Integration of Machine Learning-Based CIE Standard Skies Model With Daylight Simulation for Building Energy Performance Analysis

Emmanuel Imuetinyan Aghimien – Hong Kong Metropolitan University – emmanuel.aghimien@utah.edu
Ernest Kin-wai Tsang – Hong Kong Metropolitan University – ekwtsang@hkmu.edu.hk
Danny Hin-wa Li† – City University of Hong Kong
Zhenyu Wang – City University of Hong Kong – zwang598-c@my.cityu.edu.hk

Abstract

Daylight illuminance data is required for daylight scheme evaluations.while the adoption of daylight-linked control (DLLC) systems is a useful strategy for attaining energy savings on lighting. For evaluations of these daylighting schemes and DLLC, determining the sky conditions through sky luminance distributions is required. Moreover, through these distributions, the 15 CIE standard skies can be identified and daylight illuminance for any surface of interest can be derived. The crucial issue is that sky luminance data is sparingly measured. Recent studies have shown that the use of accessible climatic data and machine learning (ML) models for determining the standard skies can be viable alternatives. In this study, an ensembled-based Light Gradient Boosting Machine (LGBM) was used to identify the standard sky types in Hong Kong. The predictions of the LGBM model were then integrated with RADIANCE and EnergPlus for daylight and building energy simulations of a generic shopping mall. The simulation was carried out by comparing the Best fit, ASRC 1992 and the All-weather models against the measured data. Findings show that when tested, the LGBM model correctly classified the sky types over 70% of the time. Similarly, when used for daylight and energy simulations acceptable predictions were obtained from all models. Finally, it was found that the impact of the sky luminance distribution model on illuminance prediction is higher than that for energy estimation.

1. Introduction

In Hong Kong, energy spent on lighting contributes 10% to the electricity demand (Electrical & Mechanical Services Department [EMSD], 2023). Hence, attaining a balance between the provision of adequate lighting and the reduction of building energy consumption on lighting is crucial for energy-efficient designs (Aghimien et al., 2021). Even when daylight is admitted into buildings, occupants tend to rely on artificial lighting. Technically, daylight is preferable because it provides visual comfort and energy savings (Aghimien et al., 2022). One of the approaches for attaining daylight energy savings is the use of daylight-linked control (DLLC) systems since these can reduce building energy consumption without compromising the visual comfort in a room (Bellia et al., 2016). In cases where measurements are not possible to evaluate these controls, simulation techniques are used.

Furthermore, for daylight assessment, the daylight illuminance in a room is hugely dependent on the sky luminance distributions. Unfortunately, ground measurement of luminance distributions is rarely carried out, and the sky scanners used for such measurements are expensive (Granados-López et al., 2021). In 2003, the International Commission on Illumination (CIE) proposed using the 15 CIE standard skies which covers the whole spectrum of skies in the world to classify the skies (CIE, 2003). Apart from actual measurement, daylight illuminance data can be obtained and the sky condition, determined. However, the criteria for determining the sky conditions using this approach are sometimes based on the vertical and zenith illuminance or irradiance data which are not readily available (Lou et al., 2017). Moreso, meteorological parameters can be used for identifying the skies. However, whether these parameters can correctly identify the sky conditions is still a major point of discussion (Aghimien et al., 2022). In this study, an ensemble-based

Part of
Pernigotto, G., Ballarini, I., Patuzzi, F., Prada, A., Corrado, V., & Gasparella, A. (Eds.). 2025. Building simulation applications BSA 2024. bu,press.
https://doi.org/10.13124/9788860462022

