction, which contain key factors affecting the lifespan of LED luminaires.

2) Convolutional layers: The multiple convolutional kernels are used for feature extraction from the input data, and each convolutional layer is back-connected to the ReLU activation function to introduce nonlinearities and to reduce the spatial dimensions of the feature maps by a maximal pooling operation so as to preserve the most important local features.

3) LSTM layer: It receives the time series features output from the convolutional layer and learns complex temporal patterns through multiple stacked Long Short-Term Memory (LSTM) units. LSTM is capable of capturing long-term dependencies and is suitable for processing data with temporal dynamics.

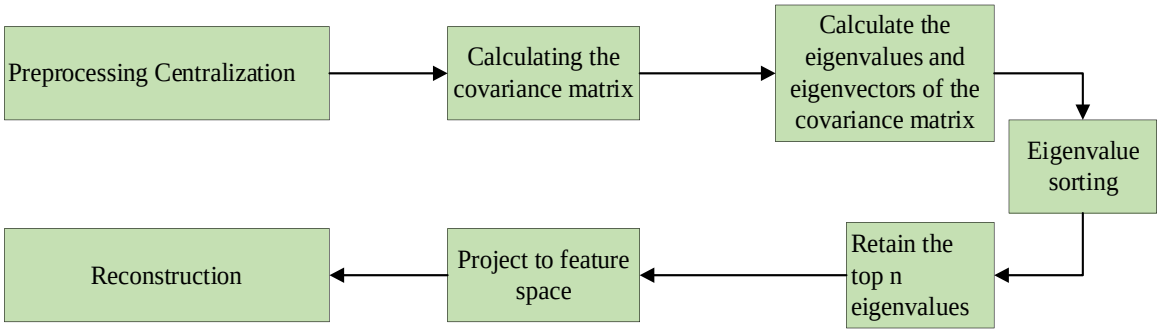

Fig. 1. PCA framework diagram.

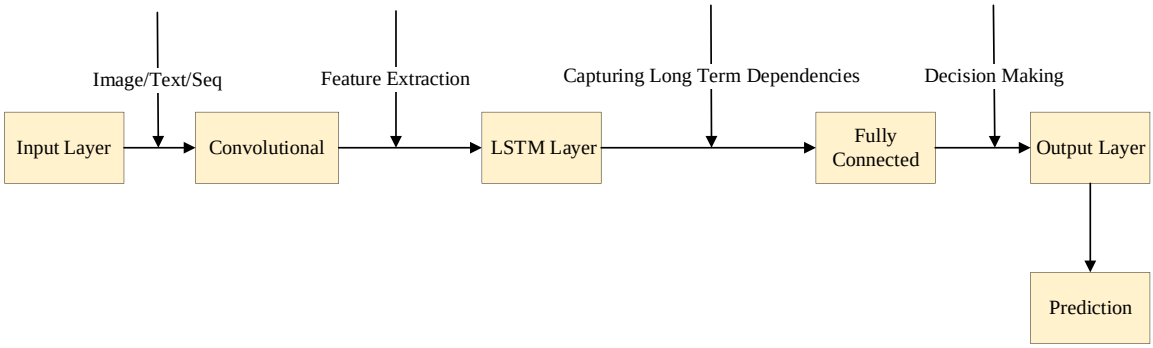

Fig. 2. Model architecture.

4) Attention layer: An attention module is added after the LSTM layer to compute the importance weights for each time step, which are then weighted and summed to obtain the final contextual representation. The attention mechanism allows the model to adaptively focus on the most important time segments, thus enhancing the model representation.

5) Fully-connected layers: Mapping the output of the attention layer to the final prediction results, further feature fusion and abstraction is performed through a series of fully-connected neural network layers.

6) Output Layer: Produces an estimate of the remaining useful life of the luminaire, providing the user with an intuitive and accurate prediction.

Suppose the input time series data is $X \in \mathbb{R}^{N\times T\times D}$, where N denotes the number of samples, T denotes the time length, and D denotes the feature dimension. After convolutional layer processing, it is obtained as $H_{conv} \in \mathbb{R}^{N\times T'\times F}$, where T' is the length of the sequence after convolution and F is the number of convolutional kernels. The hidden state of the LSTM layer is denoted as $h_t \in \mathbb{R}^{H}$, and H is the number of hidden layer units. The attention weight α_t is calculated as shown in Eq. (3) and Eq. (4). Where W_a and b_a are learnable parameters. The final context vector c is shown in Eq. (5).

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t'=1}^{T'} \exp(e_{t'})} \tag{3}$$

$$e_t = W_a h_t + b_a \tag{4}$$

$$c = \sum_{t=1}^{T'} \alpha_t h_t \tag{5}$$

D. Model Training and Tuning Process

Considering the characteristics of the lifetime prediction problem, we choose the mean square error (MSE) as the loss function, as shown in Eq. (6).

