Supervised Machine Learning Methods for the Prediction of Discomfort due to Glare

Marshal Maskarenj PhD[a]* and Sergio Altomonte PhD[a]

[a]*Architecture et Climat, Louvain Research Institute for Landscape, Architecture, Built Environment, Université catholique de Louvain, Louvain-la-Neuve, Belgium*

*maskarenj.marshal@uclouvain.be

Batiment Vinci a.017, L5.05.04,

Place du Levant 1, 1348 Louvain-la-Neuve, BELGIUM

Acknowledgements:

The authors acknowledge Dr Michael Kent for responding to queries on setup design.

Supervised Machine Learning Methods for the Prediction of Discomfort due to Glare

Supervised Machine Learning (ML) is emerging as a powerful tool for handling complex multivariate datasets and capturing non-linear relationships between predictors and outcomes. Unlike traditional statistical techniques that only enable to identify predefined functional relationships, supervised ML methods can robustly detect intricate non-linear interactions without explicit assumptions. The capacity for integrating diverse predictors makes these tools particularly suitable for challenges where traditional modelling show limitations. Prediction of discomfort due to glare is one such case, where current models often fail to explain variance in data. These models mostly focus on objective factors related to the scene, while neglecting subjective factors that can vary between- and within-users. These limitations contribute to significant residual variance between model-predicted and reported discomfort scores, which cannot be explained without accounting for inter- and intra-individual differences. To improve these predictions, this paper proposes a method based on supervised machine learning techniques where multiple predictors and their non-linear relationships are integrated towards improving model accuracy and robustness. Via preprocessing, features of a dataset collected from a lighting experiment were refined, and prediction models were developed using eight state-of-the-art ML algorithms. Through inverse regression, the best performing model was further deployed for optimising the characteristics of delivered lighting, ensuring that user perception was maintained within a target zone when all other variables were constrained. Our method demonstrates a practical approach for data handling, multivariate non-linear modelling, and user-centred optimization of delivered lighting, with potential application for problems beyond prediction of discomfort due to glare.

Keywords: supervised machine learning, discomfort glare, subjective factors, inverse regression, user-centric lighting

1. Introduction

Machine learning (ML) techniques have proven more efficient for treating large amounts of data, and solving complex multidimensional problems, compared to traditional methods for performing statistical analysis [1, 2]. ML algorithms are now enabling systems to learn from data, and to improve predictions without being explicitly programmed[3]. These algorithms can broadly be classified under four main categories: *Supervised* ML, *Semi-supervised* ML, *Unsupervised* ML, and *Reinforcement* ML [4]. Supervised ML algorithms learn to make predictions from labelled data, and are typically employed for regression and classification problems – where regression refers to predicting continuous outcomes, while classification includes categorising data into discrete classes [5]. Unsupervised ML algorithms do not need labelled data, and typically address problems such as clustering, association, or pattern detection [6, 7]. Reinforcement ML is yet another technique, where an agent learns to make decisions by interacting with an environment, and receives feedback either as rewards or as penalties [8]. This technique is used for designing strategies that maximize the cumulative reward over time [9, 10]. Each of these ML techniques offer specific advantages for data analysis, and the technique selection depends both on the application as well as on the data structure, i.e. whether the training data is labelled (supervised), unlabelled (unsupervised), or a mix of both (semi-supervised) [11]. Within supervised ML, predictors and outcomes are treated as input-output pairs, and the dataset is split into a subset used for training, and another for testing and internal validation. During the training process, the algorithm iteratively adapts itself to minimize errors in its prediction, and the model built on the dataset's training split can be used for making predictions on the test split, as well as on new datasets having the same feature configuration. Applications of supervised ML within the sciences of the built environment have been reported in the literature to address various challenges, such as for energy use modelling [6], for predicting human factors of subjective comfort [12, 13] and

satisfaction [14], and for controlling the transparency of dynamic glazing, which is a classification problem of supervised ML [15].

In lighting research, many challenges have not yet been fully resolved, among which is the complete characterisation and prediction of discomfort due to glare. Discomfort glare refers to the psychological sensation which occurs when bright light in our field-of-view causes annoyance or disturbance, without necessarily impairing our visual ability [16]. Several models have been proposed to describe subjective perception of discomfort due to glare [17-22]. In this context, Carlucci et al. [23] comprehensively reviewed 34 indices of visual comfort, subdividing them into those associated with *(i)* the amount of light, *(ii)* the quality of light, *(iii)* the distribution of light, and *(iv)* those associated with discomfort glare. It is evident from this review that the last category has expanded over the years. One potential reason justifying this increase, despite an existing abundance of prediction models, is that discomfort glare is still not fully understood, and there are large variances between model-predicted and subjective responses [24]. Although the sources of this variance cannot be entirely explained [25], a large proportion has been attributed to factors related either to the personal characteristics of the users [26], to the underlying contexts from where data were collected [27], or to the data collection procedures themselves [28]. Additionally, factors related to the lighting scene must also be considered as additional sources of uncertainty in field studies (such as those involving daylight), particularly in cases where the ranges of exposure cannot be tightly controlled.

Existing discomfort glare models only account for objective factors related to the scene, uniquely performing correlations that are either linear, quadratic, or exponential. Since the outcome can simultaneously be influenced by multiple predictors having a subjective nature, these limitations might disregard any underlying relationship between inter/intra-individual and/or contextual factors, significantly reducing the model's capacity to explain the variance in responses. Considering the limitations in traditional statistical techniques used by current

models, it is necessary to explore more robust methods for analysing data generated from lighting experiments, and for modelling discomfort due to glare.

For addressing this challenge, the data-driven approach of supervised ML can prove particularly effective, and can significantly improve predictions especially in cases where subjective or contextual factors dominate the variance in a dataset. The ability of supervised ML for performing multivariate non-linear regressions, while incorporating multiple predictors, offers significant potential for improving the accuracy and robustness of discomfort glare prediction models. Towards this aim, this paper wishes to demonstrate the use of data-driven supervised machine learning tools for analysing a multi-feature dataset from a lighting experiment, and for training a robust ML-based model for predicting discomfort glare. The various steps taken for treating and analysing the data, as well as for training and deploying the model for user-centred lighting application, can guide further investigations as well as encourage a wider application of this protocol, for example in the domain of outdoor lighting.

Compared to prior studies that have applied machine learning to visual discomfort, glare prediction, or similar applications, our study introduces various new contributions. Prior models, such as those by Nourkojouri et al. [29] and Li et al. [30], have relied on simulated lighting environments and objective glare indices, without incorporating human responses related to subjective perception (or discomfort evaluation). Other real-world studies, including Tzouvaras et al. [31] and Wagdy et al. [32], have either not included subjective ratings, or have reduced user feedback to binary responses – hence limiting their granularity and capacity for delivering personalized lighting recommendations. One of the novelties of our study consists in training models based on psychophysical data collected from human participants using continuous visual analog scales. The inclusion of multiple contextual and personal predictors – such as luminous history, chronotype, fatigue, mood, caffeine intake, time of day, etc.– allows us to model both inter- and intra-individual differences in discomfort glare perception – an

aspect that has not yet been consistently addressed in the literature. Additionally, we compare eight supervised ML regressors within a pipeline, together with 5-fold cross-validated performance evaluation, analysis of stability in feature importance, and an optimization routine that allows our model optimization framework to recommend personal lighting conditions for specific user personas and profiles. These elements collectively aim to move away from generic prediction of discomfort due to glare, toward a modelling approach that is grounded in psychophysical data, and is specific to the user and its context.

2. Methods

The following sections provide: *(i)* a brief description of the procedure used for the collection of the data used in this paper; *(ii)* an illustration of the variables featured in the dataset; and, *(iii)* a presentation of the process used for data treatment and analysis. A more detailed description of the testing apparatus and of the experimental procedure is part of a further ongoing publication by the authors[33, 34].

2.1 Data collection procedure

The subjective data used in this paper was collected from a lighting experiment conducted in Belgium during the spring months of April to June, which aimed to characterize the effect of contextual factors on the evaluation of discomfort due to glare. These factors included time of day and the spectrum characteristics of the glare source.

A chamber was built to generate pre-set electric lighting scenes that were designed to correspond to pre-calibrated thresholds of discomfort due to glare. A total of 38 participants took part in this controlled experiment. The participants reported their subjective discomfort sensation on two visual analogue scales (VAS, provided in the Appendix A1) over three

different times of day when being exposed to the glare source. Participants maintained a fixed head position defined by a chin- and forehead-rest [35], and faced a fixation point in their field of view while being exposed to the lighting scenes. Appendix A2 presents the experimental apparatus within the test chamber, and provides a detailed timeline of the experiment

2.2 Variables

2.2.1 Variables related to the scene

The scenes were calibrated using the IES-GI model [21] to stimulate a visual discomfort anchored at the 'just perceptible' (JP), 'just noticeable' (JN), 'just uncomfortable' (JU), and 'just intolerable' (JI) glare sensation vote (GSV) [16]. The four GSV-anchored scenes were designed with two spectral states (herein, for simplicity, generally referred to as their correlated colour temperature, CCT, that is 'red' and 'blue'), and the eight scenes were maintained using commercially available Wiz A67 E27 smart lamps. This study used a category rating procedure [36] where the participants marked their perception on visual analogue scales against every predefined state of lighting. Scenes were toggled between these predefined states using the Wiz application. The characteristics of these eight lighting scenes are presented in Table 1.

[TABLE 1]

For each scene, the Correlated Colour Temperature (Tcp), vertical illuminance (Vert-illum), and the CIE chromaticity coordinates presented in Table 1 were measured using a Konica Minolta CL-70F spectrometer, where the device's bulb was positioned to approximate the eye's position when the head is supported by the chin- and forehead-rest. The 1-nm separated spectral irradiances (between 380-780 nm range) measured by the spectrometer were then used to calculate the α-opic EDIs and the circadian metrics for each scene, respectively using the CIE-s026 [37, 38] and LRC's calculators [39].

Table 1. Photometric and radiometric measurements for the eight scenes measured at eye level using a Konika Minolta CL70F spectrometer (1nm resolution between 380-780nm). The CIE S026 a-opic metrics are derived from spectral data using the CIEs026 calculator, and the circadian metrics are derived using the LRC calculator.

					CIE 1931		CIE 1976		EDI based on CIE s-026 calculator						
CCT indicator	**GSV**	**Lum (cd/m^2)**	**Tcp (K)**	**Vert Illum (lx)**	**x-coord**	**y-coord**	**u-coord**	**v-coord**	**s-cone (lx, EDI_S)**	**m-cone (lx, EDI_M)**	**l-cone (lx, EDI_L)**	**Rods (lx, EDI_R)**	**melan. (lx, EDI_{mel})**	**Circad Lighting (lx, CL_A)**	**Circad Stim.**
Red	JP	1627	2645	336	0.4691	0.4116	0.2680	0.5292	85.57	262.93	342.50	174.66	145.86	291.33	0.3116
	JN	3860	2638	345	0.4693	0.4110	0.2684	0.5289	88.51	269.94	352.09	179.02	149.27	297.94	0.3159
	JU	9150	2620	394	0.4708	0.4112	0.2693	0.5292	99.62	306.96	401.71	201.87	167.12	333.64	0.3376
	JI	21700	2609	485	0.4711	0.4104	0.2699	0.5290	123.50	376.89	494.30	245.31	200.97	401.23	0.3732
		Mean	**2628**	**390**	**0.4701**	**0.4111**	**0.269**	**0.529**	**99.3**	**304.2**	**397.6**	**200.2**	**165.8**	**331**	**0.335**
		StDev	**16.5**	**68.26**	**0.001**	**0.0005**	**9E-04**	**1E-04**	**17.23**	**52.18**	**69.46**	**32.35**	**25.23**	**50.35**	**0.028**
Blue	JP	1627	6304	336	0.3159	0.3545	0.1908	0.4818	290.57	340.84	333.97	347.79	348.18	469.75	0.4032
	JN	3860	6271	343	0.3166	0.3551	0.1910	0.4822	295.25	348.13	341.47	355.41	356.16	479.01	0.4069
	JU	9150	6210	383	0.3179	0.3558	0.1916	0.4827	326.68	387.90	381.37	395.50	396.40	533.04	0.4268
	JI	21700	6069	483	0.3209	0.3588	0.1926	0.4846	400.46	487.00	480.85	492.20	491.43	655.22	0.4636
		Mean	**6214**	**386.2**	**0.3178**	**0.3561**	**0.192**	**0.483**	**328.2**	**391**	**384.4**	**397.7**	**398**	**534.3**	**0.425**
		StDev	**104**	**67.74**	**0.0022**	**0.0019**	**8E-04**	**0.001**	**50.75**	**67.28**	**67.57**	**66.37**	**65.74**	**85.34**	**0.028**

2.2.2 Variables related to the observer (Personal Factors)

This study collected 13 variables associated with personal factors [28] . For each participant, questionnaire items asking data on chronotype [40], ethnicity, eye conditions, etc. were administered during the introduction sessions. The questionnaire used to evaluate personal factors is provided in the Appendix A3, and descriptive statistics of personal factors are presented in Table 2.

[TABLE 2]

Table 2. Descriptive statistics on data related to personal factors (N = 38 subjects)

Gender	**Age**	**Ethnicity**	**Eye conditions**	**Chronotype**	**Photosensitivity**
52.6% *Male* 47.3% *Female*	Range: 22-64 Med: 32 34.4 ± 11.6	***White***: 73.6% ***Asian***: 21% ***Mixed***: 5.3%	42% wore lenses 14% reported other vision issues	7-point scale, 2.95 ± 1.4 Median: 3	**Natural Light**: 68.4% **Artificial Light**: 60.6% **Sunglasses**: 39.4% **Blind preference**: 1.15±0.85 on the 0-4 scale **Lighting preferred**: 0.73±0.75 on a 0-3 scale
Caffeine Habit					
Yes: 67.5%					

It must be noted that photosensitivity was collected as five separate items, while eye-conditions included three separate items. Among the 38 participants, five indicated "Yes" to the item "any other vision problem", which was one of the items connected to their eye-conditions. Their reported descriptions in the follow-up included: previous use of low corrective lenses (n = 1), mild short-sightedness and presbyopia (n = 1), strabismus (n = 1), and two unspecified responses. None of the participants reported any diagnosis of conditions associated with ocular pathology – such as cataract, glaucoma, or corneal disorder, that could have significantly influenced their perception of discomfort due to glare.

2.2.3 Variables related to the context (Contextual Factors)

Prior to the scene exposure within each session, participants were asked to respond to 14 items

associated with contextual factors[28]. These are the factors that can vary within the participant across sessions, and could lead to intra-individual differences in responses to perceived discomfort. These factors were collected through established scales, including fatigue [41], sleepiness [42], mood indicators [43], etc. Two instruments were employed to collect sleepiness – the Karolinska Sleepiness Scale (KSS) and the "sleepy/awake" items within the mood indicators, both of which were administered after the scene evaluations had been completed. The KSS asked participants to circle the number that best described their current perceived state of sleepiness on a 9-point scale, hence asking test subjects to report their self-evaluated level of physiological sleepiness/alertness. Conversely, the "sleepy" and "awake" mood items, collected on 5-point scales together with other dichotomous affective descriptors (including, tense/calm, sad/happy, lively/drowsy), were included to deepen the appraisal of the participant's collective emotional states. Their simultaneous administration aimed at furthering the characterisation of the perceived level of sleepiness reported by test subjects. Descriptive statistics at session level are presented in Table 3, and the questionnaire items used for contextual data collection at each session are provided in the Appendix A4.

[TABLE 3]

Table 3. Descriptive statistics on data related to contextual factors (N=114 sessions)

Caffeine score	**Prior Light Exposure (0-7)**	**Mood Indicators (1-5)**	
21.1 ± 25.5%		***Tense***	1.91 ± 0.99
Continuous VAS	***Direct sunlight:***	***Sad***	1.53 ± 0.8
Fatigue (0-6)	5.1±1.8	***Lively***	3.37 ±0.82
2.14±1.29	***Diffuse daylight***	***Sleepy***	2.24±0.96
Samn-Perelli scale	3.9±1.6	***Calm***	3.7±1.02
Sleepiness (1-9)	***Electric light***	***Happy***	3.6±0.84
4.2±1.7	4.3±1.8	***Drowsy***	2.4±1.05
Karolinska scale		***Awake***	3.53±0.96

2.2.4 Variables related to the experimental session

The dataset also included the session order, i.e. six permutations of Morning 'M', Afternoon

'A', Evening 'E'. These defined the participation sequence for every individual, which was counterbalanced at the participant level. Also included was the order of exposure for the eight defined lighting scenes during each session. The exposure order was counterbalanced at the session level.

2.2.5 Outcome variables

During each exposure within every session, participants looked at the fixation point for the entire duration of the exposure, after which they reported the level of discomfort that they felt due to glare from the lighting scene. The responses were collected on paper-based continuous visual analogue scales (VAS), where the participants placed a cross mark corresponding to their sensation. The responses were digitized by measuring the distance of the cross mark from the left end of the VAS. The VAS used for this study have been presented in the appendix A1.

2.2.6 Potential experimental biases

Experimental biases were minimized by following recommendations for good practice, such as counterbalancing session order (i.e., time of day for the participants), counterbalancing the order of displaying the eight scenes within each session (JP/Red, JP/Blue, … JI/Blue), blindfolding participants between scene changes, etc. [44, 45].

2.3 Outline of data treatment

Three distinct steps were taken for treating the data, including feature optimization (Figure 1a), supervised Machine Learning (Figure 1b), and personalized recommendation via inverse regression (Figure 1c).

[FIGURE 1]

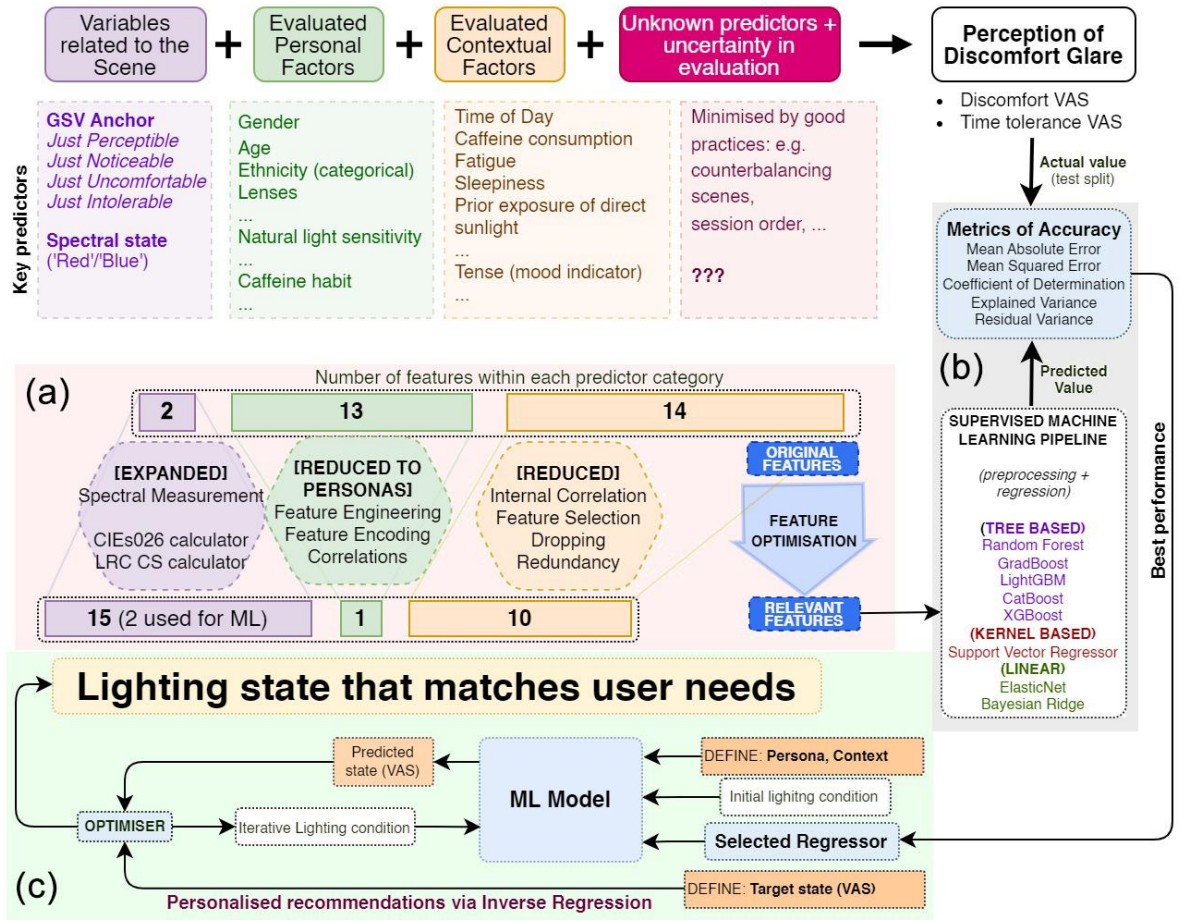

Figure 1. Data treatment in three distinct steps: (a) feature optimization, (b) supervised machine learning, (c) Inverse regression and state optimization.

The feature optimization step aimed to reduce dimensionality, since high number of features can lead to model overfitting and a reduced generalization for external datasets. In this step, the original features were transformed into relevant counterparts, after which regression models were developed using eight supervised machine learning libraries readily available in Python. The modelling method that demonstrated the highest performance in prediction (or accuracy with respect to the actual data) was identified and used to create a portable regression model. This model – which was a reusable serialised binary file of pickle *.pkl format that encapsulated the model's structure and learned parameters – was investigated and deployed for optimizing the state of lighting for any user and contextual factors, when the target zone of

(dis)comfort was specified. Effectively, the original experiment-generated dataset was transformed to train a supervised machine learning model, which could optimize the state of lighting variables – specifically the luminance and the spectral characteristics – and deliver the target zone of perceived (dis)comfort for a defined user against constrained contextual settings.

3. Results

3.1 Multi-Feature Dataset

The original dataset included 37 columns across 1824 rows, with each row corresponding to the VAS responses provided by 38 individual participants under 48 data-collection scenarios (four GSV-anchors across two spectral states, on two types of VAS scales, at three times of day). The key predictors related to the exposure were: the four distinct GSV anchor for luminance (Just Perceptible 'JP', Just Noticeable 'JN', Just Uncomfortable 'JU', Just Intolerable 'JI'), and the two spectral states of the scenes ('Red', 'Blue'). Additionally, three times of day (Morning 'M', Afternoon 'A', 'Evening 'E'), and the two VAS scales used for data collection ('discomfort scale', 'time tolerance scale') were also key objective predictors. Also included in the dataset were 27 subjective predictors related to personal factors (13) and to contextual factors (14). Against these predictors, the responses measured on the VAS scales served as the outcome variable. Effectively, the dataset contained 29 features as predictors, and one as the outcome. The outcome variable (VAS) refers to the subjective rating given by participants on one of two visual analog scales: a discomfort glare scale ('Disc') and a time-tolerance scale ('Tme'). Each participant provided ratings on both scales against every lighting condition. The type of scale that is scored in the outcome variable VAS is indicated in the 'Scale' column, which is included as a categorical predictor in all modelling processes. This is done to help the model account for potential differences in interpretation between discomfort

and time-based responses. All models are trained to predict the continuous VAS outcome, irrespective of the scale type. The remaining columns contained information that was not directly relevant for the analysis presented in this paper (e.g., participant ID, etc.). A complete list of predictor variables used in the modelling pipeline, along with their descriptions, is provided in Appendix A5 Table A5.1, which serves as a reference for all variables used in the ML modelling tasks.

3.2 Feature Optimization

For training adequate models while minimising the risk of overfitting, this first step (described in Figure 1a) reduced the dataset's complexity while retaining the relevant information. Here, overfitting refers to a common issue in supervised machine learning that prevents model generalization. A high number of features against a small dataset results in the model being trained not only on patterns in the data, but also on inherent noise [46], which must be avoided by cleaning and preprocessing the dataset, as well as by dropping redundant features. The following steps were taken to optimize the features on the predictor side.

3.2.1 Feature engineering

In this step, eight separate predictors within 'personal factors' were reduced to two via K-Means clustering and rule-based selection, respectively, which significantly reduced the dimensionality. K-Means clustering [47, 48] is an unsupervised machine learning algorithm that does not require data to be labelled, and which partitions the dataset into 'K' number of clusters (where K must be defined). The distance between data points and their corresponding cluster centroid is minimized, and each point is assigned to the nearest centroid, with centroids iteratively recalculated until convergence is reached [49]. Via K-means, five criteria nested within sensitivity to light (i.e. *sensitivity to natural light, to artificial light, affinity for using sunglasses,* and *the preferences for blind states and lighting conditions*) were reduced to a

single feature of ‘photosensitivity’, which had three possible states: High, Moderate, or Low.

For clustering process, the five individual features were used as inputs, along with the mean VAS score for each participant across the 48 scenarios. Python Scikit-learn’s ‘OneHotEncoder’ and ‘StandardScaler’ functions [50] were used to encode the dichotomous features, and to standardize the categorical features using minimum-maximum scaling, respectively. By defining ‘K’ as 3, participants were clustered into three distinct groups, which were then inferred based on feature values as having different levels of photosensitivity. The use of K-Means reduced the risk of subjective bias in manual clustering (which could have used ideal centroids of minimum, mid-point, and maximum thresholds for each feature). The use of this unsupervised machine learning technique retained the complex interactions between the individual nested features.

The three clusters generated by K-Means clustering were designated as clusters ‘0’ (11 participants), ‘1’ (17 participants), and ‘2’ (10 participants), and are summarized in Table 4 together with cluster-wise statistics, including ranges and frequency distributions. Figure 2 presents the distribution of participants within each cluster, which is visualized separately for the dichotomous (top row) and for the continuous (bottom row) predictors for each group.

[FIGURE 2]

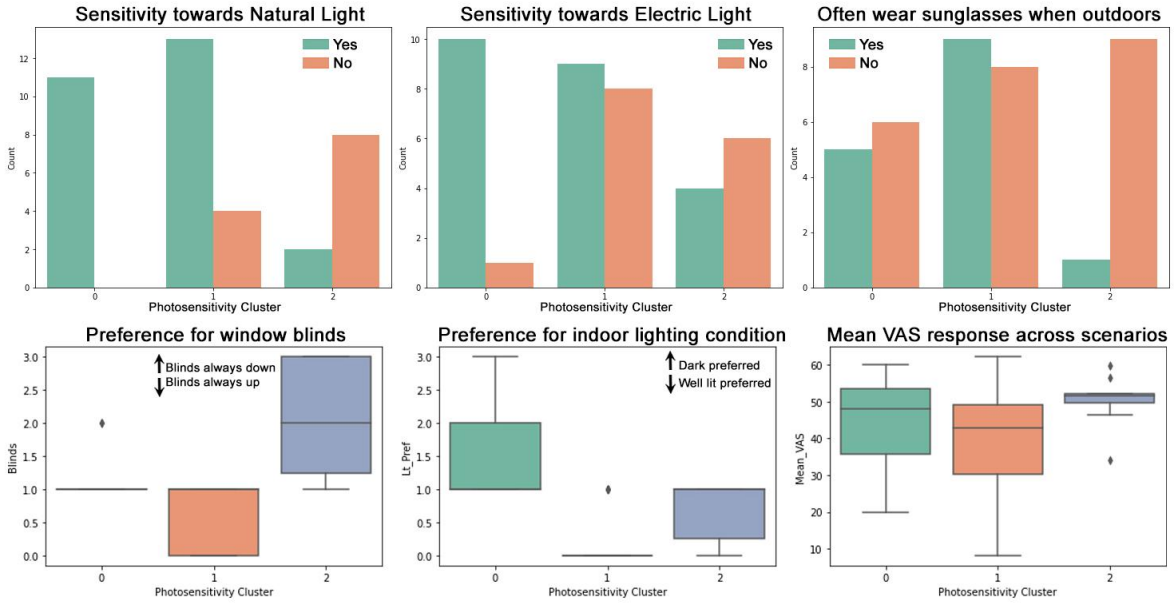

Figure 2. Distribution of participants within each of the 3 clusters for the dichotomous parameters (top row), and for the categorical variables (bottom row).

[TABLE 4]

Table 4. Frequencies and ranges of data distribution for participant attributes between clusters. The last column includes the interpreted photosensitivity, and the footnote describes the meaning of the anchors and values.