machine learning (ML) model called Light Gradient Boosting Machine (LGBM) is proposed to identify the 15 CIE standard skies using readily available climatic variables. The rationale behind using ML is that it provides an alternative approach to sky classification (Aghimien et al., 2022; Granados-López et al. 2021; Lou et al., 2016). Nevertheless, recent works have shifted focus to the use of ensemble ML models as these can provide better predictions, they can boost the performance of traditional ML and perform well even under noisy data (Mienye et al., 2022). Importantly, ensemble models perform well when the data to be predicted is imbalanced which is the case for most real-life classification problems (Khan et al., 2024; Mohammed & Kora, 2023). Nonetheless, ensembles have not been widely explored in previous CIE standard skies studies. Moreover, the previous ML works in CIE standard sky classification did not attempt to determine the energy performance of these methods in actual energy simulation scenarios (Granados-López et al., 2021; Lou et al., 2017). Neither is the comparison of the ML-based CIE standard Skies against other sky distribution models considered in the previous energy analysis. Thus, in this study, the daylight illuminance and building energy prediction from the proposed LGBM model and for a generic commercial building infused with DLLC systems were investigated. The objectives of this study were to (i) develop an LGBM CIE standard sky classification model (ii) simulate daylight illuminance based on the LGBM and other acceptable sky distribution models (iii) determine energy savings from using DLLC systems and the sky models.

2. Study Methodology

Ten-minute measurements obtained from the City University of Hong Kong measuring station between 2004 to 2005 were used in this study. These data consist of solar irradiance, daylight illuminance, sky luminance distributions, geographical and meteorological variables. The measurements were done using Kipp & Zonen CM11 thermopile pyranometers, the MS- 80 Pyranometer, the T-10M illuminance sensor and the EKO MS 300LR sky scanner. Similarly, the meteorological variables such as sunshine duration (*sun*), cloud cover (*cld*), visibility (*vis*) e.t.c were collected from the Hong Kong Observatory (HKO). These meteorological inputs also covered the same measurement period as the solar measurements.

Upon measurement, data quality control was carried out to clean the data as described in Aghimien et al. (2022). Upon cleaning, a total of 16,118 and 10,747 datasets were obtained for years 2004 and 2005, respectively. The 2004 data was further divided using the ratio of 80 to 20% for training and initial testing. This splitting ratio provides sufficient data for learning and still gives room for evaluating the model against unseen data. Next, the 15 CIE standard skies were determined as outlined in Li et al. (2013). Therefrom, the sky classification model was developed using LGBM and the 2004 input climatic data. Upon model development, the LGBM was used to classify the 15 standard skies using the 2005 test data. By using the model against the 2005 data, the ability of the model to make predictions against a whole year's worth of data was determined. For daylighting and energy calculations, a generic shopping mall was assumed and, the standard skies obtained from the LGBM were used to simulate daylight illuminance in RADIANCE using a climate-based daylight modelling (CBDM) approach. For a more robust conclusion, the LGBM classification was compared with the Best fit skies (i.e., the 15 CIE standard skies as classified by the sky luminance modelling method), the All-weather model (Perez et al., 1993), and the ASRC-1992 (Perez et al., 1992). Finally, the energy savings as obtained from the different sky models, the top-up DLLC system and other building parameters were investigated using EnergyPlus. Importantly, for the daylight and the building energy simulation, the 2005 weather data was used. The methodology flow chart is shown in Fig. 1.