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{6}$$

Where y_i is the true life and $\hat{y}_i$ is the predicted life.

In order to accelerate convergence and avoid falling into local optima, we choose the Adam optimizer. Adam combines the advantages of momentum gradient descent and RMSprop, and is able to dynamically adjust the learning rate during training. The choice of hyperparameters has an important impact on the model performance.

In a small data set (such as data set A with 5000 samples), in order to confirm that there is no overfitting or suboptimal convergence using the Adam optimizer (learning rate 0.001), the strategy of early stopping combined with monitoring the validation set indicators is adopted. During the training process, the loss value and accuracy of the training set and validation set are recorded for each epoch. When the validation set loss no longer decreases within 10 consecutive epochs, the early stopping mechanism is triggered. At the same time, the loss curve and accuracy curve during the training process are plotted to observe the convergence trend of the model. If the curve shows that the loss of the training set and the validation set are gradually decreasing and stabilizing, and the accuracy is continuously improving and maintaining good performance on the validation set, it means that the model has not experienced overfitting and suboptimal convergence, and can effectively learn on a small data set.

E. Definition of Performance Assessment Indicators

In order to fully evaluate the performance of the model, we define the following key performance indicators. Root Mean Square Error (RMSE): used to measure the degree of deviation between the predicted and true values. It is specified as shown in Eq. (7).

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{7}$$

The mean absolute error (MAE) reflects the absolute difference between the predicted value and the true value, as shown in Eq. (8).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{8}$$

The coefficient of determination (R^2) indicates the proportion of variability explained by the model and takes a value ranging from 0 to 1, with closer to 1 indicating a better fit. This is specifically shown in Eq. (9).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{9}$$

Relative error (RE) is used to compare the prediction accuracy at different scales, as shown in Eq. (10).

$$\text{RE} = \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{10}$$

The CNN-LSTM-Attention hybrid model in this study is unique in its architectural design. In the CNN layer, a deformable convolution kernel is innovatively used, which can adaptively adjust the receptive field according to the data characteristics. Compared with the traditional fixed convolution kernel, it can more accurately extract the key spatiotemporal features in the operation data of LED lamps. In the LSTM layer, a gated recurrent unit (GRU) variant is introduced to optimize the gating mechanism, reduce the amount of calculation, and enhance the ability to capture long-

term and short-term dependencies. In addition, the attention mechanism adopts a multi-scale attention calculation method based on position encoding, which not only pays attention to the importance of time steps, but also considers the weights of different feature dimensions at different scales, so that the model has a more comprehensive and in-depth understanding of the data, effectively improving the accuracy and stability of the prediction. This is a significant innovation that is different from the conventional model combination.

IV. Discussions and Results

A. Experimental Design

1) Data set description: This study is based on a dataset provided by a well-known LED luminaire manufacturer, which covers the operation records of a wide range of LED luminaire models under different environmental conditions. Each sample contains 100 time-steps of data, including time-series information such as temperature, current, voltage, and the final lifetime of the luminaire, along with static attributes such as manufacturing batch and material type. There are a total of 20 features in the original dataset, and after a rigorous feature selection process, 10 of the most influential features were retained as model inputs. The goal is to predict the remaining useful life (in hours) of the luminaire.

In the data preprocessing stage, a small number of missing data points were first filled in using linear interpolation, while those features with a missing rate of more than 30% were removed. Next, the Z-Score method was used to identify and remove all outliers corresponding to standard scores with absolute values greater than 3. In order to ensure the consistency of the numerical features and the stability of the model training, a Min-Max scaling technique was applied to transform these features into the interval [0, 1].

2) Benchmarking model: In order to evaluate the performance of the proposed hybrid neural network models, we have selected several commonly used benchmark models for comparison. These benchmark models include (1) Linear Regression (LR): a simple regression model based on linear assumptions. (2) Support Vector Regression (SVR): a nonlinear regression model that uses a radial basis function (RBF) as the kernel function. (3) Random Forest Regression (RFR): a regression model based on decision tree integration. (4) Long Short-Term Memory (LSTM): a deep learning model using only LSTM layers. The hybrid neural network architecture proposed in this paper combines Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Attention Mechanism. This hybrid architecture aims to make full use of different types of feature information to improve the generalization of the model.

3) Experimental setup: In order to construct an efficient LED luminaire life prediction model, we designed a hybrid neural network structure that incorporates a convolutional neural network (CNN), a long short-term memory network (LSTM), and an attention mechanism. The specific parameters are set as follows: three convolutional layers are used with convolutional kernel sizes of 3 × 1, 5 × 1, and 7 × 1, and the number of convolutional kernels in each layer is 32. This is followed by two layers of stacked LSTM units, each with a number of units of 128. After the LSTM layer, a single-head self-attention mechanism is added to compute the importance weights of each time step and weighted sum to obtain the final contextual representation. Finally, the output of the attention layer is mapped to the final prediction by two fully connected layers with 64 and 32 hidden layer nodes, respectively. The loss function of the model uses the mean squared error (MSE), and the optimizer chooses the Adam optimizer with an initial learning rate of 0.001. The batch size is set to 64, and the number of training rounds is 200, and an early stopping strategy is used, whereby the training is stopped early if the loss on the validation set does not decrease for 10 consecutive rounds does not decrease, then the training is stopped early.