	Range			**Frequencies**			**Photosensitivity**
Cluster	**'Blinds'**	**'Lt_Pref'**	**'Mean_VAS'**	**'Nat_Lt'**	**'Art_Lt'**	**'Sunglss'**	**(Interpretation)**
0	[1, 2]	[1, 3]	[19.77, 60.13]	0: 100%, 1 :0%	0: 90.9%, 1: 9.09%	1: 54.5%, 0: 45.4%	'High'
1	[0, 1]	[0, 1]	[8.16, 62.18]	0: 76.47%, 1: 23.52%	0: 52.9%, 1: 47.05%	0: 52.9%, 1: 47.05%	'Moderate'
2	[1, 3]	[0, 1]	[34.22, 59.88]	0: 20%, 1: 80%	0: 40%, 1: 60%	0: 10%, 1: 90%	'Low'

Feature description and interpretation:
Use of Blinds ('Blinds'): 0-4, with 4= blinds always down, 0 = blinds always up
Preference for lighting condition ('Lt_Pref'): 0-3, with 3 = dark preferred, 0 = well lit preferred
Sensitive to Natural Light ('Nat_Lt'): 0 is Yes (i.e. sensitive), 1 is Not sensitive.
Sensitive to Artificial Light ('Art_Lt'): 0 is Yes, 1 is Not sensitive.
Use of Sunglasses ('Sunglss'): 0 is Yes (meaning sensitive), 1 is No (could mean not sensitive).
Mean reported discomfort due to glare ('Mean_VAS): Higher discomfort indicates higher photosensitivity.

From the data presented in Table 4 and Figure 2, it is evident that all participants within the first cluster were sensitive to natural light, and about 90% to artificial light. This group also had an indoor lighting preference that tended towards 'dark preferred', and was therefore interpreted as having a high photosensitivity. Conversely, the third cluster had only 20% and

40% participants sensitive to natural and artificial light, respectively, which was lower than the other two clusters, in ordered sequence. The light preference for this group was towards 'well lit' – therefore this group was evaluated as the least photosensitive. The preference for wearing sunglasses also showed tendencies in the same direction, with a reducing proportion of each group indicating 'yes' along with their interpreted photosensitivity, i.e. the high photosensitive group also had the highest proportion of responses towards 'yes' (on frequent use of sunglasses). However, this aspect could be confounded by cultural habits, since the use of sunglasses is not common in Belgium. Effectively, 11, 17, and 10 subjects were categorized as having a 'high', 'moderate', and 'low' photosensitivity, respectively.

In parallel, the three features nested within eye-condition (*whether they wore corrective lenses, had functional colour vision, and reported any other vision problem*) were reduced into a singular feature via rule-based classification [51], which generated three categories of eye condition: 'poor', 'moderate', and 'good. Together, the steps within feature engineering allowed us to combine eight disparate predictors into two engineered features, while augmenting the quality of these composite features.

3.2.2 Participant profiling through feature encoding

By creating two composite features, the total number of features within the 'personal factors' category (Figure 1) was substantially reduced. This resulted in seven key features that together described the participants' personal characteristics: *age, gender, ethnicity, eye condition, chronotype, photosensitivity,* and *caffeine habits*. Among these, eye condition and photosensitivity are the composite features that were created in the feature engineering step. Chronotype refers to a biological construct that defines an individual's internal circadian rhythm as well as their natural preference for sleep and wake times, and provides an indication of when they feel most alert or tired during a 24-hour period. Three defined categories have

been used: Morning-types (M-types), Evening-types (E-types), and Neither-types (N-types), where the third category classifies individuals (about 60% of the population) that do not have clear circadian preferences, and who show intermediate features between M-types and E-types [52, 53].

With each feature having 2 to 3 possible categories, there were 972 theoretical combinations that could be identified, and participants were profiled by encoding these features. Distinct profiles were created, with a superscript above each attribute – such as photosensitivity, or gender – that defines the state of each encoded attribute. For example: *a young^{0} woman^{1} of European^{0} ethnicity with good eye condition^{2}, N-chronotype^{1}, high photosensitivity^{0}, and no caffeine habit^{1}* was designated as 'P0102101'; or, *a middle-aged^{1} man^{0} of Asian^{2} ethnicity with moderate eye condition^{1}, M-chronotype^{0}, and moderate photosensitivity^{1}, who regularly consumes caffeine^{0}* was designated as 'P1021010'. Out of the 972 theoretical profiles based on combinations of categorized features, further analysis revealed that our participants could be clustered within 35 distinct combinations. The parallel coordinate plot in Figure 3 presents the theoretical combinations, as well as those for the profiles found in our participants. The thin dark lines in the background represent 96.3% of the combinations that were not found (for example, participants having a mixed ethnicity and poor eye-condition), while the profiles of participants found in our dataset are represented using dominant lines. The encoded profiles are further connected to mean VAS scores, that were reported by each profile across the 48 scenarios.

[FIGURE 3]

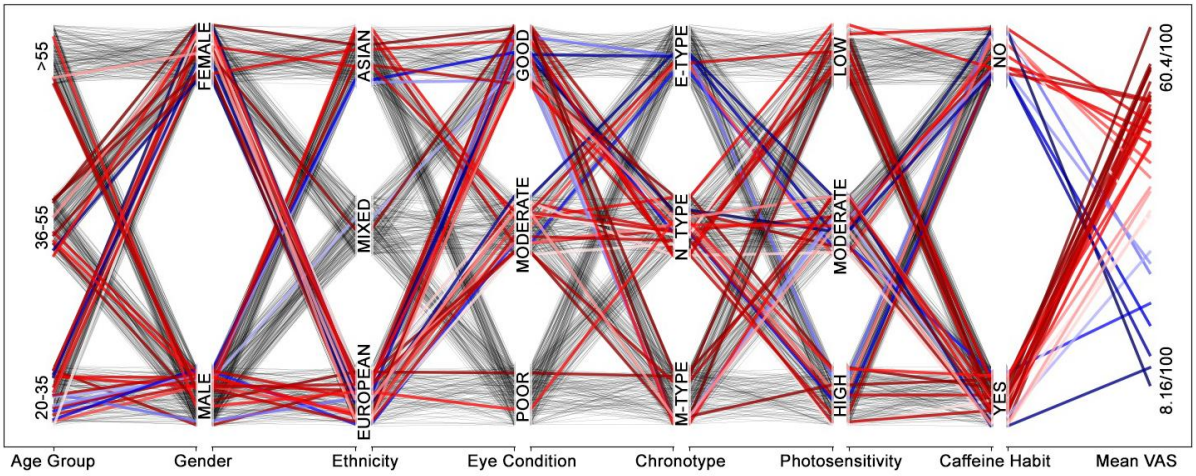

Figure 3. Parallel coordinate plot demonstrating the combinations of encoded features of personal factors based on the defined categorization. Each combination represents a profile, with all theoretical and actual profiles represented by dark connectors and dominant connectors, respectively.

3.2.3 Generating user personas

The main ambition of the feature optimization step was to aggregate personal factors into synthetic user 'personas'. Personas have already been used in the lighting literature to represent typical users, as well as to highlight their luminous needs and responses to support a user-centred design [54]. Prior steps within feature optimization transformed the raw features into meaningful composites, and unique combinations of attributes within each meaningful feature were represented as 35 encoded profiles. Through this step, the profiles were further clustered based on their similarity of response, and personas were generated as higher order representation of participants.

To achieve this clustering, the 35 distinct profiles were analysed via the Spearman rank correlation method, based on their reported discomfort due to glare across the 48 different lighting scenarios. This method performs a non-parametric measure of relationship between two variables, and describes how consistently any variable responds (i.e., increases or decreases) with respect to change in the other. This method was preferred over the Pearson

correlation method that assesses linear relationships, since Spearman rank correlation can be applied to data that may not have a normal distribution [55]. The '*spearmanr'* function in stats library of Python SciPy was used for performing this function, and the resulting matrix from pairwise correlation between groups is presented in Figure 4. Encoded profiles are presented on the two axes, where the seven sequential indices (such as '2101010') represent the individual features of personal factors within each profile.

[FIGURE 4]

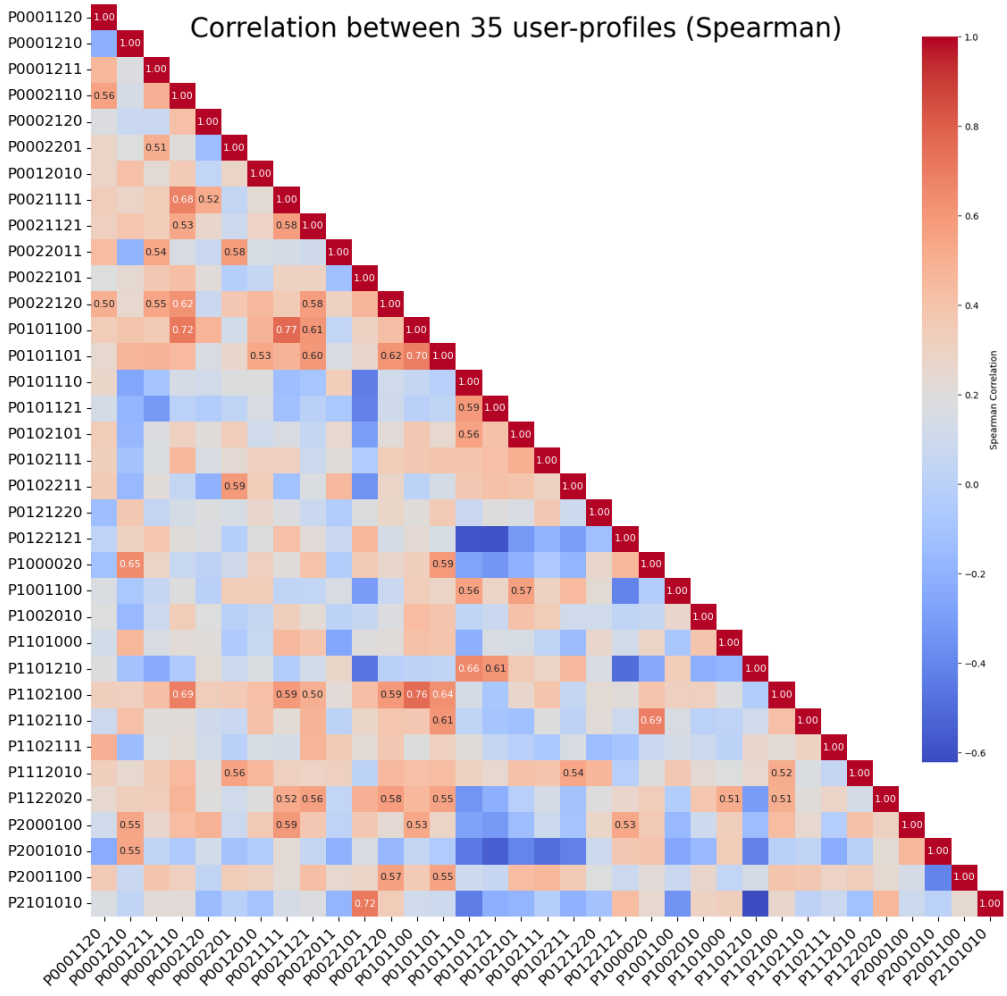

Figure 4. Correlations between profiles of attributes, drawn based on the VAS scores across the 48 scenarios. On both are the 35 profiles of users defined via encoded attributes of personal factors. Threshold for interpretation: >0.5 (Moderate). 53 of the 595 combinations demonstrate correlations above the threshold.

All combinations with moderate or higher correlations (coefficient ρ>0.5, based on [56]) were isolated, and it was observed that 53 of the 595 possible combinations (35*34/2) demonstrated internal correlations above the set threshold. Table 5 presents these combinations, where mutual pairs of profiles are presented via encoded features.

[TABLE 5]

Table 5. The 53 combinations demonstrating correlations (via Spearman method) above the ρ>0.5 threshold. These were part of three distinct sets, as indicated with the overlay colour.

#	Profile A	Profile B	Corr	#	Profile A	Profile B	Corr	#	Profile A	Profile B	Corr
1	P0002110	P0001120	0.561	19	P0101101	P0101100	0.698	37	P1112010	P0002201	0.559
2	P0002201	P0001211	0.507	20	P0101121	P0101110	0.592	38	P1112010	P0102211	0.543
3	P0021111	P0002110	0.684	21	P0102101	P0101110	0.561	39	P1112010	P1102100	0.516
4	P0021111	P0002120	0.518	22	P0102211	P0002201	0.593	40	P1122020	P0021111	0.524
5	P0021121	P0002110	0.532	23	P1000020	P0001210	0.650	41	P1122020	P0021121	0.565
6	P0021121	P0021111	0.578	24	P1000020	P0101101	0.591	42	P1122020	P0022120	0.576
7	P0022011	P0001211	0.539	25	P1001100	P0101110	0.557	43	P1122020	P0101101	0.551
8	P0022011	P0002201	0.583	26	P1001100	P0102101	0.568	44	P1122020	P1101000	0.514
9	P0022120	P0001120	0.501	27	P1101210	P0101110	0.658	45	P1122020	P1102100	0.505
10	P0022120	P0001211	0.546	28	P1101210	P0101121	0.613	46	P2000100	P0001210	0.547
11	P0022120	P0002110	0.625	29	P1102100	P0002110	0.691	47	P2000100	P0021111	0.595
12	P0022120	P0021121	0.581	30	P1102100	P0021111	0.590	48	P2000100	P0101100	0.533
13	P0101100	P0002110	0.719	31	P1102100	P0021121	0.502	49	P2000100	P0122121	0.534
14	P0101100	P0021111	0.772	32	P1102100	P0022120	0.590	50	P2001010	P0001210	0.546
15	P0101100	P0021121	0.613	33	P1102100	P0101100	0.765	51	P2001100	P0022120	0.566
16	P0101101	P0012010	0.531	34	P1102100	P0101101	0.636	52	P2001100	P0101101	0.551
17	P0101101	P0021121	0.599	35	P1102110	P0101101	0.608	53	P2101010	P0022101	0.720
18	P0101101	P0022120	0.617	36	P1102110	P1000020	0.691				
	Set 1				**Set 2**				**Set 3**		

These combinations were inspected via network analysis – a graph-based method – to identify groups that were strongly correlated. For performing this analysis, each profile was represented as a node, and the edges connected those profiles (as nodes) that had a Spearman correlation coefficient higher than the set threshold of 0.5. This method was performed using the '*nx'* function of Python's '*networkx'* library, and allowed to understand how the profiles were interconnected based on the similarity in their responses. The outcome of network analysis is presented in Figure 5, from which it was identified that 31 of 35 profiles were part of three closed networks. These sets of profiles are also highlighted in Table 5. It must be noted

that the remaining four profiles were distinct, and did not demonstrate correlation above the threshold with any other.

[FIGURE 5]

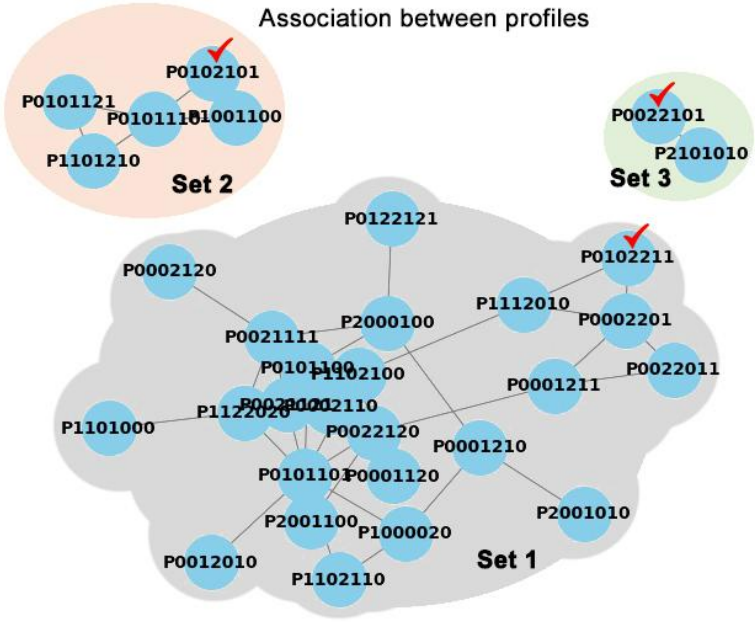

Figure 5. Network analysis for identifying association between profiles.

For each of the three networks indicated in Figure 5, a representative profile was selected (as indicated with a checkmark). By combining these with the other four distinct profiles, seven distinct personas were generated based on the similarity of their perception of lighting scenarios. This step allowed 80% of redundant profiles to be dropped, which enriched the data quality by increasing the number of data points for each included persona. These synthetic personas, each with distinct needs and responses, were designed to represent any user within our dataset, and to understand their responses while developing user-centred design [54]. The personas are presented in Table 6, which includes: the encoded profile of the representative for each persona; the number of profiles within each persona; the semantic attributes; and, an imaginary name arbitrarily assigned to represent each persona.

[TABLE 6]

Table 6. The seven distinct personas developed after feature engineering and encoding. Among these, three were representative of the sets of correlated profiles, while four represent the distinct profiles having no significant correlation with any other profile.

Encoding	n	Attributes	Name
P0022101	2	A young man of Asian ethnicity and a good eye condition, having a N-chronotype and a high photosensitivity, who typically doesn't consume caffeine	Mahendra
P0102101	5	A young woman of European ethnicity and a good eye condition, having a N-chronotype and a high photosensitivity, who typically doesn't consume caffeine	Aurora
P0102211	24	A young woman of European ethnicity and a good eye condition, having a E-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Julia
P0102111	1	A young woman of European ethnicity and a good eye condition, having a N-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Cora
P0121220	1	A young woman of Asian ethnicity and a moderate eye condition, having a E-chronotype and a low photosensitivity, who typically consumes caffeine	Dongmei
P1002010	1	A middle-aged man of European ethnicity and a good eye condition, having a M-chronotype and a high photosensitivity, who typically consumes caffeine	Johannes
P1102111	1	A middle-aged woman of European ethnicity and a good eye condition, having a N-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Elizabeth

3.2.4 Feature selection for contextual factors

Reducing features helps simplify machine learning tasks and, while features within 'personal factors' were significantly reduced (from 13 to one), the dataset also included 14 disparate contextual factors that varied between sessions within each participant. These included self-reported data specific to each session: *fatigue, sleepiness, three indicators of previous luminous exposure, caffeine consumption in the previous two hours, and eight indicators of their mood*, which have been reported to influence subjective perception of discomfort due to glare [26]. The large number of these features could still pose the risk of overfitting, and a feature selection process was conducted to reduce this risk.

An internal correlation was conducted using the Spearman correlation matrix method, and features having mutual correlation coefficient above 0.5 (interpreted as moderate correlation) [56] were filtered. The correlation matrix is presented in Figure 6.

[FIGURE 6]

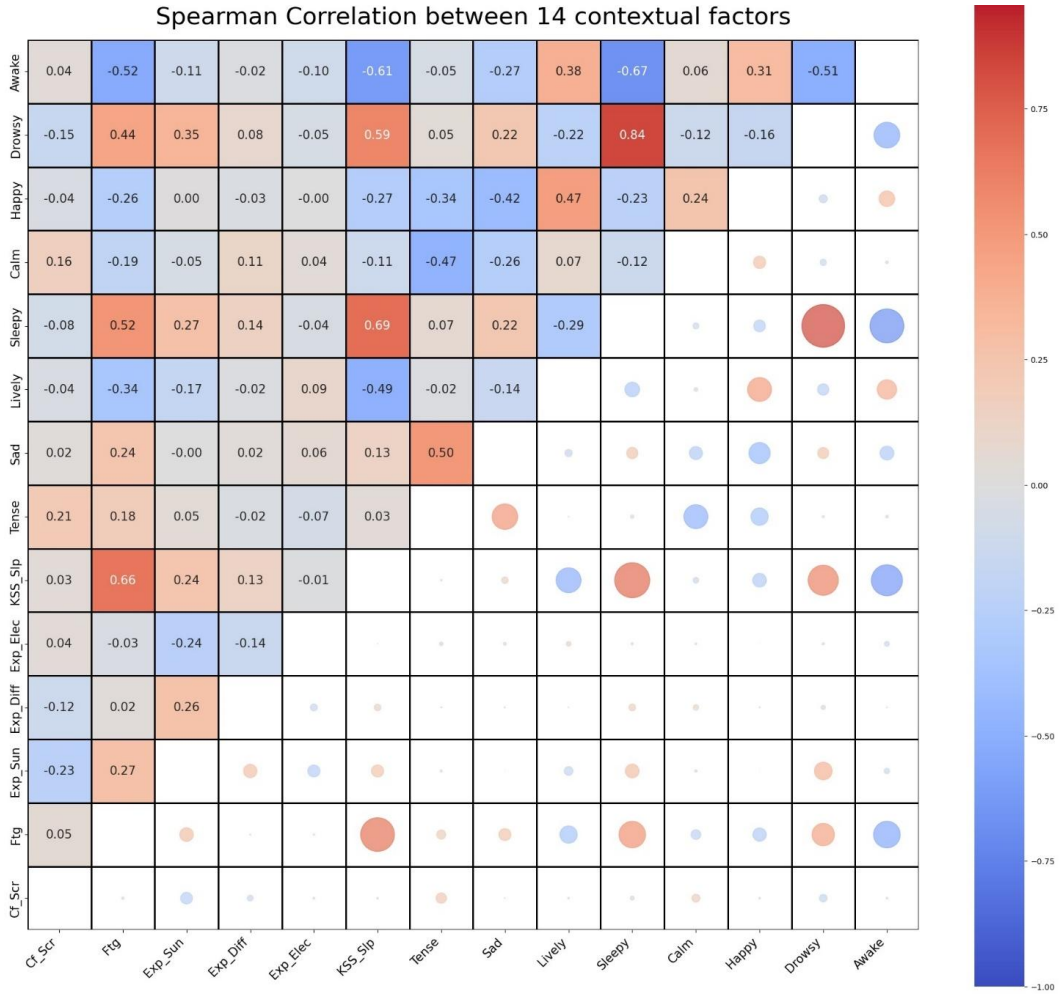

Figure 6. Correlations between the 14 contextual factors that can lead to intra-individual difference in responses. Threshold for interpretation: 0.2-0.5 RMPE (minimum effect size for a practically significant effect), 0.5-0.8 moderate, >0.8 strong.

It was observed that Fatigue (‘*Ftg’*), Sleepiness self-reported through the Karolinska scale (*‘KSS_Slp’*), and Mood indicators of ‘*Sleepy’*, ‘*Awake’*, and ‘*Drowsy’* demonstrated a ‘moderate’ (0.5) to ‘high’ (0.8) correlation. Moderate correlations were observed between the following pairs: *Fatigue/KSS* (0.66), *KSS/Sleepy* (0.69), and inversely between *Sleepy/Awake*(-0.67) and *Awake/KSS* (-0.61). Mood indicators of *Sleepy* and *Drowsy* demonstrated a high correlation, with the coefficient ρ evaluated at 0.84. This process allowed four of these five

features – *Sleepy*, *Drowsy*, *Awake*, and *KSS* – to be dropped, while Fatigue was retained to represent the others yet retaining the diversity in responses. It must be noted that between factors of prior luminous exposure, some correlations at the minimum level representing a "practically" significant effect (RPME) [56] (with a minimum threshold of 0.2) were identified, with direct sunlight/diffuse daylight presenting a correlation at $\rho = 0.26$, direct sunlight/electric light demonstrating inverse correlation at $\rho = -0.24$, while no substantive correlation was identified between diffuse daylight/electric light ($\rho = -0.14$). Beyond dropping the four correlated features, further reduction was not feasible without increasing residual (unexplained) variance. Effectively, this step allowed a reduction in contextual factors from 14 to 10 independent features.

The steps within feature optimization reduced the features of the original dataset while augmenting its quality. The transformed dataset now included seven distinct user personas as a singular feature, with ten contextual factors representing intra-individual variations. This dataset was then used for training supervised machine learning algorithms.

3.3 Supervised Machine Learning

~~In the following steps, eight state-of-the-art supervised machine learning algorithms were employed to train prediction models (Figure 1b), using the dataset enriched by the previous steps. As a first step, a simple linear regression model was implemented~~ To evaluate model performance, we first established a baseline using a simple linear regression model, using only scene-related variables: ~~—~~ ~~the~~ luminance of the glare source (GSV_{lum}) and the chromaticity category (CCT: Red/Blue) – to evaluate their capacity ~~to~~ for ~~estimate~~ estimating subjective responses to discomfort glare. This baseline provided a reference for comparing more advanced approaches. Following this step, eight state-of-the-art supervised machine-learning algorithms (Figure 1b) were trained on the enriched dataset, to assess their capacity of capturing non-linear

relationships, and for improving the accuracy of prediction.

The~~is~~ ~~linear~~ baseline model was trained on the dataset featuring 35 user profiles (N = 1680), ~~with~~ using an 80/20 train-test split~~,~~ and yielded a mean absolute error (MAE) of 20.51 with an R^2 score of 0.10, ~~this~~ indicating limited explanatory power. A log-transformed version of the model further reduced the performance (MAE = 22.08, R^2 = -0.05). These results confirm that linear models with variables associated uniquely ~~to~~ with the scene can ~~only~~ explain only a small portion of the variance in discomfort glare ratings, hence reinforcing the need to include personal and contextual predictors in advanced models.

~~Simultaneously, another approach was taken within this initial step~~In parallel, ~~where~~ the mean of the DG scores for subgroups with distinct states of source luminance ('GSV_{lum}', in four thresholds: JP, JN, JU, JI) and spectral states ('CCT', in three levels: Red, Blue, and the combin~~ing these two sub~~ed group~~s~~) were evaluated to investigate how the means changed with variables related only to the scene. While this showed that the mean discomfort glare (DG) scores increased with GSV_{lum}~~,~~ ~~–~~ ~~which validated~~ validating the alignment of GSV anchor with the VAS responses, the sensitivity of obtained mean DG VAS scores was found to be significantly reduced towards the extremes. ~~On the other hand, the~~ The effect of CCT was found to be relatively minor~~ by comparison~~, with 'Blue' spectral conditions generally associated with slightly higher responses on the mean VAS scores.

The mean discomfort glare (DG) scores on VAS against different states of the scene are presented in Appendix A5 Table A5.2. Considering the limited explanatory power of these linear regression models, further models with more predictors were developed using supervised ML techniques, as presented in the following sections.

3.3.1 Description of modelling methods

~~To evaluate the prediction performance across diverse modelling families~~Building upon the baseline results, eight supervised ML regressors were selected from three families: linear models (ElasticNet, Bayesian Ridge), kernel based (Support Vector Regression, with RBF kernel), and nonlinear tree-based ensemble learners (Random Forest, Gradient Boosting, XGBoost, CatBoost, and LightGBM). The model selection was informed by prior applications in relevant domains, such as the modelling of discomfort glare [32], visual comfort [31], lighting energy use [57], and daylight factors for regulatory compliance [58]. XGBoost, LightGBM, and CatBoost have previously demonstrated superior performance in dynamic assessment of heating and cooling loads [59], predictive HVAC control [60], solar PV prediction [61], and LED prognostics [62], while SVR and ElasticNet have served as baselines for interpretation.

In an initial implementation, default model parameters were used to prevent overfitting on a limited dataset. Hyperparameter tuning was later introduced for tree-based ensemble models (Random Forest, Gradient Boosting, LightGBM, CatBoost, and XGBoost) using randomized search over key parameter ranges (e.g., number of estimators, maximum depth, learning rate). This process incorporated 5-fold internal cross-validation, and boosting models included early stopping to avoid overfitting. All models used a fixed pseudo-random state for reproducibility. Linear models (ElasticNet, Bayesian Ridge) and SVR were retained with default settings due to their primary role as benchmarks, and due to limited advantages observed from tuning during pilot trials. The target variable for these methods was the perceived discomfort due to glare on a visual analogue scale, whereas the labelled predictors included factors related both to the observer (the persona), the lighting scene, and to the context. Within each method, the regression model was trained to predict the target variable through labelled predictors, which were either categorical or numeric.

Among tree based ensemble models, Random Forest regressor [63] is an algorithm that uses decision trees and averages outputs of each tree to minimize overfitting [64]. Gradient Boosting [65] is another method that combines the decision trees sequentially for reducing prediction errors [66]. CatBoost [67] is a variation of gradient boosting that is designed to handle categorical variables [68]. Extreme Gradient Boosting (XGBoost) [69] is another advanced implementation of gradient boosting that is optimized for scalability [70], while LightGBM is optimized for memory efficiency [71]. Among the second category, Support Vector Regression (SVR) [72] is a kernel-based method that maps data to higher dimensions, and finds a hyperplane that minimizes the error in prediction [73]. Among linear models, Bayesian Ridge Regression [74] performs probabilistic regression based on Bayesian principles [75], while ElasticNet [76] combines L1 Lasso and L2 Ridge regularization penalties into the loss function [77]. Each of these methods were used to fit the data and train separate models.