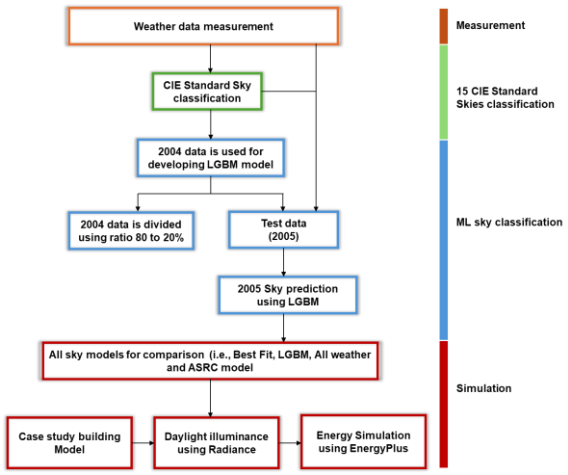

Fig. 1 – Flow chart of research methodology

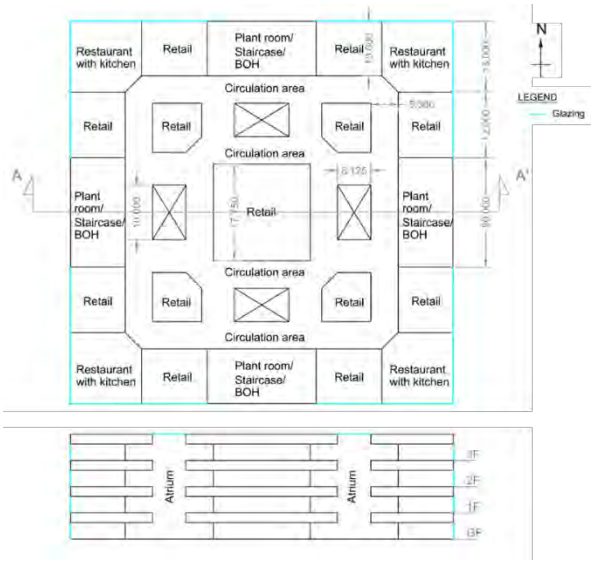

Fig. 2 – Floor plan and section of case study building

3. Case Study and Model Description

3.1 Case Study

The generic case study is a 70 m by 70 m 4-storey shopping mall in Hong Kong. The mall has typical floor plans comprising retail shops, restaurants, back-of-house, circulation areas, supporting plant rooms, and atria as shown in Fig. 2. In terms of glazing, the overall skylight-to-roof and window-to-wall ratios were 5% and 44.6%, respectivelyand the visible transmittances for the skylight and window were 0.9 and 0.8. The lighting power densities for the retail shops, restaurants and circulation areas were 10, 9 and 5 W/m², respectively. Similarly, the design illuminance for the retail shops, restaurants, and circulation areas was 300 lux, 200 lux and 100 lux, respectively. A fan-coil unit was provided for retail shops and restaurants while a variable-air-volume system was used in circulation areas. Water-cooled chillers were used for space cooling. The operating hours for the retail shops and circulation areas were 09:00-21:00 while the restaurants were 09:00-01:00. All parameters were chosen following the Hong Kong local requirements (EMSD, 2021). DLLC systems were equipped in the perimeter zones, and these were twice the window height while the reference point for lighting control was placed in the middle of the room. Finally, obstructions of 56.3° were positioned in all four major principal orientations (i.e., north, east, south and west). The atria, ground and third-floor plans were considered in this analysis.

3.2 CIE Standard Skies Classification

The CIE standard skies contain five clear, five intermediate and five overcast skies, and these cover the whole spectrum of skies in the world (CIE, 2003). These 15 CIE skies are derived from a number of mathematical expressions which are mainly composed of the relative distribution, the standard gradation function and the relative scattering indicatrix function (Li et al., 2013).

Full details of the 15 CIE standard skies and its modelling approach using the sky luminance method can be found in Aghimien and Li (2022). This method was used in this study (Section 4.1) as the baseline method for comparing the LGBM sky classification performance.

3.3 Light Gradient Boosting Machine

Ensemble models combine several ML models to build a single and more powerful model than its original constituents (Mohammed & Kora, 2023). These models are widely adopted due to their ability to reduce overfitting and efficiency when dealing with imbalance data (Khan et al., 2024). The LGBM is a highly efficient ensemble of decision trees (DT) used to minimize a loss function (Ooba et al., 2023). This model has been widely adopted and hence, used in this study.