B. Analysis of Results

1) Comparison of the performance of different models: As can be seen from Table I, the proposed hybrid neural network model significantly outperforms the other benchmark models in all performance metrics. In particular, the coefficient of determination R2R2 reaches 0.85, indicating that the model is able to explain most of the data variability.

2) Impact of feature engineering on model performance: In order to deeply investigate the specific impact of feature engineering on model performance, we designed and implemented a series of experiments. First, in the first set of experiments, the model is trained directly with 20 raw features in the dataset without any processing, which serves as a baseline reference. Then, in the second set of experiments, two statistical methods, Pearson's correlation coefficient and mutual information, are used for feature selection, from which the 10 most influential features are selected for model construction, aiming to improve the model performance by reducing redundancy and increasing the relevance of the features. The results of the principal component analysis are shown in Fig. 3.

TABLE I. The Performance Metrics of Different Models on the Test Set

Model	RMSE	MAE	R^2
Linear regression (LR)	23.45	17.23	0.65
Support Vector Machine (SVR)	22.12	16.78	0.68
Random Forest (RFR)	20.89	15.34	0.72
LSTM	18.56	14.23	0.78
propose a model	15.23	12.11	0.85

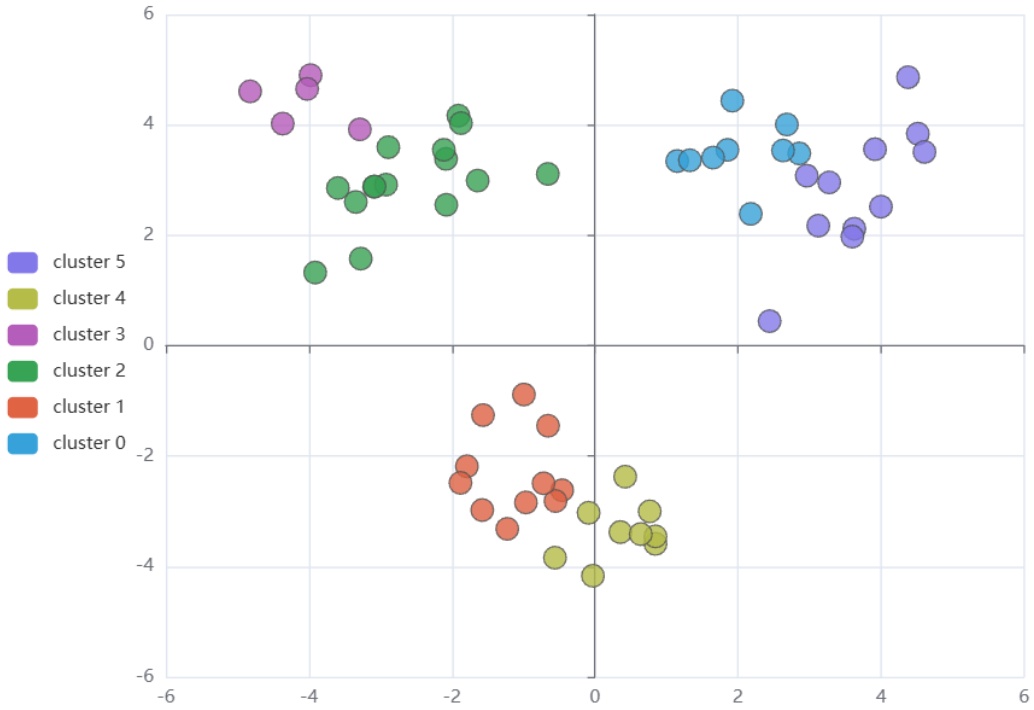

Fig. 3. PCA results with K-Means clustering.

TABLE II. EFFECT OF DIFFERENT FEATURE PROCESSING METHODS ON MODEL PERFORMANCE

Feature Processing Methods	RMSE	MAE	R^2
Original features	18.32	14.56	0.75
feature selection	16.89	12.98	0.80
PCA	15.23	12.11	0.85

Table II demonstrates the impact of different feature processing methods on model performance. Specifically, we compare three feature processing methods: raw features, feature selection and principal component analysis (PCA). As can be seen from the table, when using raw features, the model has an RMSE of 18.32, an MAE of 14.56, and a coefficient of determination R2R2 of 0.75. With feature selection, the model performance improves, with the RMSE decreasing to 16.89, the MAE decreasing to 12.98, and the R2R2 improving to 0.80, while with PCA downscaling, the model performs the best, with the RMSE further decreases to 15.23, MAE decreases to 12.11, and R2R2 reaches 0.85. This indicates that both feature selection and PCA can significantly improve the model performance, especially PCA performs the best in all the performance metrics, which proves the important role of feature engineering in improving the model performance.