3.3.2 Model training

Each model pipeline included a preprocessing step, and a regression step using the selected regressor, in this sequential order. Under the preprocessing step, the predictors were segregated into numerical and categorical (non-numeric) types [50], and together with the outcome, were divided into 80% - 20% split that were respectively used for training the regression model and for internal evaluation of accuracy. This means that 80% of the dataset was used to train the model, while the remaining was used for testing its fit on unseen data. With these settings, the model pipeline was trained to fit the training dataset, and the trained pipeline was used to predict the VAS outcome based on the predictors in the test set.

To reduce the risk of sampling bias while assessing the robustness of model generalizability, a 5-fold cross-validation (CV) procedure was implemented for all eight

supervised ML regressors. The mean and standard deviation of the key performance metrics evaluated i.e. MAE and R^2 scores across folds, are reported in Table 7. These results confirm that XGBoost and CatBoost consistently outperform other models in both datasets, supporting the selection of XGBoost for the following steps of the evaluation.

[TABLE 7]

Table 7. Metrics of accuracy for each of the eight supervised machine learning models

Rank	Model	MAE ('*ClusAvg*')	R^2 ('*ClusAvg*')	MAE ('*ClusRep*')	R^2 ('*ClusRep*')
1–2	XGBoost	8.560 ± 0.534	0.631 ± 0.084	10.231 ± 0.461	0.716 ± 0.017
1–2	CatBoost	8.305 ± 0.497	0.707 ± 0.034	10.497 ± 0.336	0.705 ± 0.012
3	LightGBM	9.972 ± 0.696	0.608 ± 0.063	12.316 ± 0.392	0.620 ± 0.022
4	Random Forest	10.243 ± 0.740	0.579 ± 0.066	13.015 ± 0.538	0.570 ± 0.029
5	Gradient Boost	10.561 ± 0.792	0.573 ± 0.090	14.918 ± 0.475	0.470 ± 0.021
6	Bayesian Ridge	13.313 ± 0.655	0.394 ± 0.129	18.905 ± 0.821	0.180 ± 0.036
7–8	ElasticNet	14.847 ± 0.959	0.300 ± 0.036	19.981 ± 0.685	0.120 ± 0.011
7–8	Support Vector	16.386 ± 1.696	0.127 ± 0.037	19.431 ± 0.811	0.160 ± 0.019

In addition to model performance, we evaluated the stability of feature importance across training folds using a 5-fold CV implementation of the XGBoost model. Feature importances were recorded for each fold, and the mean and standard deviation were computed across folds. The coefficient of variation (SD/mean) was used to quantify consistency of feature importance. Features such as GSV_{lum}, Tense, Scale_Disc, Happy, Exp_Diff, and Tcp demonstrated high reliability, with Coefficient of Variation < 0.2, hence supporting their robustness as predictive features across training subsets. This analysis complements the SHAP-based insights in Section 3.3.5. A table describing the feature consistency and relevant metrics has been included in the Appendix A5 Table A5.3.

3.3.3 Interpreting performance

Among metrics of performance, mean absolute error (MAE) defines the average of the absolute

differences between the predicted and actual values, and evaluates the magnitude of deviations between the two. A lower MAE represents a good predictive performance. The coefficient of determination (R^2) measures the proportion of variance in the dependent variable that can be predicted from the independent variables, and evaluates how well the model fits the data. A high value of R^2 (close to 1) indicates a good model fit. Conversely, the residual variance is the difference between total and explained variance and, in this case, is approximately the same as 1-R^2. A low residual variance is expected from a good predictive model.

From the tabulated values in Table 7, it is evident that all tree-based regression methods outperformed the linear models and kernel-based methods, with XGBoost and CatBoost explaining close to 70% of the variance in the data. Among these, the Support Vector Regression (SVR) method demonstrated the least predictive capacity, despite using the Radial Basis Function (RBF) Kernel that can model complex non-linear data [78]. Ideally, this should have improved performance, since the potential non-linearity between the predictors and the outcome could rather explain the lower predictive capacity of linear models. Alternatively, the low performance of the SVR/RBF method could be attributed to its high sensitivity to outliers, compared to the tree-based models that are more robust to outliers in the data. The results from the scene-only linear regression model (Section 3.3) can also be interpreted using these metrics and contextualized with the outcomes presented in Table 7. With only GSV_{lum} and CCT as predictors in linear regression models, the performance was substantially lower, with $R^2 = 0.10$ being even lower than SVR's performance, which necessitated the incorporation of additional factors, i.e. contextual and personal features, to accurately predict discomfort glare ratings.

3.3.4 Model selection and evaluation

Considering the highest accuracy demonstrated by tree-based ensemble methods, the XGBoost method was used for further training and detailed evaluations. Predictions were performed for

the dataset, and metrics of accuracy were compared for two sets of predictors: a selected set with 18 predictors (18-P), and for another set with nine key predictors (9-P). The 9-P set included factors that we aimed to investigate, as well as factors previously reported to better explain the variance [26], including: *the glare source luminance (GSVlum), correlated color temperature (Tcp), Time of day (ToD), VAS scale used (Scale), Caffeine consumption score (Cf_Scr), Fatigue (Ftg), mood indicators of 'Tense' and 'Sad',* and *the indicated persona (rep_cluster).*

With a test-train split again set at 0.2, 1344 rows of data were used for training the two separate XGBoost models (18-P and 9-P), while 336 rows were used for prediction. Both models were ported in pickle format (.pkl) using the Python 'joblib' library and were used to predict VAS scores for the test split in a separate script. The predicted values were compared with actual values from the split dataset. Figure 7 compares the predicted versus the actual VAS scores for both models, and presents the mutual comparison between outcomes also through differential accuracy of model-predicted values against actual values.

[FIGURE 7]

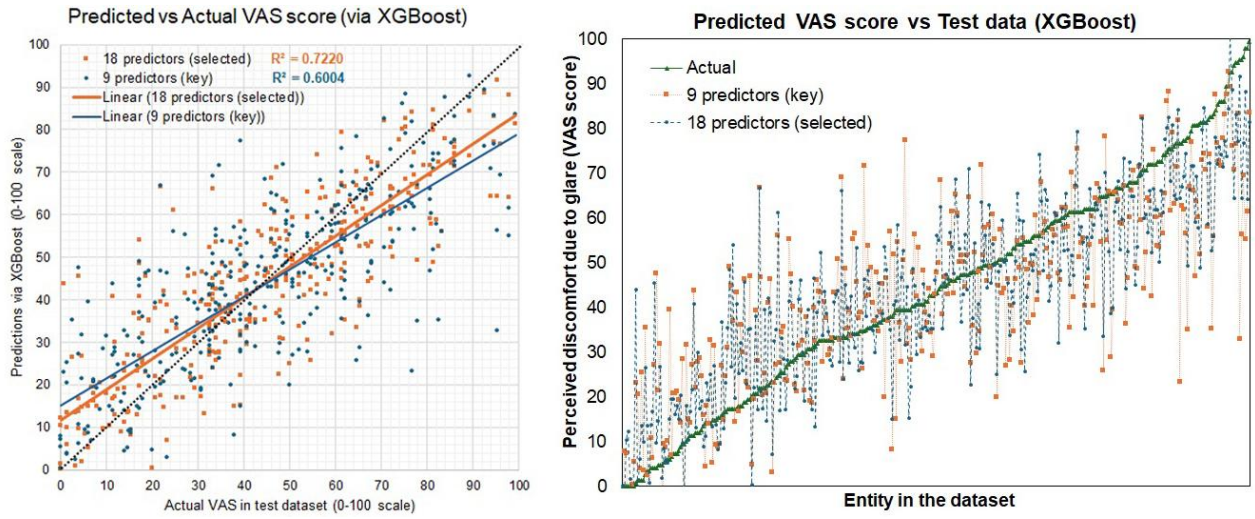

Figure 7. Comparing predicted and actual VAS scores for test dataset for two models trained on different number of predictors. (Left): Predicted values for both models against actual values; (Right):

actual VAS data (central line, 336 entities, sorted by increasing value) with deviations in the predictions made by the two models.

The 18-P model generated predictions having MAE, MSE, R^2 values and explained variance at 10.23, 173.46, 0.72, and 72% respectively, while the 9-P model demonstrated slightly reduced accuracy, with these metrics at 11.96, 251.36, 0.60, and 60%. Despite significantly reducing the number of predictors, the performance of the reduced-predictor XGBoost model was not found to decrease proportionally.

3.3.5 Feature importance

The importance of features within the 18-P model was identified using the '*feature_importances_*' function of the Scikit-Learn's '*Pipeline'* library, and SHAP (SHapley Additive exPlanations) was used to quantify feature contribution to the model using the Python 'SHAP' library. SHAP is an interoperable method of machine learning that is based on Shapley values from cooperative game theory [79]. By attributing changes in the output to individual differences in inputs, this method quantifies each feature's contribution towards the model's overall predictions [80]. SHAP has recently been used to gain insights into black-box models for discomfort glare from daylight [81]. A feature importance plot was generated using the *'Explainer'* and *'summary_plot*' functions in SHAP, while an interaction heatmap was evaluated via the *'TreeExplainer'* function to present interactions between features. The plots are presented in Figure 8.

[FIGURE 8]

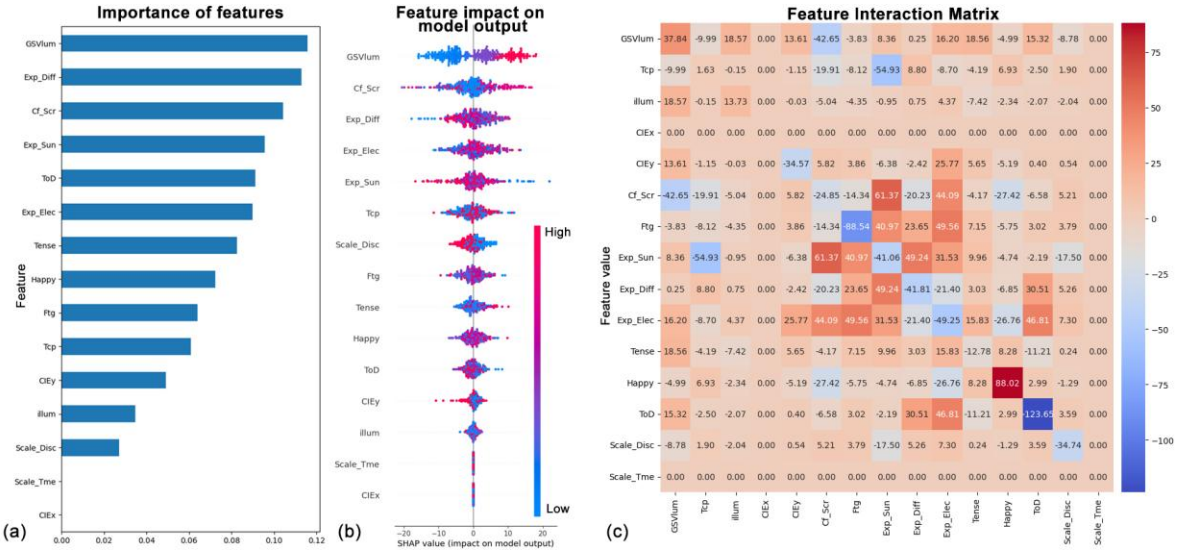

Figure 8. (a) Feature importance plot; (b) SHAP summary plot for feature importance; and, (c) SHAP interaction matrix between features, for the XGBoost model.

For the feature importance plot (Figure 8a), the X axis presents the relative importance of each feature, with the sum-total of all features being equal to 1. The SHAP summary plot (Figure 8b) presents the impact and distribution of features on the predicted value, where the magnitude of each feature on the model's outcome is shown by the horizontal axis. A positive value increases the prediction score, while a negative value reduces it. Through colour scales, each dot further represents the feature value for a specific data point. Through visual inspection, it can be observed that the SHAP summary plot closely resembles the feature importance plot, with both describing a multivariate dependence of the outcome on the predictors, i.e. no individual predictor was able to explain the outcome. Source luminance (*GSVlum*) clearly played a leading role, and the prior luminous exposures (*Exp_**) were also key to the predictions. Caffeine consumption (*Cf_Scr*) and Time of day (*ToD*) also played an important role in the model's outcomes. The CIE chromaticity coordinate and the vertical illuminance values did not significantly change between the eight scenes, which may have led to their low importance towards model prediction. The contextual predictors with the prefix '*Exp_**' identified in the SHAP analysis, i.e. *Exp_Sun*, *Exp_Diff*, and *Exp_Elec*, represent the

participants' self-reported light exposure to direct sunlight, diffuse daylight, and to electric lighting respectively – since they woke up on the day of the experiment until the time they arrived at the test environment. Since these features serve as proxies for luminance adaptation over the course of the day, their influence suggests that luminous history also plays an important role in the perception of discomfort due to glare, which highlights the relevance of incorporating intra-personal 'adaptation' within personalized lighting controls. These SHAP-derived insights are further supported by a feature stability analysis conducted across five CV folds using XGBoost, as described in Section 3.3.2, and tabulated in Appendix A5 Table A5.3.

The SHAP matrix (Figure 8c) represents the interaction between features, where the diagonal elements represent any feature's main effects on the predicted values, while mutual interactions (strength and nature) are represented by the non-diagonal elements. From the interaction matrix, it can be seen that not only were source luminance, caffeine score, and previous luminous exposures influential at the individual level, their interactions with other factors also amplified their net contributions. For example, the interactions between '*Cf_Scr'* with '*Exp_Sun*' (61.37), or of '*Cf_Sc*r' with *'Exp_Elec'* (44.09), significantly amplified the direction of the overall predictions. This step served to demonstrate insights into the black box model, and to identify which features were key to the predictions.

3.3.6 Sensitivity testing beyond the training dataset

The 18-P model was subsequently used to predict VAS scores and evaluate sensitivity of outcome in relation to change in the predictors for different personas. The contextual factors were defined as a reference set, with: Caffeine score at 10.5 (on a 0-15 continuous scale), Fatigue at 3 (on a 0-6 categorical scale); Exposure to direct sunlight/ diffuse daylight/ electric lighting all at 3 (on a 1-7 scale); and, Mood indicators (Tense, Sad, Lively, Calm, Happy) all at 2 (on a 1-5 scale). The variables related to the scene were defined as: Glare source luminance

at 5500 cd/m2 (between the 'JN' and 'JU' thresholds of 3860 and 9150 cd/m2 that ideally correspond to GSV markings between 33% and 66% on the VAS scale); Tcp at 3600 K; vertical illuminance at 400 lux; and, CIE 1931 coordinates of x at 0.4, y at 0.38 (which together yields a CCT value ~3537K). The predicted VAS scores for five personas for this reference set of predictors, as well as for individual changes in predictors, are presented in Table 8.

[TABLE 8]

Table 8. Predictions for the reference set of attributes, and those evaluated after altering one predictor each instance.

Persona	Step	Aurora	Cora	Dongmei	Johannes	Mahendra
VAS	0	49.73	49.29	54.3	55.74	49.95
Action: changing 'caffeine consumption' from reference (10.5) down to 0.5						
VAS	1	54.89	54.45	59.4	61.45	55.12
Action: *[Previous attributes]* + Changing 'fatigue' from reference (3) to 6						
VAS	2	38.52	37.36	42.85	44.48	42.93
Action: *[Previous attributes]* + Changing 'Tense' from reference (2) to 5						
VAS	3	36.15	35.45	40.91	42.67	37.42
Action: *[Previous attributes]* + Changing 'Sad' from reference (2) to 4						
VAS	4	35.61	34.99	40.41	42.14	36.88
Action: *[Previous attributes]* + Changing Scale from 'Disc' to 'Tme'						
VAS	5	40.96	39.54	40.85	42.78	39.8

For the reference set, the predicted discomfort due to glare reported by the personas on the discomfort VAS scale was observed to range between 49.73 to 55.74 (on a 0-100 scale). Changing predictors led to non-linear changes in outcome across personas. For example, changing the states of fatigue and caffeine consumption (a two-step-change from the reference state) led to a reduction between 14% (Mahendra, from 49.9 to 42.9) to 24% (Cora, from 49.2 to 37.3), with an average reduction of 20.4±3.8% compared to the baseline. Also, the influence of each change was not proportionate, and each step on average led to a relative change compared to the previous state, between 1.3% ('*sad'*, from average value of 38.52 in step 3, to 38.01 in step 4) to 27.7% ('*fatigue'*, from average value of 57.06 in step 1, to 41.22 in step 2).

While the model was sensitive to changes in predictor state, the accuracy in predictions could not be further characterised since these evaluations were conducted beyond the training dataset.

3.4 Personalized recommendation via inverse regression

As presented in Figure 1c, the trained XGBoost model was further employed to optimize lighting conditions to achieve a target perception state (expressed as a VAS score) for a specific user persona. This process identified the optimal values for each lighting related predictor, which could match a desired state of visual (dis)comfort, for any persona at any given time of day. An optimization-based prediction was performed via inverse regression [82, 83], where predictors were adjusted to produce the desired outcome. An initial state of predictors was defined based on the mean values of numerical predictors (i.e., variables defining lighting state, and those defining contextual factors), and the '*minimize'* function of SciPy's '*optimize'* class was used to iteratively reduce the difference between the targeted and achieved VAS. The XGBoost regressor model was used to calculate predicted VAS scores after each iteration, while the optimization function aimed to reduce the difference between the target and the evaluated state of lighting predictors. The L-BFGS-B optimization algorithm (limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm [84] with Box constraints) was used to fine tune the hyperparameters of the model, which ensured that the range of parameters was respected. The minimum and maximum bounds for each numerical predictor from the original dataset was defined to limit the possible range of lighting conditions that were tested by the combination of the regressor with the optimization function.

Table 9 presents the state of lighting variables, which were evaluated to match a target discomfort state of 50 (on a 1-100 scale) on the discomfort VAS at 9h, against the following set of factors: Caffeine consumption score: 10.5 (on a 1-15 scale); Fatigue: 3 (on a 0-6 scale);

Prior exposure to direct sunlight: 3, to diffuse daylight: 3, to electric light: 3 (all on a 1-7 scale); Mood indicators of '*Tense'*: 2, *'Sad'*: 2, *'Lively'*: 2, *'Calm'*: 2, *'Happy'*: 2 (all on a 1-5 scale)].

[TABLE 9]

Table 9. The state of lighting variables optimized to match a target VAS on the discomfort scale at a given time of day, for a defined set of personal and contextual factors.

#	Variable	Optimized state	Possible range (from dataset)
1.	Luminance of source	9084.25	1627 to 21700
2.	Tcp of source	4420.75	2609 to 6304
3.	ΔUV	0.01	-0.0001 to 0.0141
4.	Vertical Illuminance	388.12	336 to 485
5.	CIE 1931 x coord	0.39	0.3159 to 0.4711
6.	CIE 1931 y coord	0.38	0.3545 to 0.4116
7.	CIE 1976 u coord	0.23	0.1908 to 0.2699
8.	CIE 1976 v coord	0.51	0.4818 to 0.5292
9.	s-cone EDI	213.77	85.57 to 400.46
10.	m-cone EDI	347.57	262.93 to 486.99
11.	l-cone EDI	391.03	333.96 to 494.30
12.	Rhodopic EDI	298.97	174.65 to 492.19
13.	Melanopic EDI	281.92	145.86 to 491.43
14.	Circadian Lighting	432.65	291.33 to 655.21
15.	Circadian Stimulus	0.38	0.31 to 0.46

To demonstrate the optimizer's adaptability, these intra-individual/contextual features were manually set to simulate a user's current context – and their states were not based on cluster means. This model-based optimizer is designed to adjust its output dynamically in response to real-time user inputs, supporting personalized and context-aware lighting recommendations.

3.4.1 Limitations of the optimization process

The optimization routine of the algorithm, however, was not found to be sensitive to changes in persona, which was likely due to the limited variability of lighting states in the training dataset. The experiment only included 8 lighting scenes – JP/JN/JU/JI across 'Red' and 'Blue' spectral states – and, therefore, 8 states across all 15 lighting related predictors. Nevertheless,

when this optimization algorithm was tested on a synthetic dataset (which was generated to have a similar structure as the enriched dataset, and the 1680 constituents within each of its 34 columns had similar descriptive statistics) a sensitivity in optimized states against different personas and contextual factors was observed. Yet, since the synthetic dataset had no reliability beyond a statistical similarity to the experimental data, the optimized state of lighting predictors was not further investigated.

4. Discussion

Machine learning (ML) techniques are gaining popularity for data analysis and modelling due to their capacity for handling complex datasets and for detecting non-linear relationships between variables. Unlike conventional statistical methods that are limited to pre-specified functional relationships (such as linear, polynomial, etc.), ML models can adapt to diverse data structures and prove effective for multivariate analyses. These advantages make ML techniques particularly interesting for treating datasets from lighting research experiments, especially those that are focused on modelling discomfort due to glare.

Predicting discomfort perception is, in fact, highly challenging since the subjective nature of personal evaluations often leads to large unexplained variances. Critically, current discomfort glare models focus exclusively on objective factors related to the lighting scene, while neglecting several person- and context-related factors that have been widely reported to influence the perception of discomfort due to glare[26-28, 85-87]. By disregarding these factors, any variance that could have potentially been explained becomes residual (i.e., unexplained).

In the dataset that was used for this paper, the experimental user-testing data had been collected with the aim to analyse the concurrent effects of personal and contextual factors on the prediction of discomfort due to glare. Yet, the presence of several features that were

difficult to treat via statistical tools already used within the scientific lighting literature (e.g., analysis of variance, multilevel modelling, etc.[88-94]) made it appealing to explore the application of supervised ML techniques, a category of machine learning methods that works with labelled features. Nevertheless, ML techniques also present certain challenges, such as the risk of overfitting and the need for adequately pre-processed datasets for training the models. Additionally, the "black box" nature of ML models can make interpreting their outcomes difficult.

To address these challenges, the methods illustrated in this paper performed data preprocessing and feature optimisation via techniques such as K-means clustering, rule-based classification, feature encoding, and feature selection. Following up, synthetic Personas were generated to represent the user base, and to reduce the number of predictors from 27 to 11. This generated a high-quality dataset with reduced features, and eased the training task while reducing the risks associated with data sparseness and overfitting. Among the eight different ML methods employed for training, a wide range of accuracy was identified among the models, with Support Vector Regressor explaining as low as 18% of the variance in data, while tree-based regression models of CatBoost and XGBoost explaining as high as 70% with a 5-fold cross validation. While reducing the number of predictors in the XGBoost model did not lead to a proportional reduction in the model's performance (72% to 60% after reducing half the predictors, which demonstrates the robustness of this technique), the drop in accuracy was still significant. By evaluating feature importance and summary using SHAP method, it was identified that no feature individually dominated the importance value, and about 13 features cumulatively explained the predictions, which were also supported by a feature stability analysis conducted across 5 CV folds Sensitivity testing on an external set of conditions demonstrated non-linear change in model-predicted outcome across personas (between 14% to 24%, for a given set of conditions), and also a non-linear change in predictions across each

change in state of predictors (between 1.3% to 27% for specific sets of actions). These observations illustrate the advantage of using multivariate modelling capacities of supervised ML for making the prediction models, whereas tree-based regression methods were proven superior to kernel-based methods owing to their robustness against outliers. The capacity of explaining variance is even more relevant for this case, since the lighting scenes were designed using the IES-GI model, which does not consider subjective factors related to the user nor to the context. All variance in responses detected within each the four GSV-anchored subsets, therefore, had to be explained only via subjective factors, else they became unexplained residuals. With the 18-P XGBoost model (of which 13 features demonstrated importance) explaining up to 75% of the variance in data, the predictions exceeded our initial expectations.

The application of supervised ML for optimising the state of lighting factors via inverse regression was particularly innovative, since it can open up the avenue for wider applications for user-centric lighting based on personalised perception. However, the portable regression model did not prove sensitive to changes in personas. This inadequacy can be attributed to the limitation of our dataset, since the key predictors (including the state of lighting) were strictly maintained while collecting subjective data, that is perception of discomfort due to glare. The dataset included only 8 states of lighting variables against over 1800 reported states of discomfort. While the tight control on predictors proved particularly effective for regression-based modelling, the efficacy was much reduced in inverse regression where the state of discomfort was constrained (while the model was used to predict the optimal state of lighting variables, in combination with an iterative optimiser). The lack in sensitivity for inverse regression can, thus, be attributed to data sparseness in lighting states within our dataset. To address this, and to achieve goals such as inverse regression in future studies, data collection scenarios must combine category rating tasks (i.e., tightly controlled lighting states against subjective perception of discomfort due to glare – as in this study) with luminance adjustment

tasks (i.e., tightly controlled states of subjective perception against variable lighting states, as in [95, 96]), or data from field studies using daylight that simultaneously measure the scene variables along with subjective factors. The novelty of our study consists in its integration of psychophysical training data, rich in contextual and individual-level features, within a supervised ML framework for modelling discomfort due to glare. In contrast to simulation-driven models, such as those developed by Nourkojouri et al. (2021) and Li et al. (2025), our work anchors predictions with training data using subjective user ratings collected under controlled experimental conditions. The inclusion of contextual factors (e.g., time-of-day, caffeine intake, fatigue, etc.) together with our clustering of individual profiles into personas, enables us to account for a significant variability in individual glare response. By comparing multiple non-linear regression models through 5-fold cross-validation, and by further analysing their consistency in performance and stability in feature importance against the different training datasets, our study offers a rigorous and replicable modelling framework. Additionally, the demonstrated capacity for optimization supports not only a static evaluation of discomfort due to glare, but also an active process towards a personalized lighting control.

Although conventional models for discomfort glare rely on characteristics related to the luminous scene – such as the size and position of glare source, background luminance, etc. – the absence of such geometric and photometric variables in our models is uniquely due to the specific setup constraints of our controlled experiment. In our study, the IES-GI model relied upon set luminous conditions that achieved four thresholds of glare sensation votes, whereas the glare source's luminance was the only degree of freedom, while all other factors related to the scene were constrained. On these bases, our models use variables associated to the individual and contextual factors as additional predictors to glare source luminance. Although data on these attributes may be less readily accessible, our findings demonstrate their relevance in improving the accuracy of prediction, hence offering a pathway toward personalized lighting

control, especially considering the increase in wearable/nearable and environmental sensors, and the future scopes for acquiring such data in real time. While our results are specific to the dataset used for this study, the methods for data treatment, modelling, and optimisation, are widely transferable to other applications. These could be easily relevant to sister domains, e.g. for thermal comfort studies[97], or be upscaled to the urban domain for outdoor lighting research. As an example, for the latter, multiple contrasting priorities need to be balanced by lighting states, such as providing sufficient light for safety and visibility while giving attention towards excessive luminous exposure at nighttime and avoid phenomena of circadian disruption[98-101]. Balancing diverse demands for different stakeholders, and providing the right light to the right user (at the right time), can indeed present a prolific future field of application for the methods presented in this paper. A limitation of this study is that the models were developed and evaluated on a dataset collected from a single experimental study. Considering the absence of comparable datasets that capture both scene-related and individual/contextual factors, external validation could not be performed. While this limits the model's immediate applicability to other environments, the results highlight the potential scope for incorporating such data in modelling discomfort due to glare. Future work should focus on replicating these methods with new datasets collected in different settings.

5. Conclusions

A method based on supervised machine learning is demonstrated for analysing subjective data collected from a controlled laboratory experiment, where the sensation of discomfort due to glare was evaluated under various scenes created using tunable electric lighting. The dataset included predictors related to the lighting scene, personal variables that characterised participants, and contextual factors that were specific to each session. The data was treated via preprocessing, including feature engineering, encoding, and feature selection, and individual

test subjects were characterised into distinct personas having differing lighting needs. In parallel, redundant contextual features were identified and selected to improve feature quality. Although the models generated by this study have not been externally validated, this study intends to offer an approach that can inform future research aiming to personalize prediction and controls related to lighting conditions.

The optimised dataset was used to train eight machine learning models, which were compared based on metrics of accuracy. XGBoost emerged as the best performing model, and was further investigated for comparing its predicted scores with actual scores from the test dataset. Sensitivity analysis was conducted, and the model was seen as sensitive to changes both in personas as to those in contextual factors. The deployment of this portable model was also explored for optimizing the state of lighting variables via inverse regression, where the lighting states specific to each user were evaluated by defining the target perception zone against every contextual factor. The application of personalised controls makes this method interesting for applications beyond discomfort glare prediction. While the transferability of these models is limited by the fixed configuration of the lighting scenes in our study, and by the requirement for readily available personal and contextual data, our study lays the pathway for integrating personal and contextual variables into discomfort glare modeling.