Before model development, the relationship of the inputs was first checked using Pearson's correlation analysis. This helps to determine the model input

relationship and prevent multicollinearity. The findings show no likelihood of collinearity in the data. For the model structure, thirteen input variables were used and these consist of solar altitude angle (α), clearness index (K_t), diffuse fraction (K_d), turbidity (T_v), atmospheric pressure (*atp*), cloud cover (*cld*), sunshine duration (*sun*), visibility (*vis*), relative humidity (*rhm*), dew point temperature (*dpt*), dry bulb temperature (*dbt*), wet bulb temperature (*wbt*) and wind speed (*wsp*). While the output was the 15 CIE standard skies identified by the sky luminance method. Furthermore, the output data were encoded since they have categorical attributes, and then the data was split. The 2004 data was used for training and initial testing in the ratio of 80 to 20%. Next, the split data were separately scaled using the min-max method to prevent data leakage. Then, the Grid search method was used to optimize the model while K-fold cross-validation was used to prevent overfitting. Upon optimization, the best LGBM hyperparameters were; column subsampling by the tree: 0.9, learning rate: 0.01, number of estimators: 300, number of leaves: 60, and subsample: 0.8. After model development, the 2005 data were scaled and used as additional new sets of data to retest the model's performance. Finally, the model was assessed using accuracy (*Accu*), precision (*Pre*), recall (*Re*), F1-score (*f1*) and receiver operating characteristic (ROC). By using these arrays of metrics and setting their average as "weighted", issues related to data imbalance were catered for. Details of these evaluation metrics can be found in Hossin and Sulaiman (2015).

4. Results and Discussions

4.1 CIE Skies Luminance Classification

As shown in the frequency of occurrence (FOC) plot in Fig. 3, the overcast skies (Skies 1 to 5) represented 32.1 % of the sky condition. Partly cloudy skies (Skies 6 to 10) and clear skies (Skies 11 to 15) represented 44.3% and 23.5%, respectively. Skies number 1, 8 and 13 were the most represented sky types with FOCs of 17.5, 37.0% and 16.0%. These also represented the most represented sky types for each typical sky. Overall, the FOC result showed that the data (i.e., predicted sky types) was imbalanced as expected. Thus, making the use of the LGBM model a good alternative.

4.2 LGBM Standard Skies Classification

The statistical performance of LGBM on the 2004 test data was assessed. The ROC for the identified skies and the micro average value were presented. As shown in Fig. 4, the area under curve (AUC) of the identified skies ranged from 0.87 (i.e., Sky 5) to 0.99 (i.e., Skies 13 to 15). Similarly, the micro average ROC had an AUC of 0.98. A perfect classifier will usually have an AUC of 1.00. Hence, the findings show that the ROCs were quite close to a perfect classifier. This result means that the LGBM model can classify the 15 CIE standard skies with reasonably good accuracy and there is a tendency to obtain high recall and low false positive rate across the skies during classification.

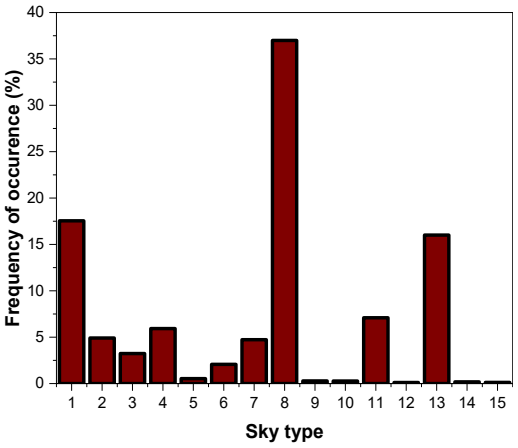

Fig. 3 – FOC of the best fit 15 CIE Standard Skies

Next, the LGBM performance was evaluated by using the confusion matrix in Fig. 5. The lighter brown box in the matrix shows instances where predictions were accurate. As expected, Skies 1, 8 and 13 were the most correctly classified skies and these had correct predictions of 482, 1098 and 2310, respectively, which is generally in line with the observed FOC of the skies (i.e., Section 4.1).