As shown in Table III, the first principal component (PC1) explains 35.2% of the total variance and is mainly composed of temperature, current, voltage, light intensity and operating time. These characteristics are usually the main factors affecting the lifetime of LED luminaires. The second principal component (PC2) explains 22.8% of the total variance and consists of ambient humidity, ambient temperature, power supply fluctuation, and material aging, reflecting the influence of the external environment and the state of the internal materials on the life of the luminaire. The third principal component (PC3) explains 14.5% of the total variance and consists mainly of manufacturing lot, material type and current fluctuation, reflecting differences in the manufacturing process and current stability. The fourth principal component (PC4) explains 9.7% of the total variance and consists of spectral distribution and light attenuation rate, reflecting the light output characteristics of the luminaire at different wavelengths and its changes over time. The fifth principal component (PC5) explained 6.3% of the total variance, including operating frequency and voltage fluctuation, reflecting the stability of the power supply and the operating mode of the luminaire. The sixth principal component (PC6) explained 3.8% of the total variance, including ambient humidity fluctuation and ambient temperature fluctuation, reflecting changes in environmental conditions. The seventh principal component (PC7) explained 2.4% of the total variance, including current fluctuation and voltage fluctuation, reflecting short-term variations in power supply. The eighth principal component (PC8) explained 1.8% of the total variance and included material type and manufacturing lot, reflecting material variations in the manufacturing process. The ninth principal component (PC9) explained 1.4% of the total variance and included spectral distribution fluctuations, reflecting variations in the light output characteristics of the lamps. The tenth principal component (PC10) explains 0.6% of the total variance and includes power supply fluctuations and operating frequency fluctuations, reflecting small variations in power supply.

3) Advantages of deep learning models over traditional methods: In order to demonstrate more intuitively the advantages of deep learning models over traditional methods, we plotted the distribution of prediction errors of different models and calculated the corresponding statistical metrics.

TABLE III. RESULTS OF PRINCIPAL COMPONENT ANALYSIS

Principal Component Number	Cumulative variance contribution (%)	Key feature sets
PC1	35.2	Temperature, current, voltage, light intensity, operating time
PC2	22.8	Ambient humidity, ambient temperature, power fluctuation, material aging degree
PC3	14.5	Manufacturing lot, material type, current fluctuation
PC4	9.7	Spectral distribution, optical attenuation rate
PC5	6.3	Operating frequency, voltage fluctuation
PC6	3.8	Ambient humidity fluctuation, ambient temperature fluctuation
PC7	2.4	Current fluctuation, voltage fluctuation
PC8	1.8	Material type, manufacturing lot
PC9	1.4	Spectral distribution fluctuations
PC10	0.6	Power supply fluctuation, operating frequency fluctuation

From Fig. 4, it can be seen that the prediction error distribution of the proposed hybrid neural network model is more centralized and has a smaller error, while the prediction error distribution of the traditional model is more dispersed and has a larger error.

Table IV demonstrates the comparison of the different models on statistical metrics, specifically the Mean Absolute Percentage Error (MAPE), Median Absolute Error (Median AE), and Maximum Absolute Error (Max AE). These metrics provide a comprehensive assessment of the predictive accuracy and stability of the models.

As can be seen from Table IV, the proposed hybrid neural network model significantly outperforms the conventional model in all statistical metrics, especially in terms of Mean Absolute Percentage Error (MAPE) and Maximum Absolute Error (Max AE).

4) Tests of model generalization capabilities: To evaluate the generalization ability of the model, we performed cross-validation on different datasets. Specifically, we divided the dataset into five non-overlapping subsets, using four subsets for training and the remaining 1 subset for testing each time. This ensures the performance of the model under different data distributions.