Nomenclature

CIE:	Commission internationale de l'éclairage (International Commission on Illumination)
GSV:	Glare Sensation Vote
IES-GI:	Illuminating Engineering Society Glare Index
JI:	Just Intolerable
JN:	Just Noticeable
JP:	Just Perceptible
JU:	Just Uncomfortable
LRC:	Lighting Research Centre
MAE:	Mean Absolute Error
ML:	Machine Learning
MSE:	Mean Squared Error
RBF:	Radial Basis Function
RMPE:	Recommended minimum effect size representing a “practically” significant effect
SHAP:	SHapley Additive exPlanations
SVR:	Support Vector Regression
Tcp/CCT:	Correlated Colour Temperature
ToD:	Time of Day
XGBoost:	Extreme Gradient Boosting
VAS:	Visual Analog Scale

Appendices: (attached at the end)

1) Visual Analogue Scale used for evaluating perception of discomfort due to glare.
2) Details on the experiment (setup and timeline)
3) Questionnaire for collecting personal factors. (Background Questionnaire)
4) Questionnaire for collecting contextual factors. (In-study Questionnaire)
5) Appendix tables

Funding details:

This work was supported by FNRS - Fonds de la Recherche Scientifique (Belgium) through fellowships granted to the first author. Data collection was supported by the postdoctoral fellowship (n. 40000322) granted for the project SCALE (Shading Control Algorithms from Luminance-based Evaluations). Data analysis and manuscript preparation was supported by scientific collaborator fellowship (n. 40024409) granted for the project SPARKLE (Spectral Profiles for Agent-based Responsive networK of Lighting Energy).

Declaration of Generative AI and AI-Assisted Technologies:

For the data processing steps, the author(s) used ChatGPT, based on OpenAI's GPT-4 model, to refine and debug Python codes which were originally written by the author(s). The author(s) reviewed, tested, and adapted the suggestions as needed, and take full responsibility for the final implementation of the codes, and for the content of the publication.

Ethical Approval:

This study was approved by the IPSY Ethics Committee at UCLouvain (Project 2022-65)

Disclosure Statement:

The authors report there are no competing interests to declare.

References

1. Jung, D., et al., *Visual fatigue prediction using classification model based on physiological responses of occupants under office lightings.* Journal of Building Engineering, 2024. **98**: p. 111146.
2. Makridakis, S., et al., *Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward.* Journal of the Operational Research Society, 2023. **74**(3): p. 840-859.
3. Nichols, J.A., H.W. Herbert Chan, and M.A.B. Baker, *Machine learning: applications of artificial intelligence to imaging and diagnosis.* Biophys Rev, 2019. **11**(1): p. 111-118.
4. Sarker, I.H., *Machine Learning: Algorithms, Real-World Applications and Research Directions.* SN Computer Science, 2021. **2**(3): p. 160.
5. Manandhar, P., H. Rafiq, and E. Rodriguez-Ubinas, *Current status, challenges, and prospects of data-driven urban energy modeling: A review of machine learning methods.* Energy Reports, 2023. **9**: p. 2757-2776.

6. Ali, U., et al., *Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach.* Energy and Buildings, 2024. **303**: p. 113768.
7. Sun, Y., F. Haghighat, and B.C.M. Fung, *A review of the-state-of-the-art in data-driven approaches for building energy prediction.* Energy and Buildings, 2020. **221**: p. 110022.
8. Sutton, R.S., *Reinforcement learning: An introduction.* A Bradford Book, 2018.
9. Wu, Z., et al., *Reinforcement learning in building controls: A comparative study of algorithms considering model availability and policy representation.* Journal of Building Engineering, 2024. **90**: p. 109497.
10. Weinberg, D., et al., *A Review of Reinforcement Learning for Controlling Building Energy Systems From a Computer Science Perspective.* Sustainable Cities and Society, 2023. **89**: p. 104351.
11. Shan Lee, V.L., et al., *Semi-supervised Learning for Sentiment Classification using Small Number of Labeled Data.* Procedia Computer Science, 2019. **161**: p. 577-584.
12. Han, X., et al., *Prediction of human thermal comfort preference based on supervised learning.* Journal of Thermal Biology, 2023. **112**: p. 103484.
13. Putrada, A.G., et al., *Machine Learning Methods in Smart Lighting Toward Achieving User Comfort: A Survey.* IEEE Access, 2022. **10**: p. 45137-45178.
14. Zhuang, D., et al., *Supervised learning-based assessment of office layout satisfaction in academic buildings.* Building and Environment, 2022. **216**: p. 109032.
15. Mathew, V., C. Pearl Kurian, and N. Augustine, *Climate-responsive machine learning-based control of switchable glazing towards human-centric lighting.* Solar Energy, 2023. **260**: p. 49-60.
16. Kent, M.G., S. Fotios, and S. Altomonte, *Discomfort glare evaluation: The influence of anchor bias in luminance adjustments.* Lighting Research & Technology, 2019. **51**: p. 131-146.
17. Akashi, Y., R. Muramatsu, and S. Kanaya, *Unified Glare Rating (UGR) and subjective appraisal of discomfort glare.* http://dx.doi.org/10.1177/14771535960280040501, 1996. **28**(4): p. 199-206.
18. Chauvel, P., et al., *Glare from windows: current views of the problem.* http://dx.doi.org/10.1177/096032718201400103, 1982. **14**(1): p. 31-46.
19. Einhorn, H.D., *Discomfort glare: a formula to bridge differences.* http://dx.doi.org/10.1177/14771535790110020401, 1979. **11**(2): p. 90-94.
20. Petherbridge, P. and R.G. Hopkinson, *Discomfort Glare and the Lighting of Buildings.* Lighting Research and Technology, 1950. **15**(2 IEStrans): p. 39-79.
21. Robinson, W., et al., *The Development of the IES Glare Index System.* http://dx.doi.org/10.1177/147715356202700103, 1962. **27**(1 IEStrans): p. 9-26.
22. Wienold, J. and J. Christoffersen, *Evaluation methods and development of a new glare prediction model for daylight environments with the use of CCD cameras.* Energy and Buildings, 2006. **38**(7): p. 743-757.
23. Carlucci, S., et al., *A review of indices for assessing visual comfort with a view to their use in optimization processes to support building integrated design.* Renewable and Sustainable Energy Reviews, 2015. **47**: p. 1016-1033.
24. Kent, M.G., et al., *Discomfort glare and time of day.* Lighting Research & Technology, 2014. **47**(6): p. 641-657.
25. Rodriquez, R.G. and A. Pattini, *Tolerance of discomfort glare from a large area source for work on a visual display.* Lighting Research & Technology, 2012. **46**(2): p. 157-170.
26. Kent, M.G., et al., *Temporal variables and personal factors in glare sensation.* Lighting Research & Technology, 2015. **48**(6): p. 689-710.
27. Altomonte, S., et al., *Visual task difficulty and temporal influences in glare response.* Building and Environment, 2016. **95**: p. 209-226.
28. Pierson, C., J. Wienold, and M. Bodart, *Review of Factors Influencing Discomfort Glare Perception from Daylight.* LEUKOS, 2018. **14**(3): p. 111-148.

29. Nourkojouri, H., et al., *A machine-learning framework for daylight and visual comfort assessment in early design stages*. 2021.
30. Li, X., et al., *Highly generalized surrogate models for indoor glare performance based on multimodal deep learning networks at the early design stage.* Building and Environment, 2025. **273**: p. 112706.
31. Tzouvaras, C., et al. *A Novel Dynamic Approach for Determining Real-Time Interior Visual Comfort Exploiting Machine Learning Techniques*. Applied Sciences, 2023. **13**, DOI: 10.3390/app13126975.
32. Wagdy, A., et al., *Open-plan Glare Evaluator (OGE): A Demonstration of a New Glare Prediction Approach Using Machine Learning Algorithms*. 2019.
33. Maskarenj, M., M.G. Kent, and S. Altomonte, *Effect of time of day and correlated colour temperature on discomfort glare.* Lighting Research & Technology (Submitted), 2024.
34. Maskarenj, M. and S. Altomonte. *Mixed effect of time of day and correlated colour temperature on discomfort glare*. in *Bright Environments: Daylight in Sustainable Building Design*. 2023.
35. Afraz, A. and P. Cavanagh, *The gender-specific face aftereffect is based in retinotopic not spatiotopic coordinates across several natural image transformations.* Journal of vision, 2009. **9**: p. 10.1-17.
36. CIE, *CIE 212:2014. Guidance towards best practice in psychophysical procedures used when measuring relative spatial brightness*. 2014, Commission Internationale De L'Éclairage: Vienna, Austria.
37. Schlangen, L.J.M. and L.L.A. Price, *The Lighting Environment, Its Metrology, and Non-visual Responses.* Frontiers in Neurology, 2021. **12**(235).
38. CIE-Div-6, *CIE S 026 α-opic Toolbox*. 2020, CIE: Vienna.
39. LRC, *Lighting Research Centre Circadian Light Combined Calculator*. 2017: New York.
40. Roenneberg, T., et al., *Chronotype and Social Jetlag: A (Self-) Critical Review.* Biology (Basel), 2019. **8**(3).
41. Gawron, V.J., *Overview of Self-Reported Measures of Fatigue.* The International Journal of Aviation Psychology, 2016. **26**(3-4): p. 120-131.
42. Shahid, A., et al., *Karolinska Sleepiness Scale (KSS)*, in *STOP, THAT and One Hundred Other Sleep Scales*, A. Shahid, et al., Editors. 2012, Springer New York: New York, NY. p. 209-210.
43. Kompier, M.E., et al., *Effects of light transitions on measures of alertness, arousal and comfort.* Physiology & Behavior, 2020. **223**: p. 112999.
44. Kent, M.G., S. Fotios, and S. Altomonte, *Order effects when using Hopkinson's multiple criterion scale of discomfort due to glare.* Building and Environment, 2018. **136**: p. 54-61.
45. Fotios, S. and M. Kent, *Measuring Discomfort from Glare: Recommendations for Good Practice.* LEUKOS, 2021. **17**(4): p. 338-358.
46. Ying, X., *An Overview of Overfitting and its Solutions.* Journal of Physics: Conference Series, 2019. **1168**(2): p. 022022.
47. MacQueen, J. *Some methods for classification and analysis of multivariate observations*. in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*. 1967.
48. Lloyd, S., *Least squares quantization in PCM.* IEEE transactions on information theory, 1982. **28**(2): p. 129-137.
49. Li, Y. and H. Wu, *A Clustering Method Based on K-Means Algorithm.* Physics Procedia, 2012. **25**: p. 1104-1109.
50. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**(null): p. 2825–2830.
51. Hu, C.-C. and H.-L. Li, *Deducing the classification rules for thermal comfort controls using optimal method.* Building and Environment, 2016. **98**: p. 107-120.

52. Polańska, S., A. Karykowska, and Ł. Pawelec, *Associations between chronotype and physical activity and well-being in adults.* Chronobiology International, 2024. **41**(4): p. 521-529.
53. Montaruli, A., et al., *Biological rhythm and chronotype: new perspectives in health.* Biomolecules, 2021. **11**(4): p. 487.
54. Sokol, N., et al., *'Personas for lighting'. Three methods to develop personas for the indoor lighting environment.* Energy and Buildings, 2023. **278**: p. 112580.
55. Quek, G., et al., *Comparing performance of discomfort glare metrics in high and low adaptation levels.* Building and Environment, 2021. **206**: p. 108335.
56. Ferguson, C.J., *An effect size primer: A guide for clinicians and researchers.* Professional Psychology: Research and Practice, 2009. **40**(5): p. 532-538.
57. Bian, Y., et al., *Using machine learning algorithm to predict lighting energy consumption of daylight-linked lighting systems from spatial daylight autonomy.* Energy and Buildings, 2025. **341**: p. 115847.
58. Dieguez, A.P., et al., *Daylight factor prediction using machine learning: A two-way study using numerical encoding and regression models, versus image encoding and pix2pix.* Building and Environment, 2025. **274**: p. 112743.
59. Guo, J., et al., *Prediction of heating and cooling loads based on light gradient boosting machine algorithms.* Building and Environment, 2023. **236**: p. 110252.
60. Wang, H., et al., *An XGBoost-Based predictive control strategy for HVAC systems in providing day-ahead demand response.* Building and Environment, 2023. **238**: p. 110350.
61. Banik, R. and A. Biswas, *Improving Solar PV Prediction Performance with RF-CatBoost Ensemble: A Robust and Complementary Approach.* Renewable Energy Focus, 2023. **46**: p. 207-221.
62. Gupta Bakshi, B. and B. Roy, *A study on prognostics of LED luminaires based on gradient-boosting decision tree classifiers.* Lighting Research & Technology, 2025: p. 14771535251335974.
63. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
64. Linda, L., *Understanding Random Forests: An Ensemble of Decision Trees*. 2023.
65. Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial.* Frontiers in Neurorobotics, 2013. **7**(DEC): p. 21-21.
66. Baladram, S., *Gradient Boosting Regressor, Explained: A Visual Guide with Code Examples*. 2024.
67. Prokhorenkova, L., et al. *Catboost: Unbiased boosting with categorical features*. 2018.
68. Ershov, V., *CatBoost - open-source gradient boosting library*. 2018.
69. Chen, T. and C. Guestrin, *XGBoost: A scalable tree boosting system.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. **13-17-August-2016**: p. 785-794.
70. Tarwidi, D., et al., *An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach.* MethodsX, 2023. **10**: p. 102119-102119.
71. Ke, G., et al. *LightGBM: A highly efficient gradient boosting decision tree*. 2017.
72. Montesinos López, O.A., A. Montesinos López, and J. Crossa, *Support Vector Machines and Support Vector Regression.* Multivariate Statistical Machine Learning Methods for Genomic Prediction, 2022: p. 337-378.
73. Jain, S., *Support Vector Machine (SVM) Algorithm*. 2024.
74. Permai, S.D. and H. Tanty, *Linear regression model using bayesian approach for energy performance of residential building.* Procedia Computer Science, 2018. **135**: p. 671-677.
75. Koehrsen, W., *Introduction to Bayesian Linear Regression*. 2018.
76. Zou, H. and T. Hastie, *Regularization and Variable Selection Via the Elastic Net.* Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005. **67**(2): p. 301-320.
77. Jain, A., *Elastic Net Regression —Combined Features of L1 and L2 regularization*. 2024.

78. Musavi, M.T., et al., *On the training of radial basis function classifiers.* Neural Networks, 1992. **5**(4): p. 595-603.
79. Hokari, T. and W. Thomson, *Cooperative Game Theory*, in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, J.D. Wright, Editor. 2015, Elsevier: Oxford. p. 867-880.
80. Lundberg, S. and S.-I. Lee, *An unexpected unity among methods for interpreting model predictions.* arXiv [cs.AI], 2016.
81. Luo, Z., et al., *Investigation of influential variations among variables in daylighting glare metrics using machine learning and SHAP.* Building and Environment, 2024. **254**: p. 111394.
82. Devijver, E. and E. Perthame, *Prediction regions through Inverse Regression.* Journal of Machine Learning Research, 2020. **21**(113): p. 1-24.
83. Parker, P., et al., *The prediction properties of inverse and reverse regression for the simple linear calibration problem.* 2010(LF99-9222).
84. Evins, R., *A review of computational optimisation methods applied to sustainable building design.* Renewable and Sustainable Energy Reviews, 2013. **22**: p. 230-245.
85. Hamedani, Z., et al., *Lighting for work: A study of the relationships among discomfort glare, physiological responses and visual performance.* Building and Environment, 2020. **167**: p. 106478.
86. Marié, S., et al., *Evaluation of Physiological Parameters on Discomfort Glare Thresholds Using LUMIZ 100 Tool.* Transl Vis Sci Technol, 2021. **10**(8): p. 28.
87. Jain, S., et al., *Influence of macular pigment on the sensitivity to discomfort glare from daylight.* Scientific Reports, 2023. **13**(1): p. 18551.
88. Suk, J.Y., *Evaluating Daylight Glare and Identifying Its Dominant Cause in Computer-Based Office Task Environment*. 2017.
89. Kent, M., S. Fotios, and S. Altomonte, *Experimental biases in discomfort glare evaluations*. 2018.
90. Girard, J., C. Villa, and R. Brémond, *Discomfort Glare from Several Sources: A Formula for Outdoor Lighting.* LEUKOS, 2021. **17**(2): p. 108-124.
91. Pierson, C., et al., *Is there a difference in how people from different socio-environmental contexts perceive discomfort due to glare from daylight?* Lighting Research & Technology, 2021. **54**(1): p. 5-32.
92. Christoffersen, J. and J. Wienold, *Assessment of user reaction to glare with three solar shading systems*. 2008.
93. Stanke, L., L. Viktorová, and T. Dominik *Discomfort Glare Perception by Drivers—Establishing a Link between Subjective and Psychophysiological Assessment*. Applied Sciences, 2022. **12**, DOI: 10.3390/app12083847.
94. Chamilothori, K., et al., *Subjective and physiological responses towards daylit spaces with contemporary façade patterns in virtual reality: Influence of sky type, space function, and latitude.* Journal of Environmental Psychology, 2022. **82**: p. 101839.
95. Kim, W. and K. Jeong Tai, *A Position Index Formula for Evaluation of Glare Source in the Visual Field.* Indoor and Built Environment, 2010. **20**(1): p. 47-53.
96. Tuaycharoen, N. and P. Tregenza, *Discomfort glare from interesting images.* Lighting Research & Technology - LIGHTING RES TECHNOL, 2005. **37**: p. 329-341.
97. Kim, J., et al., *Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning.* Building and Environment, 2018. **129**: p. 96-106.
98. Davis, L.K., et al., *Health Effects of Disrupted Circadian Rhythms by Artificial Light at Night.* Policy Insights from the Behavioral and Brain Sciences, 2023. **10**(2): p. 229-236.
99. Wang, T., N. Kaida, and K. Kaida, *Effects of outdoor artificial light at night on human health and behavior: A literature review.* Environmental Pollution, 2023. **323**: p. 121321.

100. Lok, R., et al., *Timing of outdoor light exposure is associated with sleep-wake consolidation in community-dwelling older men.* Frontiers in Sleep, 2023. **2**.
101. Meléndez-Fernández, O.H., J.A. Liu, and R.J. Nelson *Circadian Rhythms Disrupted by Light at Night and Mistimed Food Intake Alter Hormonal Rhythms and Metabolism*. International Journal of Molecular Sciences, 2023. **24**, DOI: 10.3390/ijms24043392.

Figures

(Presented inline, attached as TIFF format at the end)

Tables

(appended below)

Table 1. Photometric and radiometric measurements for the eight scenes measured at eye level using Konika Minolta CL70F spectrometer (1nm resolution between 380-780nm). The CIE S026 a-opic metrics are derived from spectral data using the CIEs026 calculator, and the circadian metrics are derived using the LRC calculator.

					CIE 1931		CIE 1976		EDI based on CIE s-026 calculator						
CCT indicator	**GSV**	**Lum (cd/m^2)**	**Tcp (K)**	**Vert Illum (lx)**	**x-coord**	**y-coord**	**u-coord**	**v-coord**	**s-cone (lx, EDI_S)**	**m-cone (lx, EDI_M)**	**l-cone (lx, EDI_L)**	**Rods (lx, EDI_R)**	**melan. (lx, EDI_{mel})**	**Circad Lighting (lx, CL_A)**	**Circad Stim.**
Red	JP	1627	2645	336	0.4691	0.4116	0.2680	0.5292	85.57	262.93	342.50	174.66	145.86	291.33	0.3116
	JN	3860	2638	345	0.4693	0.4110	0.2684	0.5289	88.51	269.94	352.09	179.02	149.27	297.94	0.3159
	JU	9150	2620	394	0.4708	0.4112	0.2693	0.5292	99.62	306.96	401.71	201.87	167.12	333.64	0.3376
	JI	21700	2609	485	0.4711	0.4104	0.2699	0.5290	123.50	376.89	494.30	245.31	200.97	401.23	0.3732
		Mean	**2628**	**390**	**0.4701**	**0.4111**	**0.269**	**0.529**	**99.3**	**304.2**	**397.6**	**200.2**	**165.8**	**331**	**0.335**
		StDev	**16.5**	**68.26**	**0.001**	**0.0005**	**9E-04**	**1E-04**	**17.23**	**52.18**	**69.46**	**32.35**	**25.23**	**50.35**	**0.028**
Blue	JP	1627	6304	336	0.3159	0.3545	0.1908	0.4818	290.57	340.84	333.97	347.79	348.18	469.75	0.4032
	JN	3860	6271	343	0.3166	0.3551	0.1910	0.4822	295.25	348.13	341.47	355.41	356.16	479.01	0.4069
	JU	9150	6210	383	0.3179	0.3558	0.1916	0.4827	326.68	387.90	381.37	395.50	396.40	533.04	0.4268
	JI	21700	6069	483	0.3209	0.3588	0.1926	0.4846	400.46	487.00	480.85	492.20	491.43	655.22	0.4636
		Mean	**6214**	**386.2**	**0.3178**	**0.3561**	**0.192**	**0.483**	**328.2**	**391**	**384.4**	**397.7**	**398**	**534.3**	**0.425**
		StDev	**104**	**67.74**	**0.0022**	**0.0019**	**8E-04**	**0.001**	**50.75**	**67.28**	**67.57**	**66.37**	**65.74**	**85.34**	**0.028**

Table 2. Descriptive statistics on data related to personal factors. Sample size = 38 subjects

Gender	Age	Ethnicity	Eye conditions	Chronotype	Photosensitivity
52.6% *Male* 47.3% *Female*	Range: 22-64 Med: 32 34.4 ± 11.6	***White***: 73.6% ***Asian***: 21% ***Mixed***: 5.3%	42% wore lenses 14% reported other vision issues	7-point scale, 2.95 ± 1.4 Median: 3	**Natural Light**: 68.4% **Artificial Light**: 60.6% **Sunglasses**: 39.4% **Blind preference**: 1.15±0.85 on the 0-4 scale **Lighting preferred**: 0.73±0.75 on a 0-3 scale
Caffeine Habit					
Yes: 67.5%					

Table 3. Descriptive statistics on data related to contextual factors. N=114 scenes

Caffeine score	Prior Light Exposure (0-7)	Mood Indicators (1-5)	
21.1 ± 25.5%		***Tense***	1.91 ± 0.99
Continuous VAS	***Direct sunlight:***	***Sad***	1.53 ± 0.8
Fatigue (0-6)	5.1±1.8	***Lively***	3.37 ±0.82
2.14±1.29	***Diffuse daylight***	***Sleepy***	2.24±0.96
Samn-Perelli scale	3.9±1.6	***Calm***	3.7±1.02
Sleepiness (1-9)	***Electric light***	***Happy***	3.6±0.84
4.2±1.7	4.3±1.8	***Drowsy***	2.4±1.05
Karolinska scale		***Awake***	3.53±0.96

Table 4. Frequencies and ranges of data distribution for participant attributes between clusters. The last column includes the interpreted photosensitivity, and the footnote describes the meaning of the anchors and values.

	Range			Frequencies			Photosensitivity
Cluster	**'Blinds'**	**'Lt_Pref'**	**'Mean_VAS'**	**'Nat_Lt'**	**'Art_Lt'**	**'Sunglss'**	**(Interpretation)**
0	[1, 2]	[1, 3]	[19.77, 60.13]	0: 100%, 1 :0%	0: 90.9%, 1: 9.09%	1: 54.5%, 0: 45.4%	'High'
1	[0, 1]	[0, 1]	[8.16, 62.18]	0: 76.47%, 1: 23.52%	0: 52.9%, 1: 47.05%	0: 52.9%, 1: 47.05%	'Moderate'
2	[1, 3]	[0, 1]	[34.22, 59.88]	0: 20%, 1: 80%	0: 40%, 1: 60%	0: 10%, 1: 90%	'Low'

Feature description and interpretation:
Use of Blinds ('Blinds'): 0-4, with 4= blinds always down, 0 = blinds always up
Preference for lighting condition ('Lt_Pref'): 0-3, with 3 = dark preferred, 0 = well lit preferred
Sensitive to Natural Light ('Nat_Lt'): 0 is Yes (i.e. sensitive), 1 is Not sensitive.
Sensitive to Artificial Light ('Art_Lt'): 0 is Yes, 1 is Not sensitive.
Use of Sunglasses ('Sunglss'): 0 is Yes (meaning sensitive), 1 is No (could mean not sensitive).
Mean reported discomfort due to glare ('Mean_VAS): Higher discomfort indicates higher photosensitivity.

Table 5. The 53 combinations demonstrating correlations (via Spearman method) above the 0.5 threshold. These were part of three distinct sets, as indicated with the overlay color.

#	Profile A	Profile B	Corr	#	Profile A	Profile B	Corr	#	Profile A	Profile B	Corr
1	P0002110	P0001120	0.561	19	P0101101	P0101100	0.698	37	P1112010	P0002201	0.559
2	P0002201	P0001211	0.507	20	P0101121	P0101110	0.592	38	P1112010	P0102211	0.543
3	P0021111	P0002110	0.684	21	P0102101	P0101110	0.561	39	P1112010	P1102100	0.516
4	P0021111	P0002120	0.518	22	P0102211	P0002201	0.593	40	P1122020	P0021111	0.524
5	P0021121	P0002110	0.532	23	P1000020	P0001210	0.650	41	P1122020	P0021121	0.565
6	P0021121	P0021111	0.578	24	P1000020	P0101101	0.591	42	P1122020	P0022120	0.576
7	P0022011	P0001211	0.539	25	P1001100	P0101110	0.557	43	P1122020	P0101101	0.551
8	P0022011	P0002201	0.583	26	P1001100	P0102101	0.568	44	P1122020	P1101000	0.514
9	P0022120	P0001120	0.501	27	P1101210	P0101110	0.658	45	P1122020	P1102100	0.505
10	P0022120	P0001211	0.546	28	P1101210	P0101121	0.613	46	P2000100	P0001210	0.547
11	P0022120	P0002110	0.625	29	P1102100	P0002110	0.691	47	P2000100	P0021111	0.595
12	P0022120	P0021121	0.581	30	P1102100	P0021111	0.590	48	P2000100	P0101100	0.533
13	P0101100	P0002110	0.719	31	P1102100	P0021121	0.502	49	P2000100	P0122121	0.534
14	P0101100	P0021111	0.772	32	P1102100	P0022120	0.590	50	P2001010	P0001210	0.546
15	P0101100	P0021121	0.613	33	P1102100	P0101100	0.765	51	P2001100	P0022120	0.566
16	P0101101	P0012010	0.531	34	P1102100	P0101101	0.636	52	P2001100	P0101101	0.551
17	P0101101	P0021121	0.599	35	P1102110	P0101101	0.608	53	P2101010	P0022101	0.720
18	P0101101	P0022120	0.617	36	P1102110	P1000020	0.691				
	Set 1				**Set 2**				**Set 3**		

Table 6. The seven distinct personas developed after feature engineering and encoding through the previous steps. Among these, three were representative of the sets of correlated profiles, while four represent the distinct profiles having no significant correlation (defined as >0.5) with any other profile.

Encoding	n	Attributes	Name
P0022101	2	A young man of Asian ethnicity and a good eye condition, having a N-chronotype and a high photosensitivity, who typically doesn't consume caffeine	Mahendra
P0102101	5	A young woman of European ethnicity and a good eye condition, having a N-chronotype and a high photosensitivity, who typically doesn't consume caffeine	Aurora
P0102211	24	A young woman of European ethnicity and a good eye condition, having a E-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Julia
P0102111	1	A young woman of European ethnicity and a good eye condition, having a N-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Cora
P0121220	1	A young woman of Asian ethnicity and a moderate eye condition, having a E-chronotype and a low photosensitivity, who typically consumes caffeine	Dongmei
P1002010	1	A middle-aged man of European ethnicity and a good eye condition, having an M-chronotype and a high photosensitivity, who typically consumes caffeine	Johannes
P1102111	1	A middle-aged woman of European ethnicity and a good eye condition, having a N-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Elizabeth

Table 7. Metrics of accuracy for each of the eight supervised machine learning models developed in the study

Rank	Model	MAE (‘*ClusAvg’*)	R^2 (‘*ClusAvg’*)	MAE (‘*ClusRep’*)	R^2 (‘*ClusRep’*)
1–2	XGBoost	8.560 ± 0.534	0.631 ± 0.084	10.231 ± 0.461	0.716 ± 0.017
1–2	CatBoost	8.305 ± 0.497	0.707 ± 0.034	10.497 ± 0.336	0.705 ± 0.012
3	LightGBM	9.972 ± 0.696	0.608 ± 0.063	12.316 ± 0.392	0.620 ± 0.022
4	Random Forest	10.243 ± 0.740	0.579 ± 0.066	13.015 ± 0.538	0.570 ± 0.029
5	Gradient Boost	10.561 ± 0.792	0.573 ± 0.090	14.918 ± 0.475	0.470 ± 0.021
6	Bayesian Ridge	13.313 ± 0.655	0.394 ± 0.129	18.905 ± 0.821	0.180 ± 0.036
7–8	ElasticNet	14.847 ± 0.959	0.300 ± 0.036	19.981 ± 0.685	0.120 ± 0.011
7–8	Support Vector	16.386 ± 1.696	0.127 ± 0.037	19.431 ± 0.811	0.160 ± 0.019

Table 8. Predictions for the reference set of attributes, and those evaluated after altering one predictor each instance.