Lastly, the model's performance on the initial and new sets of test data (i.e., 2004 and 2005, respectively) is presented in Fig. 6. It was observed that for the 2004 data, the *Accu* was 74.7% while the *Pre Re* and *f1* were also above 70%. As expected, the

performance dropped for the new 2005 test data. Nonetheless, its *Accu* and *Re* were above 70% while the *Pre* and *f1* were 66.91 and 67.77%, respectively. Overall, the result shows that for most of the predictions, the LGBM correctly identified the 15 CIE standard skies for more than 70% of the instances. This implies that good predictions were obtained.

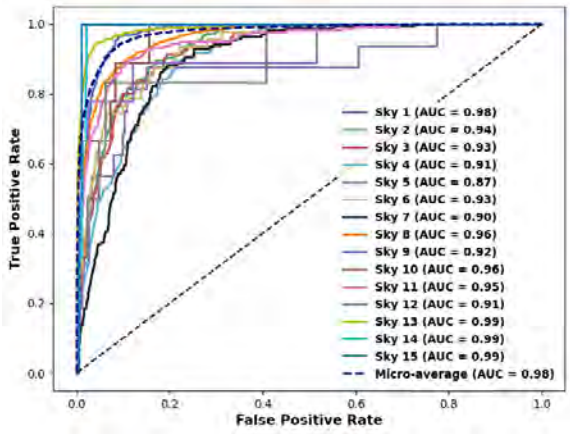

Fig. 4 – ROC Curves of all 15 standard skies

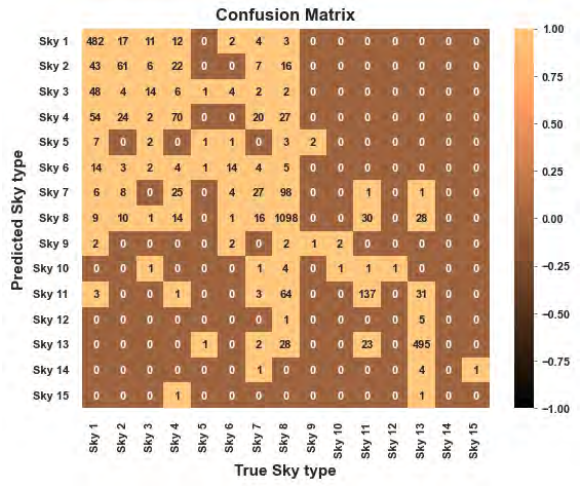

Fig. 5 – Confusion matrix of LGBM model on 2004 test data

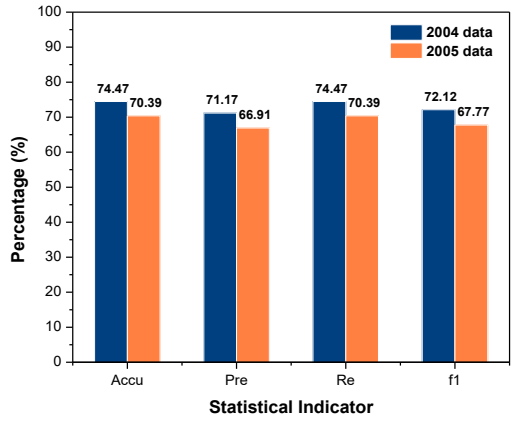

Fig. 6 – Performance of LGBM on 2004 and 2005 test data

4.2.1 LGBM feature importance

The feature importance of the LGBM model was assessed using the permutation importance method. As shown in Fig. 7, K_d, T_v and K_t were the most important inputs. Next to these, were α, *cld* and *sun*. Since these variables have been extensively used as sky clearness indicators, this result validates their high level of relevance. However, other meteorological variables had lower importance. From this analysis, the important features can be subsequently used in places with limited data for developing simpler sky models.