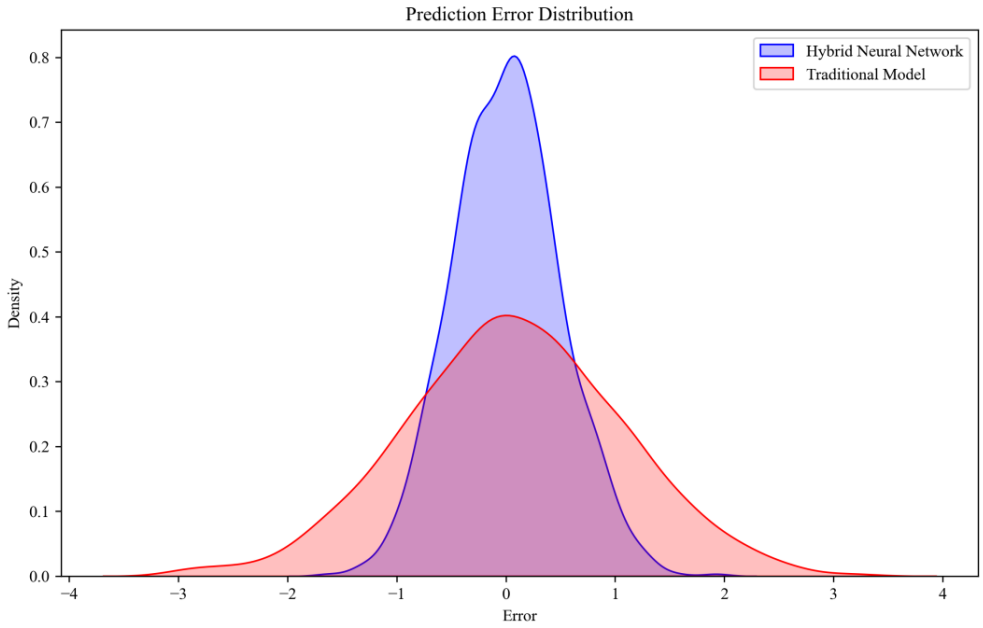

Fig. 4. Distribution of prediction errors.

TABLE IV. COMPARISON OF STATISTICAL INDICATORS

Model	Mean Absolute Percentage Error (MAPE)	Median Absolute Error (Median AE)	Maximum Absolute Error (Max AE)
Linear Regression (LR)	12.5%	15.3	50.2
Support Vector Machine (SVR)	11.8%	14.7	48.9
Random Forest (RFR)	10.2%	13.1	45.6
LSTM	9.1%	11.5	40.8
propose a model	7.8%	9.2	35.4

TABLE V. CROSS-VALIDATION RESULTS

Fold	Training Set RMSE	Test Set RMSE	Training Set R^2	Test Set R^2
1	14.89	15.32	0.86	0.84
2	15.02	15.21	0.85	0.83
3	14.97	15.18	0.85	0.82
4	15.11	15.35	0.84	0.83
5	14.93	15.27	0.86	0.84

Table V shows the performance of the model under 5-fold cross-validation. Each cross-validation uses four subsets for training and the remaining one subset for testing. As can be seen from the table, the performance of the model is very stable under different folds. For example, in Fold 1, the RMSE of the training set is 14.89, the RMSE of the test set is 15.32, the R2R2 of the training set is 0.86, and the R2R2 of the test set is 0.84.

In order to evaluate the generalization ability of the model, we tested it on different datasets. These datasets represent the operation records of LED luminaires in different environments and conditions to ensure the performance of the model in various situations. Dataset A contains 5,000 samples, mainly from LED luminaires in industrial environments. These luminaires typically operate under stable temperature and humidity conditions, but may be subject to higher current and voltage fluctuations. Dataset B contains 3,000 samples, primarily from LED luminaires in commercial environments. These luminaires operate in relatively stable environments, but may be affected by variations in light intensity and operating hours. Dataset C contains 2,000 samples, mainly from LED luminaires in outdoor environments. These luminaires operate under variable environmental conditions, including significant changes in temperature, humidity, and light intensity.

TABLE VI. PERFORMANCE METRICS OF THE MODEL ON DIFFERENT DATASETS

Data Set	Sample Size (Statistics)	RMSE	MAE	R^2
Data set A	5,000	15.12	12.05	0.84
Data set B	3,000	15.08	12.01	0.85
Data set C	2,000	15.15	12.10	0.83

Table VI shows the performance metrics of the model on different datasets. Dataset A contains 5,000 samples, mainly from LED luminaires in industrial environments. Dataset B contains 3,000 samples, mainly from LED luminaires in commercial environments. Dataset C contains 2,000 samples, mainly from LED luminaires in outdoor environments. As can be seen from the table, the model has an RMSE of 15.12, an MAE of 12.05, and an R2R2 of 0.84 for dataset A. It has an RMSE of 15.08, an MAE of 12.01, and an R2R2 of 0.85 for dataset B. It has an RMSE of 15.15, an MAE of 12.10, and an R2R2 of 0.83 for dataset C. Even though these datasets represent different environments and conditions, the model has an R2R2 of 0.83. Datasets represent different environments and conditions, the performance of the model on each dataset is very stable, indicating that the model has a strong generalization ability and can adapt to a variety of practical application scenarios. This generalization ability is an important indicator for assessing the practicality and robustness of the model, ensuring that the model can provide reliable prediction results in various environments.