Persona	Step	Aurora	Cora	Dongmei	Johannes	Mahendra
VAS	0	49.73	49.29	54.3	55.74	49.95
Action: changing ‘caffeine consumption’ from reference (10.5) down to 0.5						
VAS	1	54.89	54.45	59.4	61.45	55.12
Action: *[Previous attributes]* + Changing ‘fatigue’ from reference (3) to 6						
VAS	2	38.52	37.36	42.85	44.48	42.93
Action: *[Previous attributes]* + Changing ‘Tense’ from reference (2) to 5						
VAS	3	36.15	35.45	40.91	42.67	37.42
Action: *[Previous attributes]* + Changing ‘Sad’ from reference (2) to 4						
VAS	4	35.61	34.99	40.41	42.14	36.88
Action: *[Previous attributes]* + Changing Scale from ‘Disc’ to 'Tme'						
VAS	5	40.96	39.54	40.85	42.78	39.8

Table 9. The state of lighting variables optimized to match a target VAS on the discomfort scale at a given time of day, for a defined set of personal and contextual factors.

#	Variable	Optimized state	Possible range (from dataset)
1.	Luminance of source	9084.25	1627 to 21700
2.	Tcp of source	4420.75	2609 to 6304
3.	ΔUV	0.01	-0.0001 to 0.0141
4.	Vertical Illuminance	388.12	336 to 485
5.	CIE 1931 x coord	0.39	0.3159 to 0.4711
6.	CIE 1931 y coord	0.38	0.3545 to 0.4116
7.	CIE 1976 u coord	0.23	0.1908 to 0.2699
8.	CIE 1976 v coord	0.51	0.4818 to 0.5292
9.	s-cone EDI	213.77	85.57 to 400.46
10.	m-cone EDI	347.57	262.93 to 486.99
11.	l-cone EDI	391.03	333.96 to 494.30
12.	Rhodopic EDI	298.97	174.65 to 492.19
13.	Melanopic EDI	281.92	145.86 to 491.43
14.	Circadian Lighting	432.65	291.33 to 655.21
15.	Circadian Stimulus	0.38	0.31 to 0.46

Supervised Machine Learning Methods for the Prediction of Discomfort due to Glare

Marshal Maskarenj PhD[a]* and Sergio Altomonte PhD[a]

[a]*Architecture et Climat, Louvain Research Institute for Landscape, Architecture, Built Environment, Université catholique de Louvain, Louvain-la-Neuve, Belgium*

*maskarenj.marshal@uclouvain.be

Batiment Vinci a.017, L5.05.04,

Place du Levant 1, 1348 Louvain-la-Neuve, BELGIUM

Acknowledgements:

The authors acknowledge Dr Michael Kent for responding to queries on setup design.

Supervised Machine Learning Methods for the Prediction of Discomfort due to Glare

Supervised Machine Learning (ML) is emerging as a powerful tool for handling complex multivariate datasets and capturing non-linear relationships between predictors and outcomes. Unlike traditional statistical techniques that only enable to identify predefined functional relationships, supervised ML methods can robustly detect intricate non-linear interactions without explicit assumptions. The capacity for integrating diverse predictors makes these tools particularly suitable for challenges where traditional modelling show limitations. Prediction of discomfort due to glare is one such case, where current models often fail to explain variance in data. These models mostly focus on objective factors related to the scene, while neglecting subjective factors that can vary between- and within-users. These limitations contribute to significant residual variance between model-predicted and reported discomfort scores, which cannot be explained without accounting for inter- and intra-individual differences. To improve these predictions, this paper proposes a method based on supervised machine learning techniques where multiple predictors and their non-linear relationships are integrated towards improving model accuracy and robustness. Via preprocessing, features of a dataset collected from a lighting experiment were refined, and prediction models were developed using eight state-of-the-art ML algorithms. Through inverse regression, the best performing model was further deployed for optimising the characteristics of delivered lighting, ensuring that user perception was maintained within a target zone when all other variables were constrained. Our method demonstrates a practical approach for data handling, multivariate non-linear modelling, and user-centred optimization of delivered lighting, with potential application for problems beyond prediction of discomfort due to glare.

Keywords: supervised machine learning, discomfort glare, subjective factors, inverse regression, user-centric lighting

1. Introduction

Machine learning (ML) techniques have proven more efficient for treating large amounts of data, and solving complex multidimensional problems, compared to traditional methods for performing statistical analysis [1, 2]. ML algorithms are now enabling systems to learn from data, and to improve predictions without being explicitly programmed[3]. These algorithms can broadly be classified under four main categories: *Supervised* ML, *Semi-supervised* ML, *Unsupervised* ML, and *Reinforcement* ML [4]. Supervised ML algorithms learn to make predictions from labelled data, and are typically employed for regression and classification problems – where regression refers to predicting continuous outcomes, while classification includes categorising data into discrete classes [5]. Unsupervised ML algorithms do not need labelled data, and typically address problems such as clustering, association, or pattern detection [6, 7]. Reinforcement ML is yet another technique, where an agent learns to make decisions by interacting with an environment, and receives feedback either as rewards or as penalties [8]. This technique is used for designing strategies that maximize the cumulative reward over time [9, 10]. Each of these ML techniques offer specific advantages for data analysis, and the technique selection depends both on the application as well as on the data structure, i.e. whether the training data is labelled (supervised), unlabelled (unsupervised), or a mix of both (semi-supervised) [11]. Within supervised ML, predictors and outcomes are treated as input-output pairs, and the dataset is split into a subset used for training, and another for testing and internal validation. During the training process, the algorithm iteratively adapts itself to minimize errors in its prediction, and the model built on the dataset's training split can be used for making predictions on the test split, as well as on new datasets having the same feature configuration. Applications of supervised ML within the sciences of the built environment have been reported in the literature to address various challenges, such as for energy use modelling [6], for predicting human factors of subjective comfort [12, 13] and

satisfaction [14], and for controlling the transparency of dynamic glazing, which is a classification problem of supervised ML [15].

In lighting research, many challenges have not yet been fully resolved, among which is the complete characterisation and prediction of discomfort due to glare. Discomfort glare refers to the psychological sensation which occurs when bright light in our field-of-view causes annoyance or disturbance, without necessarily impairing our visual ability [16]. Several models have been proposed to describe subjective perception of discomfort due to glare [17-22]. In this context, Carlucci et al. [23] comprehensively reviewed 34 indices of visual comfort, subdividing them into those associated with *(i)* the amount of light, *(ii)* the quality of light, *(iii)* the distribution of light, and *(iv)* those associated with discomfort glare. It is evident from this review that the last category has expanded over the years. One potential reason justifying this increase, despite an existing abundance of prediction models, is that discomfort glare is still not fully understood, and there are large variances between model-predicted and subjective responses [24]. Although the sources of this variance cannot be entirely explained [25], a large proportion has been attributed to factors related either to the personal characteristics of the users [26], to the underlying contexts from where data were collected [27], or to the data collection procedures themselves [28]. Additionally, factors related to the lighting scene must also be considered as additional sources of uncertainty in field studies (such as those involving daylight), particularly in cases where the ranges of exposure cannot be tightly controlled.

Existing discomfort glare models only account for objective factors related to the scene, uniquely performing correlations that are either linear, quadratic, or exponential. Since the outcome can simultaneously be influenced by multiple predictors having a subjective nature, these limitations might disregard any underlying relationship between inter/intra-individual and/or contextual factors, significantly reducing the model's capacity to explain the variance in responses. Considering the limitations in traditional statistical techniques used by current

models, it is necessary to explore more robust methods for analysing data generated from lighting experiments, and for modelling discomfort due to glare.

For addressing this challenge, the data-driven approach of supervised ML can prove particularly effective, and can significantly improve predictions especially in cases where subjective or contextual factors dominate the variance in a dataset. The ability of supervised ML for performing multivariate non-linear regressions, while incorporating multiple predictors, offers significant potential for improving the accuracy and robustness of discomfort glare prediction models. Towards this aim, this paper wishes to demonstrate the use of data-driven supervised machine learning tools for analysing a multi-feature dataset from a lighting experiment, and for training a robust ML-based model for predicting discomfort glare. The various steps taken for treating and analysing the data, as well as for training and deploying the model for user-centred lighting application, can guide further investigations as well as encourage a wider application of this protocol, for example in the domain of outdoor lighting.

Compared to prior studies that have applied machine learning to visual discomfort, glare prediction, or similar applications, our study introduces various new contributions. Prior models, such as those by Nourkojouri et al. [29] and Li et al. [30], have relied on simulated lighting environments and objective glare indices, without incorporating human responses related to subjective perception (or discomfort evaluation). Other real-world studies, including Tzouvaras et al. [31] and Wagdy et al. [32], have either not included subjective ratings, or have reduced user feedback to binary responses – hence limiting their granularity and capacity for delivering personalized lighting recommendations. One of the novelties of our study consists in training models based on psychophysical data collected from human participants using continuous visual analog scales. The inclusion of multiple contextual and personal predictors – such as luminous history, chronotype, fatigue, mood, caffeine intake, time of day, etc.– allows us to model both inter- and intra-individual differences in discomfort glare perception – an

aspect that has not yet been consistently addressed in the literature. Additionally, we compare eight supervised ML regressors within a pipeline, together with 5-fold cross-validated performance evaluation, analysis of stability in feature importance, and an optimization routine that allows our model optimization framework to recommend personal lighting conditions for specific user personas and profiles. These elements collectively aim to move away from generic prediction of discomfort due to glare, toward a modelling approach that is grounded in psychophysical data, and is specific to the user and its context.

2. Methods

The following sections provide: *(i)* a brief description of the procedure used for the collection of the data used in this paper; *(ii)* an illustration of the variables featured in the dataset; and, *(iii)* a presentation of the process used for data treatment and analysis. A more detailed description of the testing apparatus and of the experimental procedure is part of a further ongoing publication by the authors[33, 34].

2.1 Data collection procedure

The subjective data used in this paper was collected from a lighting experiment conducted in Belgium during the spring months of April to June, which aimed to characterize the effect of contextual factors on the evaluation of discomfort due to glare. These factors included time of day and the spectrum characteristics of the glare source.

A chamber was built to generate pre-set electric lighting scenes that were designed to correspond to pre-calibrated thresholds of discomfort due to glare. A total of 38 participants took part in this controlled experiment. The participants reported their subjective discomfort sensation on two visual analogue scales (VAS, provided in the Appendix A1) over three

different times of day when being exposed to the glare source. Participants maintained a fixed head position defined by a chin- and forehead-rest [35], and faced a fixation point in their field of view while being exposed to the lighting scenes. Appendix A2 presents the experimental apparatus within the test chamber, and provides a detailed timeline of the experiment

2.2 Variables

2.2.1 Variables related to the scene

The scenes were calibrated using the IES-GI model [21] to stimulate a visual discomfort anchored at the 'just perceptible' (JP), 'just noticeable' (JN), 'just uncomfortable' (JU), and 'just intolerable' (JI) glare sensation vote (GSV) [16]. The four GSV-anchored scenes were designed with two spectral states (herein, for simplicity, generally referred to as their correlated colour temperature, CCT, that is 'red' and 'blue'), and the eight scenes were maintained using commercially available Wiz A67 E27 smart lamps. This study used a category rating procedure [36] where the participants marked their perception on visual analogue scales against every predefined state of lighting. Scenes were toggled between these predefined states using the Wiz application. The characteristics of these eight lighting scenes are presented in Table 1.

[TABLE 1]

For each scene, the Correlated Colour Temperature (Tcp), vertical illuminance (Vert-illum), and the CIE chromaticity coordinates presented in Table 1 were measured using a Konica Minolta CL-70F spectrometer, where the device's bulb was positioned to approximate the eye's position when the head is supported by the chin- and forehead-rest. The 1-nm separated spectral irradiances (between 380-780 nm range) measured by the spectrometer were then used to calculate the α-opic EDIs and the circadian metrics for each scene, respectively using the CIE-s026 [37, 38] and LRC's calculators [39].

Table 1. Photometric and radiometric measurements for the eight scenes measured at eye level using a Konika Minolta CL70F spectrometer (1nm resolution between 380-780nm). The CIE S026 a-opic metrics are derived from spectral data using the CIEs026 calculator, and the circadian metrics are derived using the LRC calculator.

					CIE 1931		CIE 1976		EDI based on CIE s-026 calculator						
CCT indicator	**GSV**	**Lum (cd/m^2)**	**Tcp (K)**	**Vert Illum (lx)**	**x-coord**	**y-coord**	**u-coord**	**v-coord**	**s-cone (lx, EDI_S)**	**m-cone (lx, EDI_M)**	**l-cone (lx, EDI_L)**	**Rods (lx, EDI_R)**	**melan. (lx, EDI_{mel})**	**Circad Lighting (lx, CL_A)**	**Circad Stim.**
Red	JP	1627	2645	336	0.4691	0.4116	0.2680	0.5292	85.57	262.93	342.50	174.66	145.86	291.33	0.3116
	JN	3860	2638	345	0.4693	0.4110	0.2684	0.5289	88.51	269.94	352.09	179.02	149.27	297.94	0.3159
	JU	9150	2620	394	0.4708	0.4112	0.2693	0.5292	99.62	306.96	401.71	201.87	167.12	333.64	0.3376
	JI	21700	2609	485	0.4711	0.4104	0.2699	0.5290	123.50	376.89	494.30	245.31	200.97	401.23	0.3732
		Mean	**2628**	**390**	**0.4701**	**0.4111**	**0.269**	**0.529**	**99.3**	**304.2**	**397.6**	**200.2**	**165.8**	**331**	**0.335**
		StDev	**16.5**	**68.26**	**0.001**	**0.0005**	**9E-04**	**1E-04**	**17.23**	**52.18**	**69.46**	**32.35**	**25.23**	**50.35**	**0.028**
Blue	JP	1627	6304	336	0.3159	0.3545	0.1908	0.4818	290.57	340.84	333.97	347.79	348.18	469.75	0.4032
	JN	3860	6271	343	0.3166	0.3551	0.1910	0.4822	295.25	348.13	341.47	355.41	356.16	479.01	0.4069
	JU	9150	6210	383	0.3179	0.3558	0.1916	0.4827	326.68	387.90	381.37	395.50	396.40	533.04	0.4268
	JI	21700	6069	483	0.3209	0.3588	0.1926	0.4846	400.46	487.00	480.85	492.20	491.43	655.22	0.4636
		Mean	**6214**	**386.2**	**0.3178**	**0.3561**	**0.192**	**0.483**	**328.2**	**391**	**384.4**	**397.7**	**398**	**534.3**	**0.425**
		StDev	**104**	**67.74**	**0.0022**	**0.0019**	**8E-04**	**0.001**	**50.75**	**67.28**	**67.57**	**66.37**	**65.74**	**85.34**	**0.028**

2.2.2 Variables related to the observer (Personal Factors)

This study collected 13 variables associated with personal factors [28] . For each participant, questionnaire items asking data on chronotype [40], ethnicity, eye conditions, etc. were administered during the introduction sessions. The questionnaire used to evaluate personal factors is provided in the Appendix A3, and descriptive statistics of personal factors are presented in Table 2.

[TABLE 2]

Table 2. Descriptive statistics on data related to personal factors (N = 38 subjects)

Gender	**Age**	**Ethnicity**	**Eye conditions**	**Chronotype**	**Photosensitivity**
52.6% *Male* 47.3% *Female*	Range: 22-64 Med: 32 34.4 ± 11.6	***White***: 73.6% ***Asian***: 21% ***Mixed***: 5.3%	42% wore lenses 14% reported other vision issues	7-point scale, 2.95 ± 1.4 Median: 3	**Natural Light**: 68.4% **Artificial Light**: 60.6% **Sunglasses**: 39.4% **Blind preference**: 1.15±0.85 on the 0-4 scale **Lighting preferred**: 0.73±0.75 on a 0-3 scale
Caffeine Habit					
Yes: 67.5%					

It must be noted that photosensitivity was collected as five separate items, while eye-conditions included three separate items. Among the 38 participants, five indicated “Yes” to the item “any other vision problem”, which was one of the items connected to their eye-conditions. Their reported descriptions in the follow-up included: previous use of low corrective lenses (n = 1), mild short-sightedness and presbyopia (n = 1), strabismus (n = 1), and two unspecified responses. None of the participants reported any diagnosis of conditions associated with ocular pathology – such as cataract, glaucoma, or corneal disorder, that could have significantly influenced their perception of discomfort due to glare.

2.2.3 Variables related to the context (Contextual Factors)

Prior to the scene exposure within each session, participants were asked to respond to 14 items

associated with contextual factors[28]. These are the factors that can vary within the participant across sessions, and could lead to intra-individual differences in responses to perceived discomfort. These factors were collected through established scales, including fatigue [41], sleepiness [42], mood indicators [43], etc. Two instruments were employed to collect sleepiness – the Karolinska Sleepiness Scale (KSS) and the "sleepy/awake" items within the mood indicators, both of which were administered after the scene evaluations had been completed. The KSS asked participants to circle the number that best described their current perceived state of sleepiness on a 9-point scale, hence asking test subjects to report their self-evaluated level of physiological sleepiness/alertness. Conversely, the "sleepy" and "awake" mood items, collected on 5-point scales together with other dichotomous affective descriptors (including, tense/calm, sad/happy, lively/drowsy), were included to deepen the appraisal of the participant's collective emotional states. Their simultaneous administration aimed at furthering the characterisation of the perceived level of sleepiness reported by test subjects. Descriptive statistics at session level are presented in Table 3, and the questionnaire items used for contextual data collection at each session are provided in the Appendix A4.

[TABLE 3]

Table 3. Descriptive statistics on data related to contextual factors (N=114 sessions)

Caffeine score	**Prior Light Exposure (0-7)**	**Mood Indicators (1-5)**	
21.1 ± 25.5%		***Tense***	1.91 ± 0.99
Continuous VAS	***Direct sunlight:***	***Sad***	1.53 ± 0.8
Fatigue (0-6)	5.1±1.8	***Lively***	3.37 ±0.82
2.14±1.29	***Diffuse daylight***	***Sleepy***	2.24±0.96
Samn-Perelli scale	3.9±1.6	***Calm***	3.7±1.02
Sleepiness (1-9)	***Electric light***	***Happy***	3.6±0.84
4.2±1.7	4.3±1.8	***Drowsy***	2.4±1.05
Karolinska scale		***Awake***	3.53±0.96

2.2.4 Variables related to the experimental session

The dataset also included the session order, i.e. six permutations of Morning 'M', Afternoon

'A', Evening 'E'. These defined the participation sequence for every individual, which was counterbalanced at the participant level. Also included was the order of exposure for the eight defined lighting scenes during each session. The exposure order was counterbalanced at the session level.

2.2.5 Outcome variables

During each exposure within every session, participants looked at the fixation point for the entire duration of the exposure, after which they reported the level of discomfort that they felt due to glare from the lighting scene. The responses were collected on paper-based continuous visual analogue scales (VAS), where the participants placed a cross mark corresponding to their sensation. The responses were digitized by measuring the distance of the cross mark from the left end of the VAS. The VAS used for this study have been presented in the appendix A1.

2.2.6 Potential experimental biases

Experimental biases were minimized by following recommendations for good practice, such as counterbalancing session order (i.e., time of day for the participants), counterbalancing the order of displaying the eight scenes within each session (JP/Red, JP/Blue, … JI/Blue), blindfolding participants between scene changes, etc. [44, 45].

2.3 Outline of data treatment

Three distinct steps were taken for treating the data, including feature optimization (Figure 1a), supervised Machine Learning (Figure 1b), and personalized recommendation via inverse regression (Figure 1c).

[FIGURE 1]

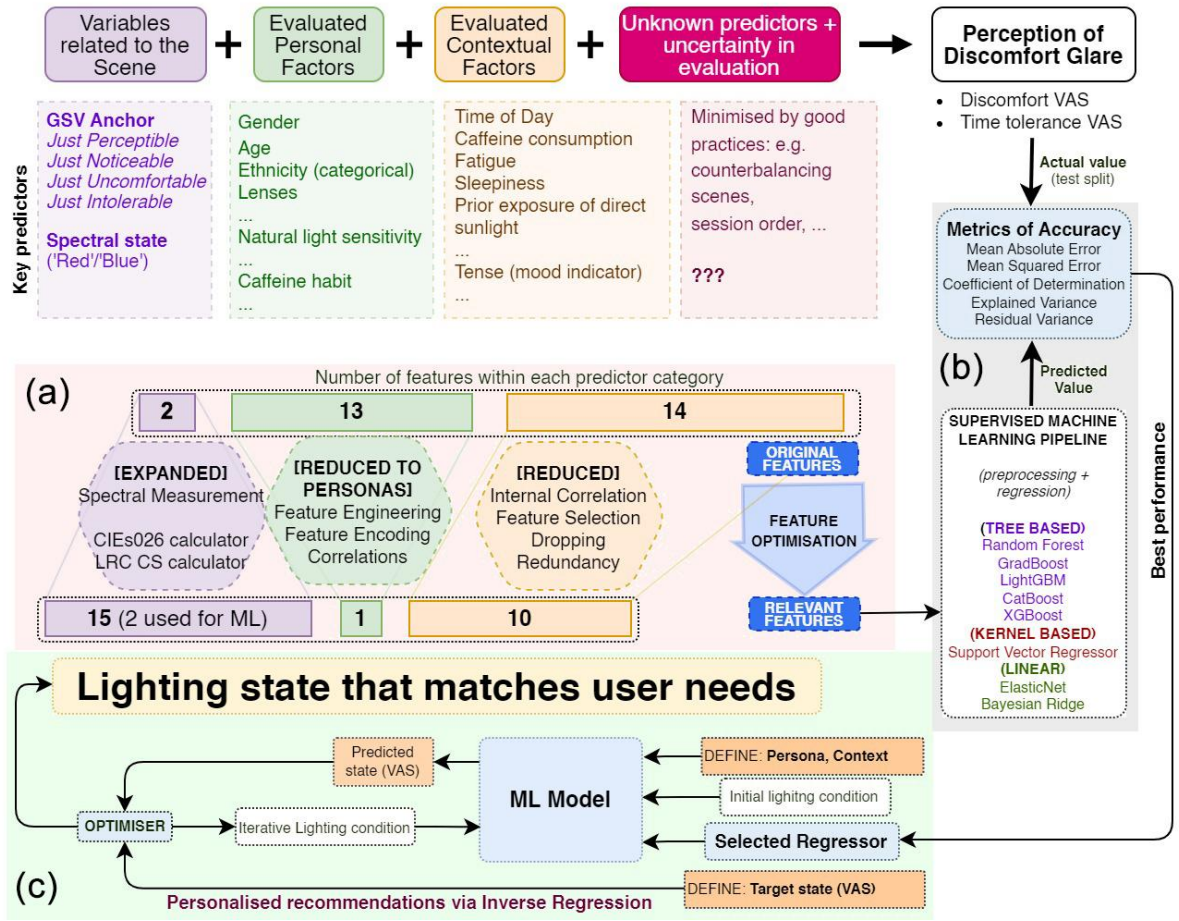

Figure 1. Data treatment in three distinct steps: (a) feature optimization, (b) supervised machine learning, (c) Inverse regression and state optimization.

The feature optimization step aimed to reduce dimensionality, since high number of features can lead to model overfitting and a reduced generalization for external datasets. In this step, the original features were transformed into relevant counterparts, after which regression models were developed using eight supervised machine learning libraries readily available in Python. The modelling method that demonstrated the highest performance in prediction (or accuracy with respect to the actual data) was identified and used to create a portable regression model. This model – which was a reusable serialised binary file of pickle *.pkl format that encapsulated the model's structure and learned parameters – was investigated and deployed for optimizing the state of lighting for any user and contextual factors, when the target zone of

(dis)comfort was specified. Effectively, the original experiment-generated dataset was transformed to train a supervised machine learning model, which could optimize the state of lighting variables – specifically the luminance and the spectral characteristics – and deliver the target zone of perceived (dis)comfort for a defined user against constrained contextual settings.

3. Results

3.1 Multi-Feature Dataset

The original dataset included 37 columns across 1824 rows, with each row corresponding to the VAS responses provided by 38 individual participants under 48 data-collection scenarios (four GSV-anchors across two spectral states, on two types of VAS scales, at three times of day). The key predictors related to the exposure were: the four distinct GSV anchor for luminance (Just Perceptible 'JP', Just Noticeable 'JN', Just Uncomfortable 'JU', Just Intolerable 'JI'), and the two spectral states of the scenes ('Red', 'Blue'). Additionally, three times of day (Morning 'M', Afternoon 'A', 'Evening 'E'), and the two VAS scales used for data collection ('discomfort scale', 'time tolerance scale') were also key objective predictors. Also included in the dataset were 27 subjective predictors related to personal factors (13) and to contextual factors (14). Against these predictors, the responses measured on the VAS scales served as the outcome variable. Effectively, the dataset contained 29 features as predictors, and one as the outcome. The outcome variable (VAS) refers to the subjective rating given by participants on one of two visual analog scales: a discomfort glare scale ('Disc') and a time-tolerance scale ('Tme'). Each participant provided ratings on both scales against every lighting condition. The type of scale that is scored in the outcome variable VAS is indicated in the 'Scale' column, which is included as a categorical predictor in all modelling processes. This is done to help the model account for potential differences in interpretation between discomfort

and time-based responses. All models are trained to predict the continuous VAS outcome, irrespective of the scale type. The remaining columns contained information that was not directly relevant for the analysis presented in this paper (e.g., participant ID, etc.). A complete list of predictor variables used in the modelling pipeline, along with their descriptions, is provided in Appendix A5 Table A5.1, which serves as a reference for all variables used in the ML modelling tasks.

3.2 Feature Optimization

For training adequate models while minimising the risk of overfitting, this first step (described in Figure 1a) reduced the dataset's complexity while retaining the relevant information. Here, overfitting refers to a common issue in supervised machine learning that prevents model generalization. A high number of features against a small dataset results in the model being trained not only on patterns in the data, but also on inherent noise [46], which must be avoided by cleaning and preprocessing the dataset, as well as by dropping redundant features. The following steps were taken to optimize the features on the predictor side.

3.2.1 Feature engineering

In this step, eight separate predictors within 'personal factors' were reduced to two via K-Means clustering and rule-based selection, respectively, which significantly reduced the dimensionality. K-Means clustering [47, 48] is an unsupervised machine learning algorithm that does not require data to be labelled, and which partitions the dataset into 'K' number of clusters (where K must be defined). The distance between data points and their corresponding cluster centroid is minimized, and each point is assigned to the nearest centroid, with centroids iteratively recalculated until convergence is reached [49]. Via K-means, five criteria nested within sensitivity to light (i.e. *sensitivity to natural light, to artificial light, affinity for using sunglasses,* and *the preferences for blind states and lighting conditions*) were reduced to a

single feature of ‘photosensitivity’, which had three possible states: High, Moderate, or Low.

For clustering process, the five individual features were used as inputs, along with the mean VAS score for each participant across the 48 scenarios. Python Scikit-learn’s ‘OneHotEncoder’ and ‘StandardScaler’ functions [50] were used to encode the dichotomous features, and to standardize the categorical features using minimum-maximum scaling, respectively. By defining ‘K’ as 3, participants were clustered into three distinct groups, which were then inferred based on feature values as having different levels of photosensitivity. The use of K-Means reduced the risk of subjective bias in manual clustering (which could have used ideal centroids of minimum, mid-point, and maximum thresholds for each feature). The use of this unsupervised machine learning technique retained the complex interactions between the individual nested features.

The three clusters generated by K-Means clustering were designated as clusters ‘0’ (11 participants), ‘1’ (17 participants), and ‘2’ (10 participants), and are summarized in Table 4 together with cluster-wise statistics, including ranges and frequency distributions. Figure 2 presents the distribution of participants within each cluster, which is visualized separately for the dichotomous (top row) and for the continuous (bottom row) predictors for each group.