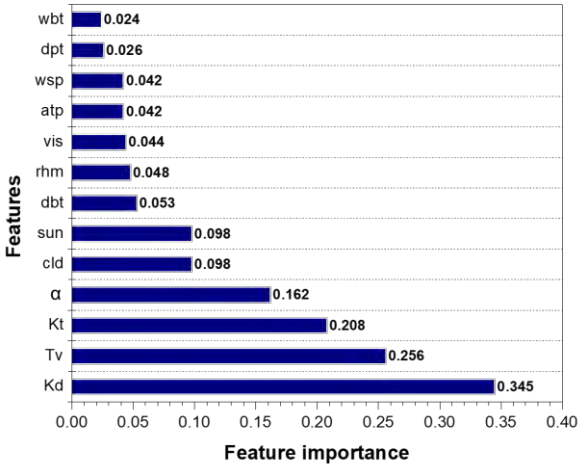

Fig. 7 – Feature importance of LGBM inputs

4.3 Daylight Illuminance Simulation

For daylight illuminance simulation, the Best fit standard skies, the LGBM-based standard skies and other luminance distribution models were integrated with RADIANCE and the analysis was conducted as described in Section 2.0. Precisely, five scenarios were considered for discussion. These scenarios include the south atria, then the north and south orientations of the ground and third floor, respectively. By selecting these scenarios, extreme and less extreme cases were covered. For example, the ground and third floors, depict daylight predictions with more and less obstruction, respectivelywhereas north and south orientations show the effect on sun-shaded and less shaded surfaces, respectively. The prediction error was evaluated using %root mean square error (*%RMSE*) and %mean bias error (*%MBE*). The *%RMSE* compares forecasting errors of different models, while the *%MBE*

determines the tendency of a model to overestimate or underestimate. Details of these evaluation metrics can be found in Despotovic et al. (2015). As observed in Table 1, all models tend to provide better predictions (i.e., lower %*RMSE*) on the ground floor compared to the third and on the north orientation compared to the other orientations. Since the model was proposed with consideration of obstruction, the reason for this might be the likelihood of the south orientation, third floor and atria being exposed to more of the sky. Overall, the models gave predictions of reasonable accuracy and this was more obvious in the Best fit and LGBM models. In fact, the Best fit prediction ranged from 10.17 to 23.73%, while the LGBM ranged from 11.5 to 26.17%. Strictly speaking, models with %*RMS*E < 20% are considered to have higher accuracy (Despotovic et al., 2015). Hence, showing the efficacy of both models. Nevertheless, the All-weather and ASRC-1992 also gave acceptable predictions with the latter performing the least. Also, the observed %*MBE* indicated that most of the time, the predictions were not so far from the daylight illuminance values. Furthermore, the ASRC-1992 and All-weather models mainly overestimated the daylight illuminance, while the Best fit and LGBM mostly underestimated. Importantly, for the proposed LGBM the %*MBE* ranged from -0.03 to -3.35%. This implies that most of the time, the average difference between the measured illuminance and the predicted value from the proposed LGBM will be around 3%.

Table 1 – Statistical performance of models in predicting daylight illuminance using RADIANCE simulation

Scenarios	Metrics	Best fit	ASRC-1992	LGBM	All-weather
Atrium (South)	*%RMSE*	23.73	33.83	26.17	28.57
	%MBE	2.08	-5.43	2.17	-1.57
GF (North)	*%RMSE*	10.17	25.67	11.5	16.93
	%MBE	0.13	8.75	-0.03	4.45
GF (South)	*%RMSE*	19.53	34.63	22.95	27.03
	%MBE	-1.38	8.88	-2.06	3.23
3F (North)	*%RMSE*	11.84	25.56	13.44	16.99
	%MBE	-0.17	8.77	-0.28	3.76
3F (South)	*%RMSE*	20.77	39.75	24.87	28.32
	%MBE	-2.49	11.04	-3.35	2.99

Note: GF represents the ground floor while 3F represents the third floor.