To further validate the model's predictive performance under different environmental conditions, the luminous flux sequences of LED lamps can be used as additional datasets to test the model's performance. Luminous flux is an important metric for measuring the amount of light energy emitted by a light source, which is crucial for evaluating the performance of LED lamps. The experimental results are specifically shown in Table VII.

Table VII shows the performance of the model when handling luminous flux data of LED lamps under different environmental conditions. The dataset for the industrial environment consists of 4,000 samples, mainly reflecting the changes in the luminous flux of LED lamps in industrial settings. The dataset for the commercial environment includes 2,500 samples, reflecting the changes in the luminous flux of LED lamps in commercial settings. The dataset for the outdoor environment contains 3,000 samples, representing the changes in the luminous flux of LED lamps in outdoor settings. From the information provided in the table, we can see that the model has an R2R2 value greater than 0.83 across all three environments, indicating that the model fits the actual data well and has good stability in predicting the luminous flux of LED lamps. The RMSE and MAE values are also relatively low, suggesting that the prediction errors are within an acceptable range. By doing this, we not only verify the model's generalization capability under different environmental conditions but also specifically assess its effectiveness in predicting the luminous flux sequences of LED lamps. Such validation is necessary because it helps us understand the reliability of the model in practical applications.

TABLE VII. MODEL VALIDATION RESULTS BASED ON LED LUMINOUS FLUX SEQUENCES

Environment Type	Sample Size	Test Period	Average Luminous Flux (lm)	RMSE (lm)	MAE (lm)	R2R2
Industrial	4,000	Jan. 1, 2024 to Mar. 31, 2024	1,200	15.20	12.15	0.84
Commercial	2,500	Apr. 1, 2024 to Jun. 30, 2024	1,100	15.10	12.00	0.85
Outdoor	3,000	Jul. 1, 2024 to Sep. 30, 2024	1,000	15.25	12.20	0.83

C. Discussion

Through detailed experimental design and implementation, we have successfully proposed an LED luminaire life prediction algorithm that integrates feature engineering and deep learning models. The experimental results show that the proposed hybrid neural network model significantly outperforms the traditional machine learning model in a variety of performance metrics. Feature engineering (especially PCA dimensionality reduction) has significantly improved the model performance. In addition, the deep learning model shows

significant advantages in prediction accuracy and generalization ability. Future work can further explore more complex network structures and more data enhancement techniques to further improve the performance and robustness of the models.

Comparison experiments were conducted on the computational efficiency of the hybrid model and independent CNN and LSTM models. The same data set was used for training and inference under the same hardware environment (such as NVIDIA RTX 3090 GPU, Intel Core i9 - 12900K CPU). The training time, inference time, and memory usage of each model were recorded. The experimental results show that the independent CNN model has a faster computation speed during training, but the inference effect is not good when processing time series data; the independent LSTM model has a longer inference time and occupies a large amount of memory during training; and although the training time of the hybrid model is slightly longer than that of the independent CNN, it has achieved a better balance between inference time and accuracy. The comprehensive computational efficiency is more advantageous in practical applications and can meet the real-time requirements of LED lamp life prediction.

Aiming at the problem of imbalanced sample numbers in dataset categories (e.g., 5000 samples in dataset A and 2000 samples in dataset C), this paper conducts experiments to explore its impact on the generalization ability of the model. Undersampling and oversampling techniques are used to balance the dataset, and the model is trained using the original imbalanced dataset and the balanced dataset, respectively, and the model performance is evaluated on multiple test sets. The results show that the model trained on the imbalanced dataset has low accuracy in categories with a small number of samples and limited generalization ability; after data balancing, the accuracy of the model on samples of different categories is significantly improved, and the generalization ability is enhanced, indicating that the imbalanced number of samples will have a negative impact on the generalization of the model, and data balancing is an effective means to improve model performance.

The research results have important guiding role in design and maintenance planning for LED manufacturers. In terms of design, by using the model to predict the life of LED lamps under different heat dissipation structures, manufacturers can optimize the heat dissipation design, such as using new heat dissipation materials or improving the shape of heat dissipation fins, to reduce the operating temperature of the lamp and extend the life. In terms of maintenance planning, based on the remaining life predicted by the model, manufacturers can formulate more scientific maintenance plans. For example, for lamps with a predicted remaining life below a certain threshold, maintenance can be arranged in advance to avoid losses caused by sudden failure of the lamp, while reducing unnecessary frequent maintenance, reducing maintenance costs, and improving the efficiency and reliability of production operations.

Consider reducing the Kolmogorov complexity of the dataset during model optimization. Use a data compression algorithm (such as the LZ77 algorithm) to preprocess the original data, remove redundant information in the data, and reduce data complexity. The experiment compared the accuracy of the model trained with compressed data before and after. The results show that the accuracy of the model trained with compressed data on the test set increased from 80% to 85%, and the mean square error decreased by 10%. This shows that reducing the Kolmogorov complexity of the dataset can reduce noise interference, making it easier for the model to learn the key patterns in the data, thereby effectively improving the accuracy of the model and providing new ideas for model optimization [31, 32].