[FIGURE 2]

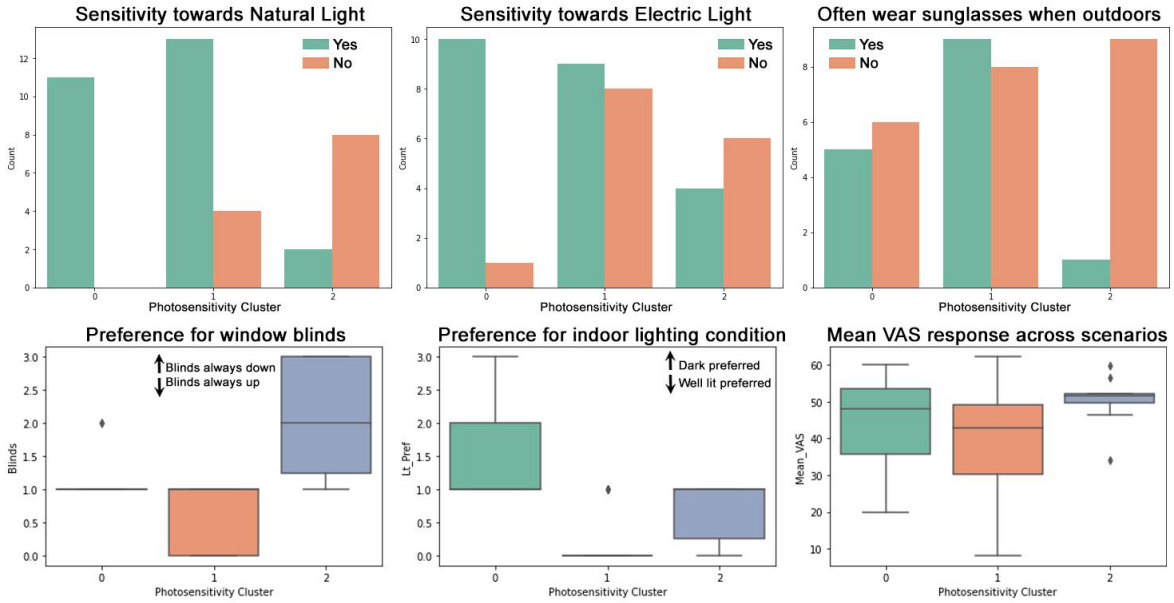

Figure 2. Distribution of participants within each of the 3 clusters for the dichotomous parameters (top row), and for the categorical variables (bottom row).

[TABLE 4]

Table 4. Frequencies and ranges of data distribution for participant attributes between clusters. The last column includes the interpreted photosensitivity, and the footnote describes the meaning of the anchors and values.

	Range			**Frequencies**			**Photosensitivity**
Cluster	**'Blinds'**	**'Lt_Pref'**	**'Mean_VAS'**	**'Nat_Lt'**	**'Art_Lt'**	**'Sunglss'**	**(Interpretation)**
0	[1, 2]	[1, 3]	[19.77, 60.13]	0: 100%, 1 :0%	0: 90.9%, 1: 9.09%	1: 54.5%, 0: 45.4%	'High'
1	[0, 1]	[0, 1]	[8.16, 62.18]	0: 76.47%, 1: 23.52%	0: 52.9%, 1: 47.05%	0: 52.9%, 1: 47.05%	'Moderate'
2	[1, 3]	[0, 1]	[34.22, 59.88]	0: 20%, 1: 80%	0: 40%, 1: 60%	0: 10%, 1: 90%	'Low'

Feature description and interpretation:
Use of Blinds ('Blinds'): 0-4, with 4= blinds always down, 0 = blinds always up
Preference for lighting condition ('Lt_Pref'): 0-3, with 3 = dark preferred, 0 = well lit preferred
Sensitive to Natural Light ('Nat_Lt'): 0 is Yes (i.e. sensitive), 1 is Not sensitive.
Sensitive to Artificial Light ('Art_Lt'): 0 is Yes, 1 is Not sensitive.
Use of Sunglasses ('Sunglss'): 0 is Yes (meaning sensitive), 1 is No (could mean not sensitive).
Mean reported discomfort due to glare ('Mean_VAS): Higher discomfort indicates higher photosensitivity.

From the data presented in Table 4 and Figure 2, it is evident that all participants within the first cluster were sensitive to natural light, and about 90% to artificial light. This group also had an indoor lighting preference that tended towards 'dark preferred', and was therefore interpreted as having a high photosensitivity. Conversely, the third cluster had only 20% and

40% participants sensitive to natural and artificial light, respectively, which was lower than the other two clusters, in ordered sequence. The light preference for this group was towards 'well lit' – therefore this group was evaluated as the least photosensitive. The preference for wearing sunglasses also showed tendencies in the same direction, with a reducing proportion of each group indicating 'yes' along with their interpreted photosensitivity, i.e. the high photosensitive group also had the highest proportion of responses towards 'yes' (on frequent use of sunglasses). However, this aspect could be confounded by cultural habits, since the use of sunglasses is not common in Belgium. Effectively, 11, 17, and 10 subjects were categorized as having a 'high', 'moderate', and 'low' photosensitivity, respectively.

In parallel, the three features nested within eye-condition (*whether they wore corrective lenses, had functional colour vision, and reported any other vision problem*) were reduced into a singular feature via rule-based classification [51], which generated three categories of eye condition: 'poor', 'moderate', and 'good. Together, the steps within feature engineering allowed us to combine eight disparate predictors into two engineered features, while augmenting the quality of these composite features.

3.2.2 Participant profiling through feature encoding

By creating two composite features, the total number of features within the 'personal factors' category (Figure 1) was substantially reduced. This resulted in seven key features that together described the participants' personal characteristics: *age, gender, ethnicity, eye condition, chronotype, photosensitivity,* and *caffeine habits*. Among these, eye condition and photosensitivity are the composite features that were created in the feature engineering step. Chronotype refers to a biological construct that defines an individual's internal circadian rhythm as well as their natural preference for sleep and wake times, and provides an indication of when they feel most alert or tired during a 24-hour period. Three defined categories have

been used: Morning-types (M-types), Evening-types (E-types), and Neither-types (N-types), where the third category classifies individuals (about 60% of the population) that do not have clear circadian preferences, and who show intermediate features between M-types and E-types [52, 53].

With each feature having 2 to 3 possible categories, there were 972 theoretical combinations that could be identified, and participants were profiled by encoding these features. Distinct profiles were created, with a superscript above each attribute – such as photosensitivity, or gender – that defines the state of each encoded attribute. For example: *a young^{0} woman^{1} of European^{0} ethnicity with good eye condition^{2}, N-chronotype^{1}, high photosensitivity^{0}, and no caffeine habit^{1}* was designated as 'P0102101'; or, *a middle-aged^{1} man^{0} of Asian^{2} ethnicity with moderate eye condition^{1}, M-chronotype^{0}, and moderate photosensitivity^{1}, who regularly consumes caffeine^{0}* was designated as 'P1021010'. Out of the 972 theoretical profiles based on combinations of categorized features, further analysis revealed that our participants could be clustered within 35 distinct combinations. The parallel coordinate plot in Figure 3 presents the theoretical combinations, as well as those for the profiles found in our participants. The thin dark lines in the background represent 96.3% of the combinations that were not found (for example, participants having a mixed ethnicity and poor eye-condition), while the profiles of participants found in our dataset are represented using dominant lines. The encoded profiles are further connected to mean VAS scores, that were reported by each profile across the 48 scenarios.

[FIGURE 3]

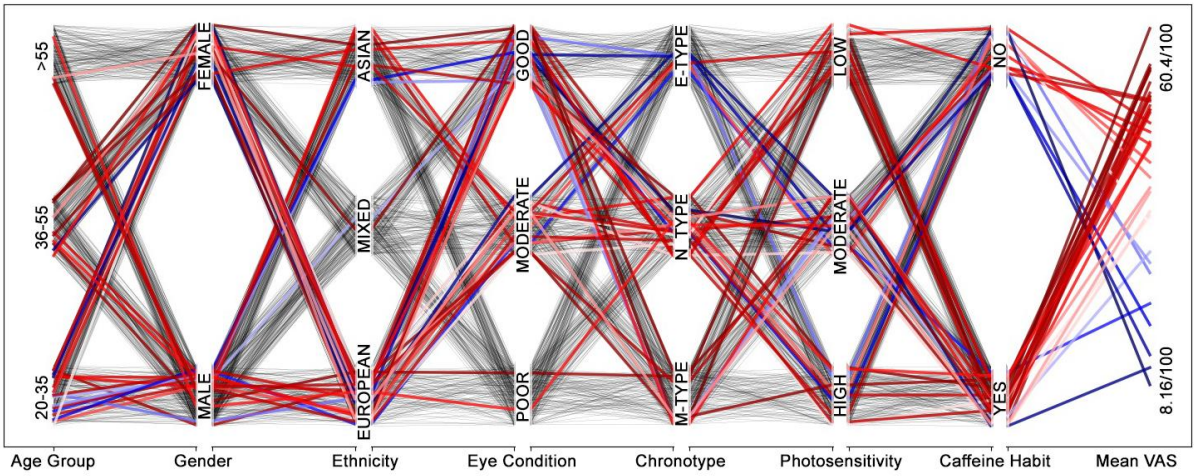

Figure 3. Parallel coordinate plot demonstrating the combinations of encoded features of personal factors based on the defined categorization. Each combination represents a profile, with all theoretical and actual profiles represented by dark connectors and dominant connectors, respectively.

3.2.3 Generating user personas

The main ambition of the feature optimization step was to aggregate personal factors into synthetic user 'personas'. Personas have already been used in the lighting literature to represent typical users, as well as to highlight their luminous needs and responses to support a user-centred design [54]. Prior steps within feature optimization transformed the raw features into meaningful composites, and unique combinations of attributes within each meaningful feature were represented as 35 encoded profiles. Through this step, the profiles were further clustered based on their similarity of response, and personas were generated as higher order representation of participants.

To achieve this clustering, the 35 distinct profiles were analysed via the Spearman rank correlation method, based on their reported discomfort due to glare across the 48 different lighting scenarios. This method performs a non-parametric measure of relationship between two variables, and describes how consistently any variable responds (i.e., increases or decreases) with respect to change in the other. This method was preferred over the Pearson

correlation method that assesses linear relationships, since Spearman rank correlation can be applied to data that may not have a normal distribution [55]. The '*spearmanr'* function in stats library of Python SciPy was used for performing this function, and the resulting matrix from pairwise correlation between groups is presented in Figure 4. Encoded profiles are presented on the two axes, where the seven sequential indices (such as '2101010') represent the individual features of personal factors within each profile.

[FIGURE 4]

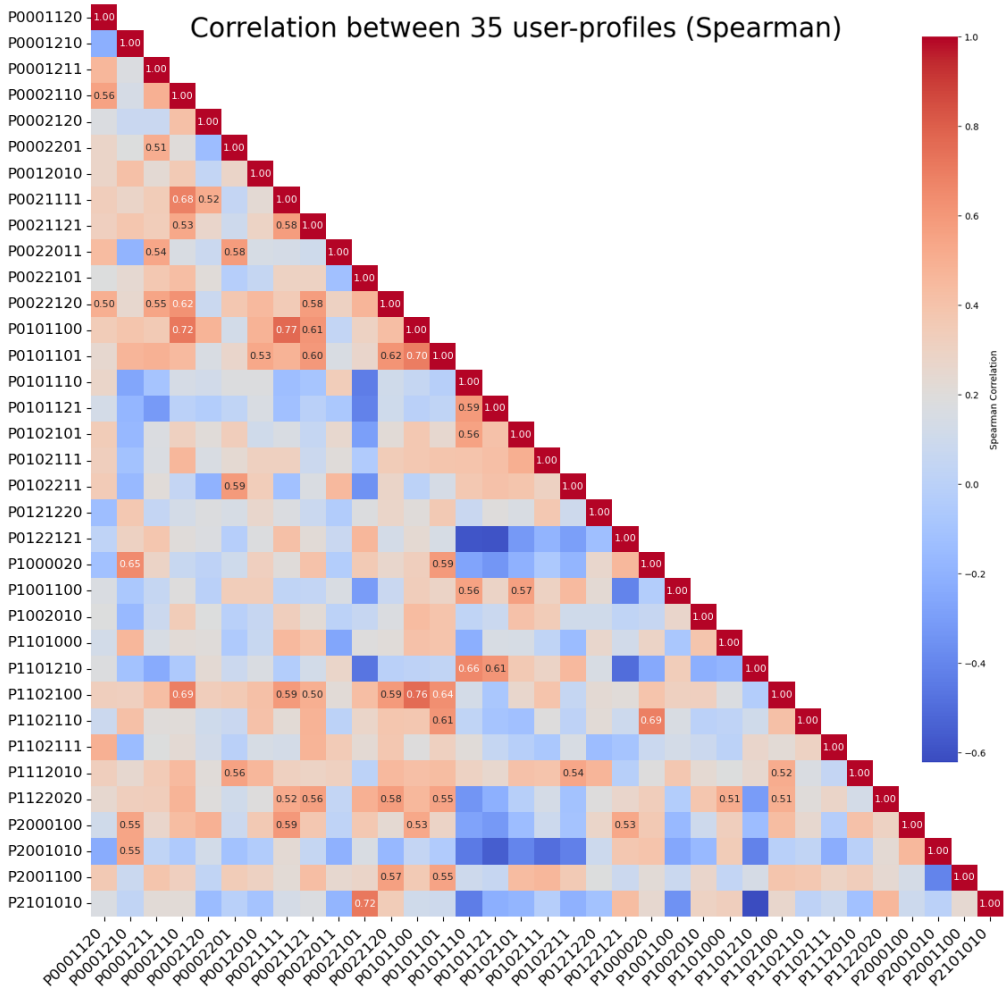

Figure 4. Correlations between profiles of attributes, drawn based on the VAS scores across the 48 scenarios. On both are the 35 profiles of users defined via encoded attributes of personal factors. Threshold for interpretation: >0.5 (Moderate). 53 of the 595 combinations demonstrate correlations above the threshold.

All combinations with moderate or higher correlations (coefficient ρ>0.5, based on [56]) were isolated, and it was observed that 53 of the 595 possible combinations (35*34/2) demonstrated internal correlations above the set threshold. Table 5 presents these combinations, where mutual pairs of profiles are presented via encoded features.

[TABLE 5]

Table 5. The 53 combinations demonstrating correlations (via Spearman method) above the ρ>0.5 threshold. These were part of three distinct sets, as indicated with the overlay colour.

#	Profile A	Profile B	Corr	#	Profile A	Profile B	Corr	#	Profile A	Profile B	Corr
1	P0002110	P0001120	0.561	19	P0101101	P0101100	0.698	37	P1112010	P0002201	0.559
2	P0002201	P0001211	0.507	20	P0101121	P0101110	0.592	38	P1112010	P0102211	0.543
3	P0021111	P0002110	0.684	21	P0102101	P0101110	0.561	39	P1112010	P1102100	0.516
4	P0021111	P0002120	0.518	22	P0102211	P0002201	0.593	40	P1122020	P0021111	0.524
5	P0021121	P0002110	0.532	23	P1000020	P0001210	0.650	41	P1122020	P0021121	0.565
6	P0021121	P0021111	0.578	24	P1000020	P0101101	0.591	42	P1122020	P0022120	0.576
7	P0022011	P0001211	0.539	25	P1001100	P0101110	0.557	43	P1122020	P0101101	0.551
8	P0022011	P0002201	0.583	26	P1001100	P0102101	0.568	44	P1122020	P1101000	0.514
9	P0022120	P0001120	0.501	27	P1101210	P0101110	0.658	45	P1122020	P1102100	0.505
10	P0022120	P0001211	0.546	28	P1101210	P0101121	0.613	46	P2000100	P0001210	0.547
11	P0022120	P0002110	0.625	29	P1102100	P0002110	0.691	47	P2000100	P0021111	0.595
12	P0022120	P0021121	0.581	30	P1102100	P0021111	0.590	48	P2000100	P0101100	0.533
13	P0101100	P0002110	0.719	31	P1102100	P0021121	0.502	49	P2000100	P0122121	0.534
14	P0101100	P0021111	0.772	32	P1102100	P0022120	0.590	50	P2001010	P0001210	0.546
15	P0101100	P0021121	0.613	33	P1102100	P0101100	0.765	51	P2001100	P0022120	0.566
16	P0101101	P0012010	0.531	34	P1102100	P0101101	0.636	52	P2001100	P0101101	0.551
17	P0101101	P0021121	0.599	35	P1102110	P0101101	0.608	53	P2101010	P0022101	0.720
18	P0101101	P0022120	0.617	36	P1102110	P1000020	0.691				
	Set 1				**Set 2**				**Set 3**		

These combinations were inspected via network analysis – a graph-based method – to identify groups that were strongly correlated. For performing this analysis, each profile was represented as a node, and the edges connected those profiles (as nodes) that had a Spearman correlation coefficient higher than the set threshold of 0.5. This method was performed using the '*nx'* function of Python's '*networkx'* library, and allowed to understand how the profiles were interconnected based on the similarity in their responses. The outcome of network analysis is presented in Figure 5, from which it was identified that 31 of 35 profiles were part of three closed networks. These sets of profiles are also highlighted in Table 5. It must be noted

that the remaining four profiles were distinct, and did not demonstrate correlation above the threshold with any other.

[FIGURE 5]

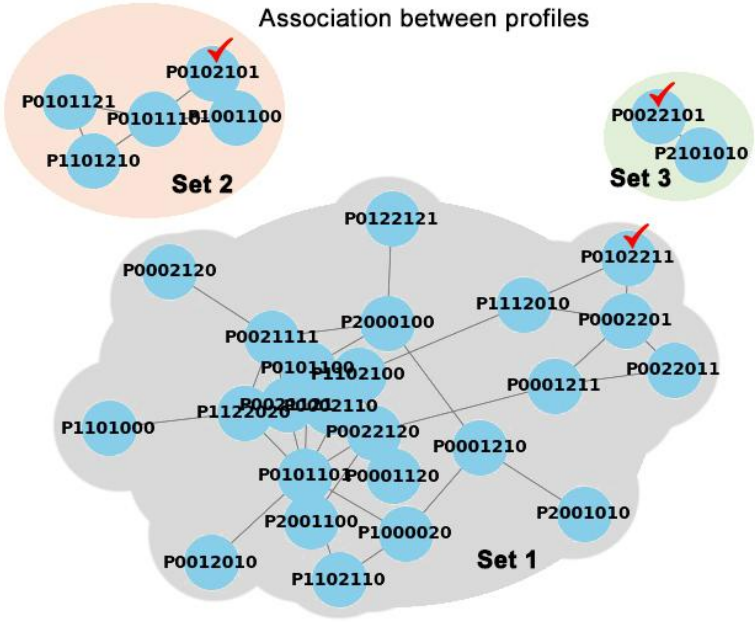

Figure 5. Network analysis for identifying association between profiles.

For each of the three networks indicated in Figure 5, a representative profile was selected (as indicated with a checkmark). By combining these with the other four distinct profiles, seven distinct personas were generated based on the similarity of their perception of lighting scenarios. This step allowed 80% of redundant profiles to be dropped, which enriched the data quality by increasing the number of data points for each included persona. These synthetic personas, each with distinct needs and responses, were designed to represent any user within our dataset, and to understand their responses while developing user-centred design [54]. The personas are presented in Table 6, which includes: the encoded profile of the representative for each persona; the number of profiles within each persona; the semantic attributes; and, an imaginary name arbitrarily assigned to represent each persona.

[TABLE 6]

Table 6. The seven distinct personas developed after feature engineering and encoding. Among these, three were representative of the sets of correlated profiles, while four represent the distinct profiles having no significant correlation with any other profile.

Encoding	n	Attributes	Name
P0022101	2	A young man of Asian ethnicity and a good eye condition, having a N-chronotype and a high photosensitivity, who typically doesn't consume caffeine	Mahendra
P0102101	5	A young woman of European ethnicity and a good eye condition, having a N-chronotype and a high photosensitivity, who typically doesn't consume caffeine	Aurora
P0102211	24	A young woman of European ethnicity and a good eye condition, having a E-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Julia
P0102111	1	A young woman of European ethnicity and a good eye condition, having a N-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Cora
P0121220	1	A young woman of Asian ethnicity and a moderate eye condition, having a E-chronotype and a low photosensitivity, who typically consumes caffeine	Dongmei
P1002010	1	A middle-aged man of European ethnicity and a good eye condition, having a M-chronotype and a high photosensitivity, who typically consumes caffeine	Johannes
P1102111	1	A middle-aged woman of European ethnicity and a good eye condition, having a N-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Elizabeth

3.2.4 Feature selection for contextual factors

Reducing features helps simplify machine learning tasks and, while features within 'personal factors' were significantly reduced (from 13 to one), the dataset also included 14 disparate contextual factors that varied between sessions within each participant. These included self-reported data specific to each session: *fatigue, sleepiness, three indicators of previous luminous exposure, caffeine consumption in the previous two hours, and eight indicators of their mood*, which have been reported to influence subjective perception of discomfort due to glare [26]. The large number of these features could still pose the risk of overfitting, and a feature selection process was conducted to reduce this risk.

An internal correlation was conducted using the Spearman correlation matrix method, and features having mutual correlation coefficient above 0.5 (interpreted as moderate correlation) [56] were filtered. The correlation matrix is presented in Figure 6.

[FIGURE 6]

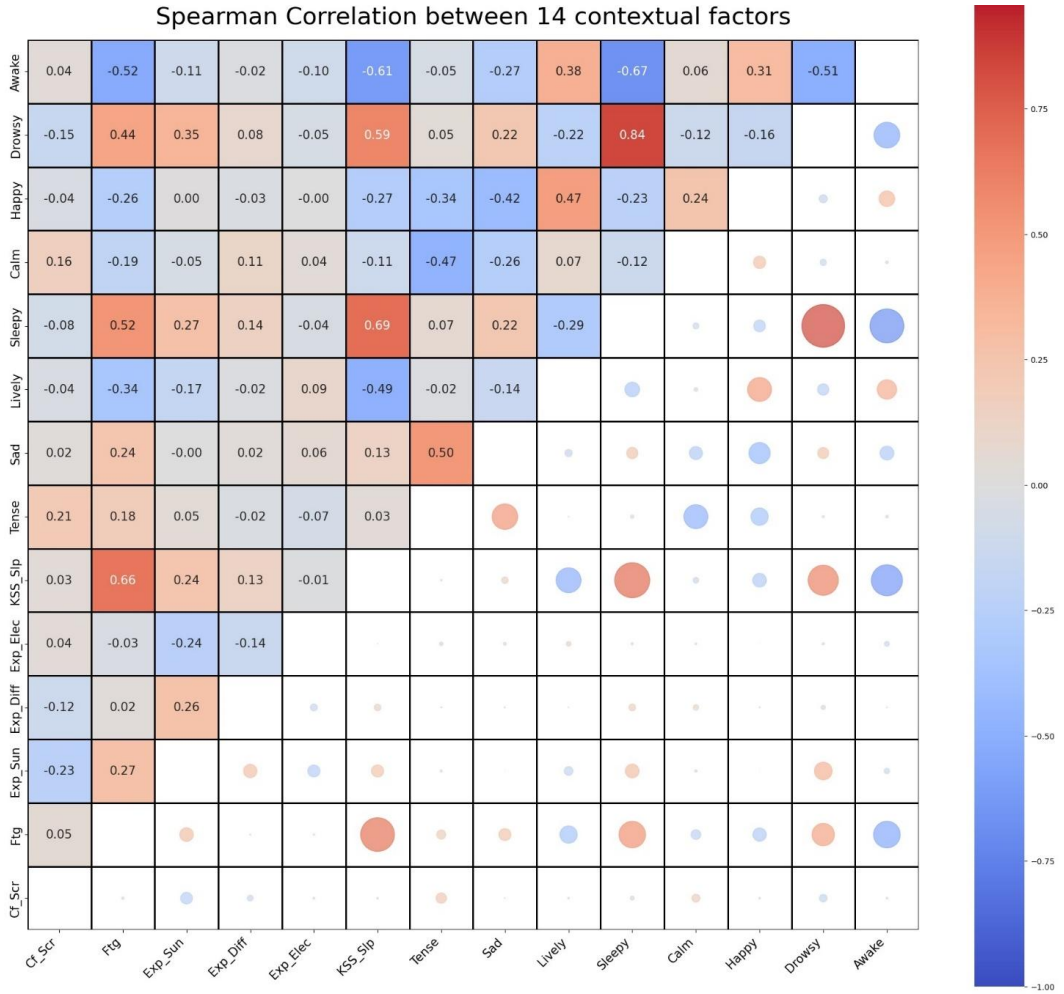

Figure 6. Correlations between the 14 contextual factors that can lead to intra-individual difference in responses. Threshold for interpretation: 0.2-0.5 RMPE (minimum effect size for a practically significant effect), 0.5-0.8 moderate, >0.8 strong.

It was observed that Fatigue (‘*Ftg’*), Sleepiness self-reported through the Karolinska scale (*‘KSS_Slp’*), and Mood indicators of ‘*Sleepy’*, ‘*Awake’*, and ‘*Drowsy’* demonstrated a ‘moderate’ (0.5) to ‘high’ (0.8) correlation. Moderate correlations were observed between the following pairs: *Fatigue/KSS* (0.66), *KSS/Sleepy* (0.69), and inversely between *Sleepy/Awake*(-0.67) and *Awake/KSS* (-0.61). Mood indicators of *Sleepy* and *Drowsy* demonstrated a high correlation, with the coefficient ρ evaluated at 0.84. This process allowed four of these five

features – *Sleepy*, *Drowsy*, *Awake*, and *KSS* – to be dropped, while Fatigue was retained to represent the others yet retaining the diversity in responses. It must be noted that between factors of prior luminous exposure, some correlations at the minimum level representing a "practically" significant effect (RPME) [56] (with a minimum threshold of 0.2) were identified, with direct sunlight/diffuse daylight presenting a correlation at $\rho = 0.26$, direct sunlight/electric light demonstrating inverse correlation at $\rho = -0.24$, while no substantive correlation was identified between diffuse daylight/electric light ($\rho = -0.14$). Beyond dropping the four correlated features, further reduction was not feasible without increasing residual (unexplained) variance. Effectively, this step allowed a reduction in contextual factors from 14 to 10 independent features.

The steps within feature optimization reduced the features of the original dataset while augmenting its quality. The transformed dataset now included seven distinct user personas as a singular feature, with ten contextual factors representing intra-individual variations. This dataset was then used for training supervised machine learning algorithms.

3.3 Supervised Machine Learning

To evaluate model performance, we first established a baseline using a simple linear regression model, using only scene-related variables: luminance of the glare source (GSV_{lum}) and the chromaticity category (CCT: Red/Blue) – to evaluate their capacity for estimating subjective responses to discomfort glare. This baseline provided a reference for comparing more advanced approaches. Following this step, eight state-of-the-art supervised machine-learning algorithms (Figure 1b) were trained on the enriched dataset, to assess their capacity of capturing non-linear relationships, and for improving the accuracy of prediction.

The baseline model was trained on the dataset featuring 35 user profiles (N = 1680), using an 80/20 train-test split, and yielded a mean absolute error (MAE) of 20.51 with an R^2 score of

0.10, indicating limited explanatory power. A log-transformed version of the model further reduced the performance (MAE = 22.08, R^2 = -0.05). These results confirm that linear models with variables associated uniquely with the scene can explain only a small portion of the variance in discomfort glare ratings, hence reinforcing the need to include personal and contextual predictors in advanced models.

In parallel, the mean of the DG scores for subgroups with distinct states of source luminance ('GSV_{lum}', in four thresholds: JP, JN, JU, JI) and spectral states ('CCT', in three levels: Red, Blue, and the combined group) were evaluated to investigate how the means changed with variables related only to the scene. While this showed that the mean discomfort glare (DG) scores increased with GSV_{lum} – validating the alignment of GSV anchor with the VAS responses, the sensitivity of obtained mean DG VAS scores was found to be significantly reduced towards the extremes. The effect of CCT was found to be relatively minor, with 'Blue' spectral conditions generally associated with slightly higher responses on the mean VAS scores.

The mean discomfort glare (DG) scores on VAS against different states of the scene are presented in Appendix A5 Table A5.2. Considering the limited explanatory power of these linear regression models, further models with more predictors were developed using supervised ML techniques, as presented in the following sections.

3.3.1 Description of modelling methods

Building upon the baseline results, eight supervised ML regressors were selected from three families: linear models (ElasticNet, Bayesian Ridge), kernel based (Support Vector Regression, with RBF kernel), and nonlinear tree-based ensemble learners (Random Forest, Gradient Boosting, XGBoost, CatBoost, and LightGBM). The model selection was informed by prior applications in relevant domains, such as the modelling of discomfort glare [32], visual

comfort [31], lighting energy use [57], and daylight factors for regulatory compliance [58]. XGBoost, LightGBM, and CatBoost have previously demonstrated superior performance in dynamic assessment of heating and cooling loads [59], predictive HVAC control [60], solar PV prediction [61], and LED prognostics [62], while SVR and ElasticNet have served as baselines for interpretation.