4.4 Energy Performance Simulation

Although indoor lamps help improve visual comfort, they dissipate heat, which will impact the indoor cooling requirement. Similarly, by setting the DLLC to the target indoor illuminance, significant energy savings can be achieved. Based on the predicted illuminance, the building energy consumption is estimated by EnergyPlus. The energy simulation result in terms of the energy savings of the measured sky and different models are presented in Table 2. As pointed out in Section 2, there were only 10,747 sets of valid data, which is equivalent to about 1,791 hours and about half of the daytime for the simulation period. Thus, it should be noted that the findings in Table 2 only cover half of the year's daylight conditions. As observed, the addition of DLLC systems caused energy savings on lighting for the measured sky luminance data and the sky models. This energy saving on lighting ranged from 38.1 (i.e., measured sky luminance data) to 38.6% (i.e., ASRC 1992). Similarly, for cooling-related end uses like fans, and heat rejection there is also a considerable energy savings actualized from the use of the DLLC system. This had maximum values of 6.0 and 4.0%, for fans and heat rejection, respectively. Other savings of about 2.5 % were derived from pumps while end use without a direct relationship with DLLC systems such as equipment and heating had no energy savings. Furthermore, it was observed that the energy savings for the different models was not so far from the measured data and all savings from these models were of similar magnitude. The reason for this might be because the lux level and visible window transmittance used in the analysis were low. Moreover, the presence of obstruction might be of concern since the analysed spaces were mainly dependent on the diffuse and reflected illuminance. Generally, shopping centers have long operating hours and relatively low illuminance requirements, hence, larger savings may be obtained if it is applied to office buildings. Nevertheless, the closeness of the predicted energy savings from the sky models to that of the measured sky luminance data shows that acceptable predictions were obtained.

5. Conclusion

This paper shows the potential of using the LGBM model and accessible climatic variables to determine the 15 CIE standard skies. The findings show that for over 70% of the time, the LGBM model could correctly classify the sky types. Hence, the proposed model could provide acceptable predictions. The important features in the LGBM model were determined. It was shown that sky clearness indicators like K_d, T_v K_t, α, *cld* and *sun*, were the most important features. Therefrom, the LGBM model alongside the Best fit, ASRC-1992 and all-weather models were used for daylight and energy simulations of a generic shopping mall. For daylight simulation, it was observed that surfaces more exposed to the skies like the south and upper floors (e.g. third floor in this study) are more prone to error during daylight predictions. In terms of %*RMSE*, the Best fit model gave the best predictions while the ASRC-1992 performed the least. It was also observed that from the %*MBE* obtained, the difference between the measured and predicted illuminance from the proposed LGBM will be around 3% on average. Finally, electricity consumption was predicted and findings show that all models gave predictions which were close to the measured data. Most of the savings only deviated within 2 MWh which is equivalent to about 1.5% of the saving. Generally, an approach for determining the skies which can be incorporated into simulation software has been proposed. Nevertheless, more work using other ensemble models, larger databases and different locations is required.

Table 2 – Energy savings for all four sky models with Top-up control

End Uses	Sky luminance	Best Fit	LGBM	All-weather	ASRC 1992
Lighting	38.1	38.2	38.2	38.5	38.6
Equipment	0.0	0.0	0.0	0.0	0.0
Fans	5.9	5.9	5.9	5.9	6.0
Heating	-3.9	-3.9	-3.9	-3.9	-4.0
Cooling	0.0	0.0	0.0	0.0	0.0
Heat Rejection	3.9	4.0	4.0	4.0	4.0
Pumps	2.4	2.4	2.4	2.5	2.4

Note: Energy savings are expressed as percentages (%)

Acknowledgement

Work described was supported by the Faculty Development Scheme from the Research Grant Council of HKSAR [Project no. UGC/FDS16/E03/20] and Research Matching Grant Scheme from the Research Grant Council of HKSAR [Ref. no. 2021/3006].