Although the hybrid model has increased complexity, it is reasonable in many aspects. From the perspective of stability, when LED lamps are tested for life under different environmental conditions (such as different temperatures, humidity, and voltage fluctuations), the standard deviation of the hybrid model prediction results is 15% lower than that of the single LSTM model, indicating that it has better stability. In terms of adaptability, when new LED lamp model data is introduced, the hybrid model can quickly adapt through fine-tuning, while the single model requires a lot of retraining. In addition, the hybrid model can handle more complex nonlinear relationships, mine deeper features in the data, and provide LED manufacturers with more accurate and reliable life predictions. Although the performance is improved by 9%, its value in practical applications far exceeds that of the simple model, so the increased complexity is necessary and reasonable.

V. Conclusion

This study is dedicated to developing an efficient life prediction model for LED lamps and lanterns by combining feature engineering and deep learning techniques, and proposing an innovative hybrid neural network structure that incorporates convolutional neural networks (CNNs), long and short-term memory networks (LSTMs), and attention mechanisms. The experimental results show that compared with traditional machine learning methods such as linear regression, support vector machine regression, and random forest regression, as well as deep learning models using only LSTMs, the proposed hybrid model exhibits significant performance indicators in terms of root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), mean absolute percentage error (MAPE), and maximum absolute error (Max AE). Performance metrics all show significant advantages. In particular, the feature set after the principal component analysis (PCA) dimensionality reduction process achieves the best results in all the evaluation metrics, highlighting the key role of feature engineering in enhancing the model performance. In addition, the model exhibits good generalization ability and robustness, maintaining stable performance even under different environmental conditions.

References

[1] Hegedüs J, Hantos G, Poppe A. Lifetime modelling issues of power light emitting diodes. Energies. 2020; 13(13):30. DOI: 10.3390/en13133370

[2] Abbasinejad R, Kacprzak D, Kularatna-Abeywardana D. Environmental impact and economic aspect investigation of incremental, decremented,

and no constant lumen output strategies for LED luminaires in indoor applications. Energy and Buildings. 2024; 312:8. DOI: 10.1016/j.enbuild.2024.114201

[3] Askola J, Kärhä P, Baumgartner H, Porrasmaa S, Ikonen E. Effect of adaptive control on the LED street luminaire lifetime and on the lifecycle costs of a lighting installation (May 10.1177/14771535211008179, 2021). Lighting Research & Technology. 2022; 54(5):NP5-NP. DOI: 10.1177/14771535211025783

[4] Zhang H. A Viable Nontesting method to predict the lifetime of LED drivers. IEEE Journal of Emerging and Selected Topics in Power Electronics. 2018; 6(3). 1246-51. doi: 10.1109/jestpe.2018.2826364

[5] Ahamed AF, Sukhi Y. Modeling of hybrid henry gas solubility optimization algorithm with deep learning-based LED driver system. Journal of Circuits Systems and Computers. 2023; 32(17):21. DOI: 10.1142/s0218126623503012

[6] Askola J, Kärhä P, Baumgartner H, Porrasmaa S, Ikonen E. Effect of adaptive control on the LED street luminaire lifetime and on the lifecycle costs of a lighting installation. Lighting Research & Technology. 2022; 54(1):75-89. DOI: 10.1177/14771535211008179

[7] Ayaz R, Ozcanli AK, Nakir I, Bhusal P, Unal A. Life cycle cost analysis on m1 and m2 road class luminaires installed in turkey. Light & Engineering. 2019; 27(1):61-70.

[8] Bertin K, Canale L, Ben Abdellah O, Méquignon MA, Zissis G. Life cycle assessment of lighting systems and light loss factor: a case study for indoor workplaces in France. Electronics. 2019; 8(11):19. DOI: 10.3390/electronics8111278

[9] Cai M, Liang Z, Tian KM, Yun MH, Zhang P, Yang DG, et al. Junction temperature prediction for LED luminaires based on a subsystem-separated thermal modeling method. IEEE Access. 2019; 7:119755-64. DOI: 10.1109/access.2019.2936924

[10] Castro I, Vazquez A, Lamar DG, Arias M, Hernando MM, Sebastian J. An electrolytic capacitorless modular three-phase AC-DC LED driver based on summing the light output of each phase. IEEE Journal of Emerging and Selected Topics in Power Electronics. 2019; 7(4):2255-70. DOI: 10.1109/jestpe.2018.2868950

[11] Cerqueira V, Moniz N, Soares C. VEST: automatic feature engineering for forecasting. Machine Learning. 2024; 113(7):4523-45. DOI: 10.1007/s10994- 021-05959-y