In an initial implementation, default model parameters were used to prevent overfitting on a limited dataset. Hyperparameter tuning was later introduced for tree-based ensemble models (Random Forest, Gradient Boosting, LightGBM, CatBoost, and XGBoost) using randomized search over key parameter ranges (e.g., number of estimators, maximum depth, learning rate). This process incorporated 5-fold internal cross-validation, and boosting models included early stopping to avoid overfitting. All models used a fixed pseudo-random state for reproducibility. Linear models (ElasticNet, Bayesian Ridge) and SVR were retained with default settings due to their primary role as benchmarks, and due to limited advantages observed from tuning during pilot trials. The target variable for these methods was the perceived discomfort due to glare on a visual analogue scale, whereas the labelled predictors included factors related both to the observer (the persona), the lighting scene, and to the context. Within each method, the regression model was trained to predict the target variable through labelled predictors, which were either categorical or numeric.

Among tree based ensemble models, Random Forest regressor [63] is an algorithm that uses decision trees and averages outputs of each tree to minimize overfitting [64]. Gradient Boosting [65] is another method that combines the decision trees sequentially for reducing prediction errors [66]. CatBoost [67] is a variation of gradient boosting that is designed to handle categorical variables [68]. Extreme Gradient Boosting (XGBoost) [69] is another advanced implementation of gradient boosting that is optimized for scalability [70], while LightGBM is optimized for memory efficiency [71]. Among the second category, Support

Vector Regression (SVR) [72] is a kernel-based method that maps data to higher dimensions, and finds a hyperplane that minimizes the error in prediction [73]. Among linear models, Bayesian Ridge Regression [74] performs probabilistic regression based on Bayesian principles [75], while ElasticNet [76] combines L1 Lasso and L2 Ridge regularization penalties into the loss function [77]. Each of these methods were used to fit the data and train separate models.

3.3.2 Model training

Each model pipeline included a preprocessing step, and a regression step using the selected regressor, in this sequential order. Under the preprocessing step, the predictors were segregated into numerical and categorical (non-numeric) types [50], and together with the outcome, were divided into 80% - 20% split that were respectively used for training the regression model and for internal evaluation of accuracy. This means that 80% of the dataset was used to train the model, while the remaining was used for testing its fit on unseen data. With these settings, the model pipeline was trained to fit the training dataset, and the trained pipeline was used to predict the VAS outcome based on the predictors in the test set.

To reduce the risk of sampling bias while assessing the robustness of model generalizability, a 5-fold cross-validation (CV) procedure was implemented for all eight supervised ML regressors. The mean and standard deviation of the key performance metrics evaluated i.e. MAE and R^2 scores across folds, are reported in Table 7. These results confirm that XGBoost and CatBoost consistently outperform other models in both datasets, supporting the selection of XGBoost for the following steps of the evaluation.

[TABLE 7]

Table 7. Metrics of accuracy for each of the eight supervised machine learning models

Rank	Model	MAE ('*ClusAvg*')	R^2 ('*ClusAvg*')	MAE ('*ClusRep*')	R^2 ('*ClusRep*')
1–2	XGBoost	8.560 ± 0.534	0.631 ± 0.084	10.231 ± 0.461	0.716 ± 0.017
1–2	CatBoost	8.305 ± 0.497	0.707 ± 0.034	10.497 ± 0.336	0.705 ± 0.012
3	LightGBM	9.972 ± 0.696	0.608 ± 0.063	12.316 ± 0.392	0.620 ± 0.022
4	Random Forest	10.243 ± 0.740	0.579 ± 0.066	13.015 ± 0.538	0.570 ± 0.029
5	Gradient Boost	10.561 ± 0.792	0.573 ± 0.090	14.918 ± 0.475	0.470 ± 0.021
6	Bayesian Ridge	13.313 ± 0.655	0.394 ± 0.129	18.905 ± 0.821	0.180 ± 0.036
7–8	ElasticNet	14.847 ± 0.959	0.300 ± 0.036	19.981 ± 0.685	0.120 ± 0.011
7–8	Support Vector	16.386 ± 1.696	0.127 ± 0.037	19.431 ± 0.811	0.160 ± 0.019

In addition to model performance, we evaluated the stability of feature importance across training folds using a 5-fold CV implementation of the XGBoost model. Feature importances were recorded for each fold, and the mean and standard deviation were computed across folds. The coefficient of variation (SD/mean) was used to quantify consistency of feature importance. Features such as GSV_{lum}, Tense, Scale_Disc, Happy, Exp_Diff, and Tcp demonstrated high reliability, with Coefficient of Variation < 0.2, hence supporting their robustness as predictive features across training subsets. This analysis complements the SHAP-based insights in Section 3.3.5. A table describing the feature consistency and relevant metrics has been included in the Appendix A5 Table A5.3.

3.3.3 Interpreting performance

Among metrics of performance, mean absolute error (MAE) defines the average of the absolute differences between the predicted and actual values, and evaluates the magnitude of deviations between the two. A lower MAE represents a good predictive performance. The coefficient of determination (R^2) measures the proportion of variance in the dependent variable that can be predicted from the independent variables, and evaluates how well the model fits the data. A high value of R^2 (close to 1) indicates a good model fit. Conversely, the residual variance is the difference between total and explained variance and, in this case, is approximately the same

as 1-R^2. A low residual variance is expected from a good predictive model.

From the tabulated values in Table 7, it is evident that all tree-based regression methods outperformed the linear models and kernel-based methods, with XGBoost and CatBoost explaining close to 70% of the variance in the data. Among these, the Support Vector Regression (SVR) method demonstrated the least predictive capacity, despite using the Radial Basis Function (RBF) Kernel that can model complex non-linear data [78]. Ideally, this should have improved performance, since the potential non-linearity between the predictors and the outcome could rather explain the lower predictive capacity of linear models. Alternatively, the low performance of the SVR/RBF method could be attributed to its high sensitivity to outliers, compared to the tree-based models that are more robust to outliers in the data. The results from the scene-only linear regression model (Section 3.3) can also be interpreted using these metrics and contextualized with the outcomes presented in Table 7. With only GSV_{lum} and CCT as predictors in linear regression models, the performance was substantially lower, with $R^2 = 0.10$ being even lower than SVR's performance, which necessitated the incorporation of additional factors, i.e. contextual and personal features, to accurately predict discomfort glare ratings.

3.3.4 Model selection and evaluation

Considering the highest accuracy demonstrated by tree-based ensemble methods, the XGBoost method was used for further training and detailed evaluations. Predictions were performed for the dataset, and metrics of accuracy were compared for two sets of predictors: a selected set with 18 predictors (18-P), and for another set with nine key predictors (9-P). The 9-P set included factors that we aimed to investigate, as well as factors previously reported to better explain the variance [26], including: *the glare source luminance (GSVlum), correlated color temperature (Tcp), Time of day (ToD), VAS scale used (Scale), Caffeine consumption score (Cf_Scr), Fatigue (Ftg), mood indicators of 'Tense' and 'Sad',* and *the indicated persona*

(rep_cluster).

With a test-train split again set at 0.2, 1344 rows of data were used for training the two separate XGBoost models (18-P and 9-P), while 336 rows were used for prediction. Both models were ported in pickle format (.pkl) using the Python 'joblib' library and were used to predict VAS scores for the test split in a separate script. The predicted values were compared with actual values from the split dataset. Figure 7 compares the predicted versus the actual VAS scores for both models, and presents the mutual comparison between outcomes also through differential accuracy of model-predicted values against actual values.

[FIGURE 7]

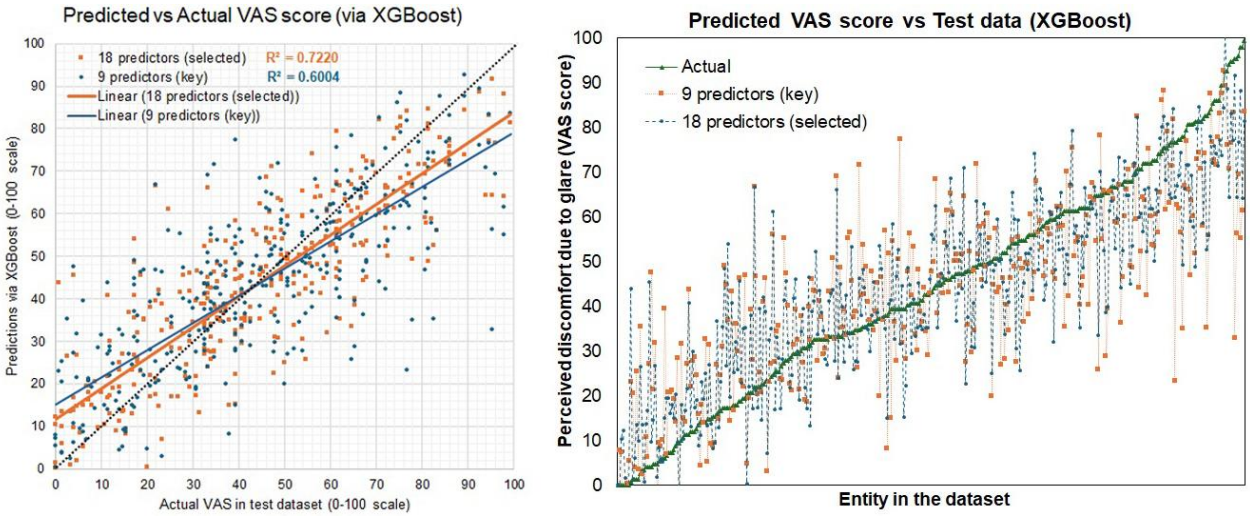

Figure 7. Comparing predicted and actual VAS scores for test dataset for two models trained on different number of predictors. (Left): Predicted values for both models against actual values; (Right): actual VAS data (central line, 336 entities, sorted by increasing value) with deviations in the predictions made by the two models.

The 18-P model generated predictions having MAE, MSE, R^2 values and explained variance at 10.23, 173.46, 0.72, and 72% respectively, while the 9-P model demonstrated slightly reduced accuracy, with these metrics at 11.96, 251.36, 0.60, and 60%. Despite

significantly reducing the number of predictors, the performance of the reduced-predictor XGBoost model was not found to decrease proportionally.

3.3.5 Feature importance

The importance of features within the 18-P model was identified using the '*feature_importances_*' function of the Scikit-Learn's '*Pipeline'* library, and SHAP (SHapley Additive exPlanations) was used to quantify feature contribution to the model using the Python 'SHAP' library. SHAP is an interoperable method of machine learning that is based on Shapley values from cooperative game theory [79]. By attributing changes in the output to individual differences in inputs, this method quantifies each feature's contribution towards the model's overall predictions [80]. SHAP has recently been used to gain insights into black-box models for discomfort glare from daylight [81]. A feature importance plot was generated using the *'Explainer'* and *'summary_plot*' functions in SHAP, while an interaction heatmap was evaluated via the *'TreeExplainer'* function to present interactions between features. The plots are presented in Figure 8.

[FIGURE 8]

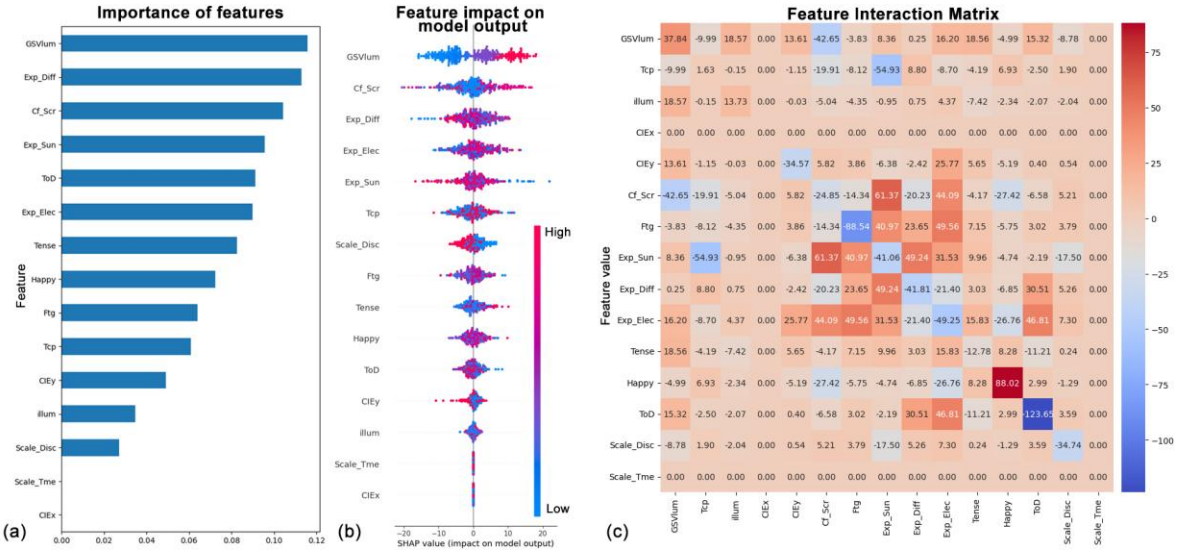

Figure 8. (a) Feature importance plot; (b) SHAP summary plot for feature importance; and, (c) SHAP interaction matrix between features, for the XGBoost model.

For the feature importance plot (Figure 8a), the X axis presents the relative importance of each feature, with the sum-total of all features being equal to 1. The SHAP summary plot (Figure 8b) presents the impact and distribution of features on the predicted value, where the magnitude of each feature on the model's outcome is shown by the horizontal axis. A positive value increases the prediction score, while a negative value reduces it. Through colour scales, each dot further represents the feature value for a specific data point. Through visual inspection, it can be observed that the SHAP summary plot closely resembles the feature importance plot, with both describing a multivariate dependence of the outcome on the predictors, i.e. no individual predictor was able to explain the outcome. Source luminance (*GSVlum*) clearly played a leading role, and the prior luminous exposures (*Exp_**) were also key to the predictions. Caffeine consumption (*Cf_Scr*) and Time of day (*ToD*) also played an important role in the model's outcomes. The CIE chromaticity coordinate and the vertical illuminance values did not significantly change between the eight scenes, which may have led to their low importance towards model prediction. The contextual predictors with the prefix '*Exp_**' identified in the SHAP analysis, i.e. *Exp_Sun*, *Exp_Diff*, and *Exp_Elec*, represent the

participants' self-reported light exposure to direct sunlight, diffuse daylight, and to electric lighting respectively – since they woke up on the day of the experiment until the time they arrived at the test environment. Since these features serve as proxies for luminance adaptation over the course of the day, their influence suggests that luminous history also plays an important role in the perception of discomfort due to glare, which highlights the relevance of incorporating intra-personal 'adaptation' within personalized lighting controls. These SHAP-derived insights are further supported by a feature stability analysis conducted across five CV folds using XGBoost, as described in Section 3.3.2, and tabulated in Appendix A5 Table A5.3.

The SHAP matrix (Figure 8c) represents the interaction between features, where the diagonal elements represent any feature's main effects on the predicted values, while mutual interactions (strength and nature) are represented by the non-diagonal elements. From the interaction matrix, it can be seen that not only were source luminance, caffeine score, and previous luminous exposures influential at the individual level, their interactions with other factors also amplified their net contributions. For example, the interactions between '*Cf_Scr'* with '*Exp_Sun*' (61.37), or of '*Cf_Sc*r' with *'Exp_Elec'* (44.09), significantly amplified the direction of the overall predictions. This step served to demonstrate insights into the black box model, and to identify which features were key to the predictions.

3.3.6 Sensitivity testing beyond the training dataset

The 18-P model was subsequently used to predict VAS scores and evaluate sensitivity of outcome in relation to change in the predictors for different personas. The contextual factors were defined as a reference set, with: Caffeine score at 10.5 (on a 0-15 continuous scale), Fatigue at 3 (on a 0-6 categorical scale); Exposure to direct sunlight/ diffuse daylight/ electric lighting all at 3 (on a 1-7 scale); and, Mood indicators (Tense, Sad, Lively, Calm, Happy) all at 2 (on a 1-5 scale). The variables related to the scene were defined as: Glare source luminance

at 5500 cd/m2 (between the 'JN' and 'JU' thresholds of 3860 and 9150 cd/m2 that ideally correspond to GSV markings between 33% and 66% on the VAS scale); Tcp at 3600 K; vertical illuminance at 400 lux; and, CIE 1931 coordinates of x at 0.4, y at 0.38 (which together yields a CCT value ~3537K). The predicted VAS scores for five personas for this reference set of predictors, as well as for individual changes in predictors, are presented in Table 8.

[TABLE 8]

Table 8. Predictions for the reference set of attributes, and those evaluated after altering one predictor each instance.

Persona	Step	Aurora	Cora	Dongmei	Johannes	Mahendra
VAS	0	49.73	49.29	54.3	55.74	49.95
Action: changing 'caffeine consumption' from reference (10.5) down to 0.5						
VAS	1	54.89	54.45	59.4	61.45	55.12
Action: *[Previous attributes]* + Changing 'fatigue' from reference (3) to 6						
VAS	2	38.52	37.36	42.85	44.48	42.93
Action: *[Previous attributes]* + Changing 'Tense' from reference (2) to 5						
VAS	3	36.15	35.45	40.91	42.67	37.42
Action: *[Previous attributes]* + Changing 'Sad' from reference (2) to 4						
VAS	4	35.61	34.99	40.41	42.14	36.88
Action: *[Previous attributes]* + Changing Scale from 'Disc' to 'Tme'						
VAS	5	40.96	39.54	40.85	42.78	39.8

For the reference set, the predicted discomfort due to glare reported by the personas on the discomfort VAS scale was observed to range between 49.73 to 55.74 (on a 0-100 scale). Changing predictors led to non-linear changes in outcome across personas. For example, changing the states of fatigue and caffeine consumption (a two-step-change from the reference state) led to a reduction between 14% (Mahendra, from 49.9 to 42.9) to 24% (Cora, from 49.2 to 37.3), with an average reduction of 20.4±3.8% compared to the baseline. Also, the influence of each change was not proportionate, and each step on average led to a relative change compared to the previous state, between 1.3% ('*sad'*, from average value of 38.52 in step 3, to 38.01 in step 4) to 27.7% ('*fatigue'*, from average value of 57.06 in step 1, to 41.22 in step 2).

While the model was sensitive to changes in predictor state, the accuracy in predictions could not be further characterised since these evaluations were conducted beyond the training dataset.

3.4 Personalized recommendation via inverse regression

As presented in Figure 1c, the trained XGBoost model was further employed to optimize lighting conditions to achieve a target perception state (expressed as a VAS score) for a specific user persona. This process identified the optimal values for each lighting related predictor, which could match a desired state of visual (dis)comfort, for any persona at any given time of day. An optimization-based prediction was performed via inverse regression [82, 83], where predictors were adjusted to produce the desired outcome. An initial state of predictors was defined based on the mean values of numerical predictors (i.e., variables defining lighting state, and those defining contextual factors), and the '*minimize'* function of SciPy's '*optimize'* class was used to iteratively reduce the difference between the targeted and achieved VAS. The XGBoost regressor model was used to calculate predicted VAS scores after each iteration, while the optimization function aimed to reduce the difference between the target and the evaluated state of lighting predictors. The L-BFGS-B optimization algorithm (limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm [84] with Box constraints) was used to fine tune the hyperparameters of the model, which ensured that the range of parameters was respected. The minimum and maximum bounds for each numerical predictor from the original dataset was defined to limit the possible range of lighting conditions that were tested by the combination of the regressor with the optimization function.

Table 9 presents the state of lighting variables, which were evaluated to match a target discomfort state of 50 (on a 1-100 scale) on the discomfort VAS at 9h, against the following set of factors: Caffeine consumption score: 10.5 (on a 1-15 scale); Fatigue: 3 (on a 0-6 scale);

Prior exposure to direct sunlight: 3, to diffuse daylight: 3, to electric light: 3 (all on a 1-7 scale); Mood indicators of '*Tense'*: 2, *'Sad'*: 2, *'Lively'*: 2, *'Calm'*: 2, *'Happy'*: 2 (all on a 1-5 scale)].

[TABLE 9]

Table 9. The state of lighting variables optimized to match a target VAS on the discomfort scale at a given time of day, for a defined set of personal and contextual factors.

#	Variable	Optimized state	Possible range (from dataset)
1.	Luminance of source	9084.25	1627 to 21700
2.	Tcp of source	4420.75	2609 to 6304
3.	ΔUV	0.01	-0.0001 to 0.0141
4.	Vertical Illuminance	388.12	336 to 485
5.	CIE 1931 x coord	0.39	0.3159 to 0.4711
6.	CIE 1931 y coord	0.38	0.3545 to 0.4116
7.	CIE 1976 u coord	0.23	0.1908 to 0.2699
8.	CIE 1976 v coord	0.51	0.4818 to 0.5292
9.	s-cone EDI	213.77	85.57 to 400.46
10.	m-cone EDI	347.57	262.93 to 486.99
11.	l-cone EDI	391.03	333.96 to 494.30
12.	Rhodopic EDI	298.97	174.65 to 492.19
13.	Melanopic EDI	281.92	145.86 to 491.43
14.	Circadian Lighting	432.65	291.33 to 655.21
15.	Circadian Stimulus	0.38	0.31 to 0.46

To demonstrate the optimizer's adaptability, these intra-individual/contextual features were manually set to simulate a user's current context – and their states were not based on cluster means. This model-based optimizer is designed to adjust its output dynamically in response to real-time user inputs, supporting personalized and context-aware lighting recommendations.

3.4.1 Limitations of the optimization process

The optimization routine of the algorithm, however, was not found to be sensitive to changes in persona, which was likely due to the limited variability of lighting states in the training dataset. The experiment only included 8 lighting scenes – JP/JN/JU/JI across 'Red' and 'Blue' spectral states – and, therefore, 8 states across all 15 lighting related predictors. Nevertheless,

when this optimization algorithm was tested on a synthetic dataset (which was generated to have a similar structure as the enriched dataset, and the 1680 constituents within each of its 34 columns had similar descriptive statistics) a sensitivity in optimized states against different personas and contextual factors was observed. Yet, since the synthetic dataset had no reliability beyond a statistical similarity to the experimental data, the optimized state of lighting predictors was not further investigated.

4. Discussion

Machine learning (ML) techniques are gaining popularity for data analysis and modelling due to their capacity for handling complex datasets and for detecting non-linear relationships between variables. Unlike conventional statistical methods that are limited to pre-specified functional relationships (such as linear, polynomial, etc.), ML models can adapt to diverse data structures and prove effective for multivariate analyses. These advantages make ML techniques particularly interesting for treating datasets from lighting research experiments, especially those that are focused on modelling discomfort due to glare.

Predicting discomfort perception is, in fact, highly challenging since the subjective nature of personal evaluations often leads to large unexplained variances. Critically, current discomfort glare models focus exclusively on objective factors related to the lighting scene, while neglecting several person- and context-related factors that have been widely reported to influence the perception of discomfort due to glare[26-28, 85-87]. By disregarding these factors, any variance that could have potentially been explained becomes residual (i.e., unexplained).

In the dataset that was used for this paper, the experimental user-testing data had been collected with the aim to analyse the concurrent effects of personal and contextual factors on the prediction of discomfort due to glare. Yet, the presence of several features that were

difficult to treat via statistical tools already used within the scientific lighting literature (e.g., analysis of variance, multilevel modelling, etc.[88-94]) made it appealing to explore the application of supervised ML techniques, a category of machine learning methods that works with labelled features. Nevertheless, ML techniques also present certain challenges, such as the risk of overfitting and the need for adequately pre-processed datasets for training the models. Additionally, the "black box" nature of ML models can make interpreting their outcomes difficult.

To address these challenges, the methods illustrated in this paper performed data preprocessing and feature optimisation via techniques such as K-means clustering, rule-based classification, feature encoding, and feature selection. Following up, synthetic Personas were generated to represent the user base, and to reduce the number of predictors from 27 to 11. This generated a high-quality dataset with reduced features, and eased the training task while reducing the risks associated with data sparseness and overfitting. Among the eight different ML methods employed for training, a wide range of accuracy was identified among the models, with Support Vector Regressor explaining as low as 18% of the variance in data, while tree-based regression models of CatBoost and XGBoost explaining as high as 70% with a 5-fold cross validation. While reducing the number of predictors in the XGBoost model did not lead to a proportional reduction in the model's performance (72% to 60% after reducing half the predictors, which demonstrates the robustness of this technique), the drop in accuracy was still significant. By evaluating feature importance and summary using SHAP method, it was identified that no feature individually dominated the importance value, and about 13 features cumulatively explained the predictions, which were also supported by a feature stability analysis conducted across 5 CV folds Sensitivity testing on an external set of conditions demonstrated non-linear change in model-predicted outcome across personas (between 14% to 24%, for a given set of conditions), and also a non-linear change in predictions across each

change in state of predictors (between 1.3% to 27% for specific sets of actions). These observations illustrate the advantage of using multivariate modelling capacities of supervised ML for making the prediction models, whereas tree-based regression methods were proven superior to kernel-based methods owing to their robustness against outliers. The capacity of explaining variance is even more relevant for this case, since the lighting scenes were designed using the IES-GI model, which does not consider subjective factors related to the user nor to the context. All variance in responses detected within each the four GSV-anchored subsets, therefore, had to be explained only via subjective factors, else they became unexplained residuals. With the 18-P XGBoost model (of which 13 features demonstrated importance) explaining up to 75% of the variance in data, the predictions exceeded our initial expectations.

The application of supervised ML for optimising the state of lighting factors via inverse regression was particularly innovative, since it can open up the avenue for wider applications for user-centric lighting based on personalised perception. However, the portable regression model did not prove sensitive to changes in personas. This inadequacy can be attributed to the limitation of our dataset, since the key predictors (including the state of lighting) were strictly maintained while collecting subjective data, that is perception of discomfort due to glare. The dataset included only 8 states of lighting variables against over 1800 reported states of discomfort. While the tight control on predictors proved particularly effective for regression-based modelling, the efficacy was much reduced in inverse regression where the state of discomfort was constrained (while the model was used to predict the optimal state of lighting variables, in combination with an iterative optimiser). The lack in sensitivity for inverse regression can, thus, be attributed to data sparseness in lighting states within our dataset. To address this, and to achieve goals such as inverse regression in future studies, data collection scenarios must combine category rating tasks (i.e., tightly controlled lighting states against subjective perception of discomfort due to glare – as in this study) with luminance adjustment

tasks (i.e., tightly controlled states of subjective perception against variable lighting states, as in [95, 96]), or data from field studies using daylight that simultaneously measure the scene variables along with subjective factors. The novelty of our study consists in its integration of psychophysical training data, rich in contextual and individual-level features, within a supervised ML framework for modelling discomfort due to glare. In contrast to simulation-driven models, such as those developed by Nourkojouri et al. (2021) and Li et al. (2025), our work anchors predictions with training data using subjective user ratings collected under controlled experimental conditions. The inclusion of contextual factors (e.g., time-of-day, caffeine intake, fatigue, etc.) together with our clustering of individual profiles into personas, enables us to account for a significant variability in individual glare response. By comparing multiple non-linear regression models through 5-fold cross-validation, and by further analysing their consistency in performance and stability in feature importance against the different training datasets, our study offers a rigorous and replicable modelling framework. Additionally, the demonstrated capacity for optimization supports not only a static evaluation of discomfort due to glare, but also an active process towards a personalized lighting control.

Although conventional models for discomfort glare rely on characteristics related to the luminous scene – such as the size and position of glare source, background luminance, etc. – the absence of such geometric and photometric variables in our models is uniquely due to the specific setup constraints of our controlled experiment. In our study, the IES-GI model relied upon set luminous conditions that achieved four thresholds of glare sensation votes, whereas the glare source's luminance was the only degree of freedom, while all other factors related to the scene were constrained. On these bases, our models use variables associated to the individual and contextual factors as additional predictors to glare source luminance. Although data on these attributes may be less readily accessible, our findings demonstrate their relevance in improving the accuracy of prediction, hence offering a pathway toward personalized lighting

control, especially considering the increase in wearable/nearable and environmental sensors, and the future scopes for acquiring such data in real time. While our results are specific to the dataset used for this study, the methods for data treatment, modelling, and optimisation, are widely transferable to other applications. These could be easily relevant to sister domains, e.g. for thermal comfort studies[97], or be upscaled to the urban domain for outdoor lighting research. As an example, for the latter, multiple contrasting priorities need to be balanced by lighting states, such as providing sufficient light for safety and visibility while giving attention towards excessive luminous exposure at nighttime and avoid phenomena of circadian disruption[98-101]. Balancing diverse demands for different stakeholders, and providing the right light to the right user (at the right time), can indeed present a prolific future field of application for the methods presented in this paper. A limitation of this study is that the models were developed and evaluated on a dataset collected from a single experimental study. Considering the absence of comparable datasets that capture both scene-related and individual/contextual factors, external validation could not be performed. While this limits the model's immediate applicability to other environments, the results highlight the potential scope for incorporating such data in modelling discomfort due to glare. Future work should focus on replicating these methods with new datasets collected in different settings.