References

Aghimien, E. I., D. H. W. Li, and E.K.W. Tsang 2021. "Bioclimatic architecture and its energy saving potentials: a review and future directions". *Engineering, Construction and Architectural Management* 29(2): 961–988. https://doi.org/10.1108/ECAM-11-2020-0928

Aghimien, E. I., and D.H.W. Li. 2022. "Application of Luminous Efficacies for Daylight Illuminance Data Generation in Subtropical Hong Kong." *Smart and Sustainable Built Environment* 11 (2): 271–93. https://doi.org/10.1108/SASBE-08-2021-0146

Bellia, L., F. Fragliasso, and E. Stefanizzi. 2016. "Why Are Daylight-Linked Controls (DLCs) Not so Spread? A Literature Review." *Building and Environment* 106 (September): 301–12. https://doi.org/10.1016/j.buildenv.2016.06.040

CIE S 011/E. 2003. "CIE 2003 Spatial Distribution of Daylight – CIE Standard General Sky", *Standard, C.I.E. central bureau: Vienna* 1 – 7.

Despotovic, M., V. Nedic, D. Despotovic, and S. Cvetanovic. 2015. "Review and statistical analysis of different global solar radiation sunshine models". *Renewable and Sustainable Energy Reviews* 52: 1869–1880. https://doi.org/10.1016/j.rser.2015.08.035

Electrical and Mechanical Services Department (EMSD). 2023. "Hong Kong Energy End-Use Data 2023." Hong Kong SAR Government.

Electrical and Mechanical Services Department (EMSD). 2021. "Code of Practice for Energy Efficiency of Building Services Installation." Hong Kong SAR Government.

Granados-López, D., A. Suárez-García, M. Díez-Mediavilla, and C. Alonso-Tristán. 2021. "Feature selection for CIE standard sky classification". *Solar Energy* 218: 95–107. https://doi.org/10.1016/j.solener.2021.02.039

Hossin, M. and M.N. Sulaiman. 2015. "A review on evaluation metrics for data classification

evaluations". *International Journal of Data Mining & Knowledge Management Process* 5:2. https://doi.org/10.5121/IJDKP.2015.5201

Khan, A., O. Chaudhari, and R. Chandra. 2024. "A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced Problems: Combination, Implementation And Evaluation". *Expert Systems With Applications* 244:122778.

Li, D.H.W., N.T.C. Chau., and K.K.W. Wan. 2013. Predicting Daylight Illuminance and Solar Irradiance on Vertical Surfaces Based on Classified Standard Skies. *Energy* 53: 252–258, 2013. https://doi.org/10.1016/j.energy.2013.02.049.

Lou, S., D.H.W. Li, and J.C. Lam. 2017. CIE Standard Sky classification by accessible climatic indices. *Renewable Energy* 113: 347–356. https://doi.org/10.1016/j.renene.2017.06.013

Mienye, I. D., Y. Sun, and Z. Wang. 2019. Prediction Performance of Improved Decision Tree-based Algorithms: A Review. *Procedia Manufacturing* 35: 698–703. https://doi.org/10.1016/j.promfg.2019.06.011

Mohammed, A., and R. Kora. 2023. A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges. *Journal of King Saud University –Computer and Information Sciences* 35: 757- 774. https://doi.org/10.1016/j.jksuci.2023.01.014

Ooba, H., J. Maki, T. Tabuchi, and H. Masuyama, . 2023. Partner Relationships, Hopelessness, and Health Status Strongly Predict Maternal Well-Being: An Approach using Light Gradient Boosting Machine. *Science Reports* 13: 17032. https://doi.org/10.1038/s41598-023-44410-1

Perez, R., J. Michalshy, and R. Seals. 1992. "Modeling Sky Luminance Angular Distribution for Real Sky Conditions: Experimental Evaluation of Existing Algorithms". *Journal of the Illuminating Engineering Society* 21(2): 84–92. https://doi.org/10.1080/00994480.1992.10748005

Perez, R., R. Seals, and J. Michalsky. 1993." All-weather model for sky luminance distribution—Preliminary configuration and validation. " *Solar Energy* 50(3): 235–245.