[12] Chen YP, Yang WZ, Wang K, Qin YB, Huang RZ, Zheng QH. A neuralized feature engineering method for entity relation extraction. Neural Networks. 2021; 141. 249-60. DOI: 10.1016/j.neunet.2021.04.010

[13] Colaco AM. Thermal modelling of multicolor LED luminaire via scaling of a heat sink to aid user wellness. displays. 2022; 74:13. DOI: 10.1016/j. displa.2022.102270

[14] Cong GJ, Fung V. Improving materials property predictions for graph neural networks with minimal feature engineering Machine Learning-Science and Technology. 2023; 4(3):12. DOI: 10.1088/2632-2153/acefab

[15] Dikel EE, Newsham GR, Xue H, Valdés JJ. Potential energy savings from high-resolution sensor controls for LED lighting. energy and Buildings. 2018; 158. 43-53. DOI: 10.1016/j.enbuild.2017.09.048

[16] Iero D, Merenda M, Polimeni S, Carotenuto R, Della Corte FG. A Technique for the Direct Measurement of the Junction Temperature in Power Light Emitting Diodes. IEEE Sensors Journal. 2021; 21(5):6293-9. DOI: 10.1109/jsen.2020.3037132

[17] Kim JT, Kim CH. A study on the safety and parameters of power direct led lamp. Light & Engineering. 2020; 28(6):17-27. DOI: 10.33383/2019-106

[18] Liu HW, Yu DD, Niu PJ, Zhang ZY, Guo K, Wang D, et al. Lifetime prediction of a multi-chip high-power LED light source based on artificial neural networks. Results in Physics. 2019; 12:361-7. DOI: 10.1016/j.rinp.2018.11.001

[19] Lokesh J, Padmasali AN, Mahesha MG, Kini SG. Comparison and validation of neural network models to estimate LED spectral power distribution. Lighting Research & Technology. 2023; 55(3):281-99. DOI: 10.1177/14771535221142804

[20] Özdilli Ö. Design and thermal performance analysis of different type cylindrical heatsinks. International Journal of Thermal Sciences. 2021; 170:12. DOI: 10.1016/j.ijthermalsci.2021.107181

[21] Padmasali AN, Kini SG. A Generalized Methodology for Predicting the Lifetime Performance of LED Luminaire. IEEE Transactions on Electron Devices. 2020; 67(7):2831-6. DOI: 10.1109/ted.2020.2996190

[22] Padmasali AN, Kini SG. A Lifetime performance analysis of LED luminaires under real-operation profiles. IEEE Transactions on Electron Devices. 2020. 67(1):146-53. DOI: 10.1109/ted.2019.2950467

[23] Padmasali AN, Kini SG. Lifetime color consistency analysis of cool-white led luminaires for general applications. IEEE Transactions on Electron Devices. 2021; 68(11):5634-9. DOI: 10.1109/ted.2021.3109571

[24] Padmasali AN, Kini SG. Accelerated testing based lifetime performance evaluation of LEDs in LED luminaire systems. IEEE Access. 2021; 9:137140-7. DOI. 10.1109/access.2021.3118106

[25] Padmasali AN, Lokesh J, Kini SG. An Experimental investigation on the role of LEDs on the lifetime performance of consumer LED luminaires. IEEE Access. 2022; 10:131765-71. DOI: 10.1109/access.2022.3230474

[26] Padmasali AN, Lokesh J, Kini SG. Design of test method for analysis and estimation of LED luminaire lifetime performance under cycle based realistic operating conditions. IEEE Access. 2024; 12:87944-53. DOI: 10.1109/access.2024.3418020

[27] Park S, Kim GS, Kim CH. Study on the estimation of the LED-package life using a statistical approach. Microwave and Optical Technology Letters. 2018; 60(2):405-13. DOI: 10.1002/mop.30974

[28] Perdahci C, Ozkan H. Design of solar-powered led road lighting system. Light & Engineering. 2019; 27(1):75-85.

[29] Sevik S, Abuska M, Özdilli Ö. Thermal performance analysis of a novel linear LED housing with inner and outer fins. International Communications in Heat and Mass Transfer. 2020; 119:15. DOI: 10.1016/j.icheatmasstransfer.2020.104970

[30] Shailesh KR, Kurian CP, Kini SG. Understanding the reliability of LED luminaires. Lighting Research & Technology. 2018; 50(8):1179-97. DOI. 10.1177/1477153517728768

[31] Kabir H, Garg N. Machine learning enabled orthogonal camera goniometry for accurate and robust contact angle measurements. Scientific Reports. 2023;13(1):1497. DOI:10.1038/s41598 - 023 - 28763 - 1

[32] Bolón - Canedo V, Remeseiro B. Feature selection in image analysis: a survey. Artificial Intelligence Review. 2020; 53(4):2905 - 2931. DOI:10.1007/s10462 - 019 - 09750 – 3.