5. Conclusions

A method based on supervised machine learning is demonstrated for analysing subjective data collected from a controlled laboratory experiment, where the sensation of discomfort due to glare was evaluated under various scenes created using tunable electric lighting. The dataset included predictors related to the lighting scene, personal variables that characterised participants, and contextual factors that were specific to each session. The data was treated via preprocessing, including feature engineering, encoding, and feature selection, and individual

test subjects were characterised into distinct personas having differing lighting needs. In parallel, redundant contextual features were identified and selected to improve feature quality. Although the models generated by this study have not been externally validated, this study intends to offer an approach that can inform future research aiming to personalize prediction and controls related to lighting conditions.

The optimised dataset was used to train eight machine learning models, which were compared based on metrics of accuracy. XGBoost emerged as the best performing model, and was further investigated for comparing its predicted scores with actual scores from the test dataset. Sensitivity analysis was conducted, and the model was seen as sensitive to changes both in personas as to those in contextual factors. The deployment of this portable model was also explored for optimizing the state of lighting variables via inverse regression, where the lighting states specific to each user were evaluated by defining the target perception zone against every contextual factor. The application of personalised controls makes this method interesting for applications beyond discomfort glare prediction. While the transferability of these models is limited by the fixed configuration of the lighting scenes in our study, and by the requirement for readily available personal and contextual data, our study lays the pathway for integrating personal and contextual variables into discomfort glare modeling.

Nomenclature

CIE:	Commission internationale de l'éclairage (International Commission on Illumination)
GSV:	Glare Sensation Vote
IES-GI:	Illuminating Engineering Society Glare Index
JI:	Just Intolerable
JN:	Just Noticeable
JP:	Just Perceptible
JU:	Just Uncomfortable
LRC:	Lighting Research Centre
MAE:	Mean Absolute Error
ML:	Machine Learning
MSE:	Mean Squared Error
RBF:	Radial Basis Function
RMPE:	Recommended minimum effect size representing a “practically” significant effect
SHAP:	SHapley Additive exPlanations
SVR:	Support Vector Regression
Tcp/CCT:	Correlated Colour Temperature
ToD:	Time of Day
XGBoost:	Extreme Gradient Boosting
VAS:	Visual Analog Scale

Appendices: (attached at the end)

1) Visual Analogue Scale used for evaluating perception of discomfort due to glare.
2) Details on the experiment (setup and timeline)
3) Questionnaire for collecting personal factors. (Background Questionnaire)
4) Questionnaire for collecting contextual factors. (In-study Questionnaire)
5) Appendix tables

Funding details:

This work was supported by FNRS - Fonds de la Recherche Scientifique (Belgium) through fellowships granted to the first author. Data collection was supported by the postdoctoral fellowship (n. 40000322) granted for the project SCALE (Shading Control Algorithms from Luminance-based Evaluations). Data analysis and manuscript preparation was supported by scientific collaborator fellowship (n. 40024409) granted for the project SPARKLE (Spectral Profiles for Agent-based Responsive networK of Lighting Energy).

Declaration of Generative AI and AI-Assisted Technologies:

For the data processing steps, the author(s) used ChatGPT, based on OpenAI's GPT-4 model, to refine and debug Python codes which were originally written by the author(s). The author(s) reviewed, tested, and adapted the suggestions as needed, and take full responsibility for the final implementation of the codes, and for the content of the publication.

Ethical Approval:

This study was approved by the IPSY Ethics Committee at UCLouvain (Project 2022-65)

Disclosure Statement:

The authors report there are no competing interests to declare.

References

1. Jung, D., et al., *Visual fatigue prediction using classification model based on physiological responses of occupants under office lightings.* Journal of Building Engineering, 2024. **98**: p. 111146.
2. Makridakis, S., et al., *Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward.* Journal of the Operational Research Society, 2023. **74**(3): p. 840-859.
3. Nichols, J.A., H.W. Herbert Chan, and M.A.B. Baker, *Machine learning: applications of artificial intelligence to imaging and diagnosis.* Biophys Rev, 2019. **11**(1): p. 111-118.
4. Sarker, I.H., *Machine Learning: Algorithms, Real-World Applications and Research Directions.* SN Computer Science, 2021. **2**(3): p. 160.
5. Manandhar, P., H. Rafiq, and E. Rodriguez-Ubinas, *Current status, challenges, and prospects of data-driven urban energy modeling: A review of machine learning methods.* Energy Reports, 2023. **9**: p. 2757-2776.

6. Ali, U., et al., *Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach.* Energy and Buildings, 2024. **303**: p. 113768.
7. Sun, Y., F. Haghighat, and B.C.M. Fung, *A review of the-state-of-the-art in data-driven approaches for building energy prediction.* Energy and Buildings, 2020. **221**: p. 110022.
8. Sutton, R.S., *Reinforcement learning: An introduction.* A Bradford Book, 2018.
9. Wu, Z., et al., *Reinforcement learning in building controls: A comparative study of algorithms considering model availability and policy representation.* Journal of Building Engineering, 2024. **90**: p. 109497.
10. Weinberg, D., et al., *A Review of Reinforcement Learning for Controlling Building Energy Systems From a Computer Science Perspective.* Sustainable Cities and Society, 2023. **89**: p. 104351.
11. Shan Lee, V.L., et al., *Semi-supervised Learning for Sentiment Classification using Small Number of Labeled Data.* Procedia Computer Science, 2019. **161**: p. 577-584.
12. Han, X., et al., *Prediction of human thermal comfort preference based on supervised learning.* Journal of Thermal Biology, 2023. **112**: p. 103484.
13. Putrada, A.G., et al., *Machine Learning Methods in Smart Lighting Toward Achieving User Comfort: A Survey.* IEEE Access, 2022. **10**: p. 45137-45178.
14. Zhuang, D., et al., *Supervised learning-based assessment of office layout satisfaction in academic buildings.* Building and Environment, 2022. **216**: p. 109032.
15. Mathew, V., C. Pearl Kurian, and N. Augustine, *Climate-responsive machine learning-based control of switchable glazing towards human-centric lighting.* Solar Energy, 2023. **260**: p. 49-60.
16. Kent, M.G., S. Fotios, and S. Altomonte, *Discomfort glare evaluation: The influence of anchor bias in luminance adjustments.* Lighting Research & Technology, 2019. **51**: p. 131-146.
17. Akashi, Y., R. Muramatsu, and S. Kanaya, *Unified Glare Rating (UGR) and subjective appraisal of discomfort glare.* http://dx.doi.org/10.1177/14771535960280040501, 1996. **28**(4): p. 199-206.
18. Chauvel, P., et al., *Glare from windows: current views of the problem.* http://dx.doi.org/10.1177/096032718201400103, 1982. **14**(1): p. 31-46.
19. Einhorn, H.D., *Discomfort glare: a formula to bridge differences.* http://dx.doi.org/10.1177/14771535790110020401, 1979. **11**(2): p. 90-94.
20. Petherbridge, P. and R.G. Hopkinson, *Discomfort Glare and the Lighting of Buildings.* Lighting Research and Technology, 1950. **15**(2 IEStrans): p. 39-79.
21. Robinson, W., et al., *The Development of the IES Glare Index System.* http://dx.doi.org/10.1177/147715356202700103, 1962. **27**(1 IEStrans): p. 9-26.
22. Wienold, J. and J. Christoffersen, *Evaluation methods and development of a new glare prediction model for daylight environments with the use of CCD cameras.* Energy and Buildings, 2006. **38**(7): p. 743-757.
23. Carlucci, S., et al., *A review of indices for assessing visual comfort with a view to their use in optimization processes to support building integrated design.* Renewable and Sustainable Energy Reviews, 2015. **47**: p. 1016-1033.
24. Kent, M.G., et al., *Discomfort glare and time of day.* Lighting Research & Technology, 2014. **47**(6): p. 641-657.
25. Rodriquez, R.G. and A. Pattini, *Tolerance of discomfort glare from a large area source for work on a visual display.* Lighting Research & Technology, 2012. **46**(2): p. 157-170.
26. Kent, M.G., et al., *Temporal variables and personal factors in glare sensation.* Lighting Research & Technology, 2015. **48**(6): p. 689-710.
27. Altomonte, S., et al., *Visual task difficulty and temporal influences in glare response.* Building and Environment, 2016. **95**: p. 209-226.
28. Pierson, C., J. Wienold, and M. Bodart, *Review of Factors Influencing Discomfort Glare Perception from Daylight.* LEUKOS, 2018. **14**(3): p. 111-148.

29. Nourkojouri, H., et al., *A machine-learning framework for daylight and visual comfort assessment in early design stages*. 2021.
30. Li, X., et al., *Highly generalized surrogate models for indoor glare performance based on multimodal deep learning networks at the early design stage.* Building and Environment, 2025. **273**: p. 112706.
31. Tzouvaras, C., et al. *A Novel Dynamic Approach for Determining Real-Time Interior Visual Comfort Exploiting Machine Learning Techniques*. Applied Sciences, 2023. **13**, DOI: 10.3390/app13126975.
32. Wagdy, A., et al., *Open-plan Glare Evaluator (OGE): A Demonstration of a New Glare Prediction Approach Using Machine Learning Algorithms*. 2019.
33. Maskarenj, M., M.G. Kent, and S. Altomonte, *Effect of time of day and correlated colour temperature on discomfort glare.* Lighting Research & Technology (Submitted), 2024.
34. Maskarenj, M. and S. Altomonte. *Mixed effect of time of day and correlated colour temperature on discomfort glare*. in *Bright Environments: Daylight in Sustainable Building Design*. 2023.
35. Afraz, A. and P. Cavanagh, *The gender-specific face aftereffect is based in retinotopic not spatiotopic coordinates across several natural image transformations.* Journal of vision, 2009. **9**: p. 10.1-17.
36. CIE, *CIE 212:2014. Guidance towards best practice in psychophysical procedures used when measuring relative spatial brightness*. 2014, Commission Internationale De L'Éclairage: Vienna, Austria.
37. Schlangen, L.J.M. and L.L.A. Price, *The Lighting Environment, Its Metrology, and Non-visual Responses.* Frontiers in Neurology, 2021. **12**(235).
38. CIE-Div-6, *CIE S 026 α-opic Toolbox*. 2020, CIE: Vienna.
39. LRC, *Lighting Research Centre Circadian Light Combined Calculator*. 2017: New York.
40. Roenneberg, T., et al., *Chronotype and Social Jetlag: A (Self-) Critical Review.* Biology (Basel), 2019. **8**(3).
41. Gawron, V.J., *Overview of Self-Reported Measures of Fatigue.* The International Journal of Aviation Psychology, 2016. **26**(3-4): p. 120-131.
42. Shahid, A., et al., *Karolinska Sleepiness Scale (KSS)*, in *STOP, THAT and One Hundred Other Sleep Scales*, A. Shahid, et al., Editors. 2012, Springer New York: New York, NY. p. 209-210.
43. Kompier, M.E., et al., *Effects of light transitions on measures of alertness, arousal and comfort.* Physiology & Behavior, 2020. **223**: p. 112999.
44. Kent, M.G., S. Fotios, and S. Altomonte, *Order effects when using Hopkinson's multiple criterion scale of discomfort due to glare.* Building and Environment, 2018. **136**: p. 54-61.
45. Fotios, S. and M. Kent, *Measuring Discomfort from Glare: Recommendations for Good Practice.* LEUKOS, 2021. **17**(4): p. 338-358.
46. Ying, X., *An Overview of Overfitting and its Solutions.* Journal of Physics: Conference Series, 2019. **1168**(2): p. 022022.
47. MacQueen, J. *Some methods for classification and analysis of multivariate observations*. in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*. 1967.
48. Lloyd, S., *Least squares quantization in PCM.* IEEE transactions on information theory, 1982. **28**(2): p. 129-137.
49. Li, Y. and H. Wu, *A Clustering Method Based on K-Means Algorithm.* Physics Procedia, 2012. **25**: p. 1104-1109.
50. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**(null): p. 2825–2830.
51. Hu, C.-C. and H.-L. Li, *Deducing the classification rules for thermal comfort controls using optimal method.* Building and Environment, 2016. **98**: p. 107-120.

52. Polańska, S., A. Karykowska, and Ł. Pawelec, *Associations between chronotype and physical activity and well-being in adults.* Chronobiology International, 2024. **41**(4): p. 521-529.
53. Montaruli, A., et al., *Biological rhythm and chronotype: new perspectives in health.* Biomolecules, 2021. **11**(4): p. 487.
54. Sokol, N., et al., *'Personas for lighting'. Three methods to develop personas for the indoor lighting environment.* Energy and Buildings, 2023. **278**: p. 112580.
55. Quek, G., et al., *Comparing performance of discomfort glare metrics in high and low adaptation levels.* Building and Environment, 2021. **206**: p. 108335.
56. Ferguson, C.J., *An effect size primer: A guide for clinicians and researchers.* Professional Psychology: Research and Practice, 2009. **40**(5): p. 532-538.
57. Bian, Y., et al., *Using machine learning algorithm to predict lighting energy consumption of daylight-linked lighting systems from spatial daylight autonomy.* Energy and Buildings, 2025. **341**: p. 115847.
58. Dieguez, A.P., et al., *Daylight factor prediction using machine learning: A two-way study using numerical encoding and regression models, versus image encoding and pix2pix.* Building and Environment, 2025. **274**: p. 112743.
59. Guo, J., et al., *Prediction of heating and cooling loads based on light gradient boosting machine algorithms.* Building and Environment, 2023. **236**: p. 110252.
60. Wang, H., et al., *An XGBoost-Based predictive control strategy for HVAC systems in providing day-ahead demand response.* Building and Environment, 2023. **238**: p. 110350.
61. Banik, R. and A. Biswas, *Improving Solar PV Prediction Performance with RF-CatBoost Ensemble: A Robust and Complementary Approach.* Renewable Energy Focus, 2023. **46**: p. 207-221.
62. Gupta Bakshi, B. and B. Roy, *A study on prognostics of LED luminaires based on gradient-boosting decision tree classifiers.* Lighting Research & Technology, 2025: p. 14771535251335974.
63. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
64. Linda, L., *Understanding Random Forests: An Ensemble of Decision Trees*. 2023.
65. Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial.* Frontiers in Neurorobotics, 2013. **7**(DEC): p. 21-21.
66. Baladram, S., *Gradient Boosting Regressor, Explained: A Visual Guide with Code Examples*. 2024.
67. Prokhorenkova, L., et al. *Catboost: Unbiased boosting with categorical features*. 2018.
68. Ershov, V., *CatBoost - open-source gradient boosting library*. 2018.
69. Chen, T. and C. Guestrin, *XGBoost: A scalable tree boosting system.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. **13-17-August-2016**: p. 785-794.
70. Tarwidi, D., et al., *An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach.* MethodsX, 2023. **10**: p. 102119-102119.
71. Ke, G., et al. *LightGBM: A highly efficient gradient boosting decision tree*. 2017.
72. Montesinos López, O.A., A. Montesinos López, and J. Crossa, *Support Vector Machines and Support Vector Regression.* Multivariate Statistical Machine Learning Methods for Genomic Prediction, 2022: p. 337-378.
73. Jain, S., *Support Vector Machine (SVM) Algorithm*. 2024.
74. Permai, S.D. and H. Tanty, *Linear regression model using bayesian approach for energy performance of residential building.* Procedia Computer Science, 2018. **135**: p. 671-677.
75. Koehrsen, W., *Introduction to Bayesian Linear Regression*. 2018.
76. Zou, H. and T. Hastie, *Regularization and Variable Selection Via the Elastic Net.* Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005. **67**(2): p. 301-320.
77. Jain, A., *Elastic Net Regression —Combined Features of L1 and L2 regularization*. 2024.

78. Musavi, M.T., et al., *On the training of radial basis function classifiers.* Neural Networks, 1992. **5**(4): p. 595-603.
79. Hokari, T. and W. Thomson, *Cooperative Game Theory*, in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, J.D. Wright, Editor. 2015, Elsevier: Oxford. p. 867-880.
80. Lundberg, S. and S.-I. Lee, *An unexpected unity among methods for interpreting model predictions.* arXiv [cs.AI], 2016.
81. Luo, Z., et al., *Investigation of influential variations among variables in daylighting glare metrics using machine learning and SHAP.* Building and Environment, 2024. **254**: p. 111394.
82. Devijver, E. and E. Perthame, *Prediction regions through Inverse Regression.* Journal of Machine Learning Research, 2020. **21**(113): p. 1-24.
83. Parker, P., et al., *The prediction properties of inverse and reverse regression for the simple linear calibration problem.* 2010(LF99-9222).
84. Evins, R., *A review of computational optimisation methods applied to sustainable building design.* Renewable and Sustainable Energy Reviews, 2013. **22**: p. 230-245.
85. Hamedani, Z., et al., *Lighting for work: A study of the relationships among discomfort glare, physiological responses and visual performance.* Building and Environment, 2020. **167**: p. 106478.
86. Marié, S., et al., *Evaluation of Physiological Parameters on Discomfort Glare Thresholds Using LUMIZ 100 Tool.* Transl Vis Sci Technol, 2021. **10**(8): p. 28.
87. Jain, S., et al., *Influence of macular pigment on the sensitivity to discomfort glare from daylight.* Scientific Reports, 2023. **13**(1): p. 18551.
88. Suk, J.Y., *Evaluating Daylight Glare and Identifying Its Dominant Cause in Computer-Based Office Task Environment*. 2017.
89. Kent, M., S. Fotios, and S. Altomonte, *Experimental biases in discomfort glare evaluations*. 2018.
90. Girard, J., C. Villa, and R. Brémond, *Discomfort Glare from Several Sources: A Formula for Outdoor Lighting.* LEUKOS, 2021. **17**(2): p. 108-124.
91. Pierson, C., et al., *Is there a difference in how people from different socio-environmental contexts perceive discomfort due to glare from daylight?* Lighting Research & Technology, 2021. **54**(1): p. 5-32.
92. Christoffersen, J. and J. Wienold, *Assessment of user reaction to glare with three solar shading systems*. 2008.
93. Stanke, L., L. Viktorová, and T. Dominik *Discomfort Glare Perception by Drivers—Establishing a Link between Subjective and Psychophysiological Assessment*. Applied Sciences, 2022. **12**, DOI: 10.3390/app12083847.
94. Chamilothori, K., et al., *Subjective and physiological responses towards daylit spaces with contemporary façade patterns in virtual reality: Influence of sky type, space function, and latitude.* Journal of Environmental Psychology, 2022. **82**: p. 101839.
95. Kim, W. and K. Jeong Tai, *A Position Index Formula for Evaluation of Glare Source in the Visual Field.* Indoor and Built Environment, 2010. **20**(1): p. 47-53.
96. Tuaycharoen, N. and P. Tregenza, *Discomfort glare from interesting images.* Lighting Research & Technology - LIGHTING RES TECHNOL, 2005. **37**: p. 329-341.
97. Kim, J., et al., *Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning.* Building and Environment, 2018. **129**: p. 96-106.
98. Davis, L.K., et al., *Health Effects of Disrupted Circadian Rhythms by Artificial Light at Night.* Policy Insights from the Behavioral and Brain Sciences, 2023. **10**(2): p. 229-236.
99. Wang, T., N. Kaida, and K. Kaida, *Effects of outdoor artificial light at night on human health and behavior: A literature review.* Environmental Pollution, 2023. **323**: p. 121321.

100. Lok, R., et al., *Timing of outdoor light exposure is associated with sleep-wake consolidation in community-dwelling older men.* Frontiers in Sleep, 2023. **2**.
101. Meléndez-Fernández, O.H., J.A. Liu, and R.J. Nelson *Circadian Rhythms Disrupted by Light at Night and Mistimed Food Intake Alter Hormonal Rhythms and Metabolism*. International Journal of Molecular Sciences, 2023. **24**, DOI: 10.3390/ijms24043392.

Figures

(Presented inline, attached as TIFF format at the end)

Tables

(appended below)

Table 1. Photometric and radiometric measurements for the eight scenes measured at eye level using Konika Minolta CL70F spectrometer (1nm resolution between 380-780nm). The CIE S026 a-opic metrics are derived from spectral data using the CIEs026 calculator, and the circadian metrics are derived using the LRC calculator.

					CIE 1931		CIE 1976		EDI based on CIE s-026 calculator						
CCT indicator	**GSV**	**Lum (cd/m^2)**	**Tcp (K)**	**Vert Illum (lx)**	**x-coord**	**y-coord**	**u-coord**	**v-coord**	**s-cone (lx, EDI_S)**	**m-cone (lx, EDI_M)**	**l-cone (lx, EDI_L)**	**Rods (lx, EDI_R)**	**melan. (lx, EDI_{mel})**	**Circad Lighting (lx, CL_A)**	**Circad Stim.**
Red	JP	1627	2645	336	0.4691	0.4116	0.2680	0.5292	85.57	262.93	342.50	174.66	145.86	291.33	0.3116
	JN	3860	2638	345	0.4693	0.4110	0.2684	0.5289	88.51	269.94	352.09	179.02	149.27	297.94	0.3159
	JU	9150	2620	394	0.4708	0.4112	0.2693	0.5292	99.62	306.96	401.71	201.87	167.12	333.64	0.3376
	JI	21700	2609	485	0.4711	0.4104	0.2699	0.5290	123.50	376.89	494.30	245.31	200.97	401.23	0.3732
		Mean	**2628**	**390**	**0.4701**	**0.4111**	**0.269**	**0.529**	**99.3**	**304.2**	**397.6**	**200.2**	**165.8**	**331**	**0.335**
		StDev	**16.5**	**68.26**	**0.001**	**0.0005**	**9E-04**	**1E-04**	**17.23**	**52.18**	**69.46**	**32.35**	**25.23**	**50.35**	**0.028**
Blue	JP	1627	6304	336	0.3159	0.3545	0.1908	0.4818	290.57	340.84	333.97	347.79	348.18	469.75	0.4032
	JN	3860	6271	343	0.3166	0.3551	0.1910	0.4822	295.25	348.13	341.47	355.41	356.16	479.01	0.4069
	JU	9150	6210	383	0.3179	0.3558	0.1916	0.4827	326.68	387.90	381.37	395.50	396.40	533.04	0.4268
	JI	21700	6069	483	0.3209	0.3588	0.1926	0.4846	400.46	487.00	480.85	492.20	491.43	655.22	0.4636
		Mean	**6214**	**386.2**	**0.3178**	**0.3561**	**0.192**	**0.483**	**328.2**	**391**	**384.4**	**397.7**	**398**	**534.3**	**0.425**
		StDev	**104**	**67.74**	**0.0022**	**0.0019**	**8E-04**	**0.001**	**50.75**	**67.28**	**67.57**	**66.37**	**65.74**	**85.34**	**0.028**

Table 2. Descriptive statistics on data related to personal factors. Sample size = 38 subjects

Gender	Age	Ethnicity	Eye conditions	Chronotype	Photosensitivity
52.6% *Male* 47.3% *Female*	Range: 22-64 Med: 32 34.4 ± 11.6	***White***: 73.6% ***Asian***: 21% ***Mixed***: 5.3%	42% wore lenses 14% reported other vision issues	7-point scale, 2.95 ± 1.4 Median: 3	**Natural Light**: 68.4% **Artificial Light**: 60.6% **Sunglasses**: 39.4% **Blind preference**: 1.15±0.85 on the 0-4 scale **Lighting preferred**: 0.73±0.75 on a 0-3 scale
Caffeine Habit					
Yes: 67.5%					

Table 3. Descriptive statistics on data related to contextual factors. N=114 scenes

Caffeine score	Prior Light Exposure (0-7)	Mood Indicators (1-5)	
21.1 ± 25.5%		***Tense***	1.91 ± 0.99
Continuous VAS	***Direct sunlight:***	***Sad***	1.53 ± 0.8
Fatigue (0-6)	5.1±1.8	***Lively***	3.37 ±0.82
2.14±1.29	***Diffuse daylight***	***Sleepy***	2.24±0.96
Samn-Perelli scale	3.9±1.6	***Calm***	3.7±1.02
Sleepiness (1-9)	***Electric light***	***Happy***	3.6±0.84
4.2±1.7	4.3±1.8	***Drowsy***	2.4±1.05
Karolinska scale		***Awake***	3.53±0.96

Table 4. Frequencies and ranges of data distribution for participant attributes between clusters. The last column includes the interpreted photosensitivity, and the footnote describes the meaning of the anchors and values.

	Range			Frequencies			Photosensitivity
Cluster	**'Blinds'**	**'Lt_Pref'**	**'Mean_VAS'**	**'Nat_Lt'**	**'Art_Lt'**	**'Sunglss'**	**(Interpretation)**
0	[1, 2]	[1, 3]	[19.77, 60.13]	0: 100%, 1 :0%	0: 90.9%, 1: 9.09%	1: 54.5%, 0: 45.4%	'High'
1	[0, 1]	[0, 1]	[8.16, 62.18]	0: 76.47%, 1: 23.52%	0: 52.9%, 1: 47.05%	0: 52.9%, 1: 47.05%	'Moderate'
2	[1, 3]	[0, 1]	[34.22, 59.88]	0: 20%, 1: 80%	0: 40%, 1: 60%	0: 10%, 1: 90%	'Low'

Feature description and interpretation:
Use of Blinds ('Blinds'): 0-4, with 4= blinds always down, 0 = blinds always up
Preference for lighting condition ('Lt_Pref'): 0-3, with 3 = dark preferred, 0 = well lit preferred
Sensitive to Natural Light ('Nat_Lt'): 0 is Yes (i.e. sensitive), 1 is Not sensitive.
Sensitive to Artificial Light ('Art_Lt'): 0 is Yes, 1 is Not sensitive.
Use of Sunglasses ('Sunglss'): 0 is Yes (meaning sensitive), 1 is No (could mean not sensitive).
Mean reported discomfort due to glare ('Mean_VAS): Higher discomfort indicates higher photosensitivity.

Table 5. The 53 combinations demonstrating correlations (via Spearman method) above the 0.5 threshold. These were part of three distinct sets, as indicated with the overlay color.

#	Profile A	Profile B	Corr	#	Profile A	Profile B	Corr	#	Profile A	Profile B	Corr
1	P0002110	P0001120	0.561	19	P0101101	P0101100	0.698	37	P1112010	P0002201	0.559
2	P0002201	P0001211	0.507	20	P0101121	P0101110	0.592	38	P1112010	P0102211	0.543
3	P0021111	P0002110	0.684	21	P0102101	P0101110	0.561	39	P1112010	P1102100	0.516
4	P0021111	P0002120	0.518	22	P0102211	P0002201	0.593	40	P1122020	P0021111	0.524
5	P0021121	P0002110	0.532	23	P1000020	P0001210	0.650	41	P1122020	P0021121	0.565
6	P0021121	P0021111	0.578	24	P1000020	P0101101	0.591	42	P1122020	P0022120	0.576
7	P0022011	P0001211	0.539	25	P1001100	P0101110	0.557	43	P1122020	P0101101	0.551
8	P0022011	P0002201	0.583	26	P1001100	P0102101	0.568	44	P1122020	P1101000	0.514
9	P0022120	P0001120	0.501	27	P1101210	P0101110	0.658	45	P1122020	P1102100	0.505
10	P0022120	P0001211	0.546	28	P1101210	P0101121	0.613	46	P2000100	P0001210	0.547
11	P0022120	P0002110	0.625	29	P1102100	P0002110	0.691	47	P2000100	P0021111	0.595
12	P0022120	P0021121	0.581	30	P1102100	P0021111	0.590	48	P2000100	P0101100	0.533
13	P0101100	P0002110	0.719	31	P1102100	P0021121	0.502	49	P2000100	P0122121	0.534
14	P0101100	P0021111	0.772	32	P1102100	P0022120	0.590	50	P2001010	P0001210	0.546
15	P0101100	P0021121	0.613	33	P1102100	P0101100	0.765	51	P2001100	P0022120	0.566
16	P0101101	P0012010	0.531	34	P1102100	P0101101	0.636	52	P2001100	P0101101	0.551
17	P0101101	P0021121	0.599	35	P1102110	P0101101	0.608	53	P2101010	P0022101	0.720
18	P0101101	P0022120	0.617	36	P1102110	P1000020	0.691				
	Set 1				**Set 2**				**Set 3**		

Table 6. The seven distinct personas developed after feature engineering and encoding through the previous steps. Among these, three were representative of the sets of correlated profiles, while four represent the distinct profiles having no significant correlation (defined as >0.5) with any other profile.

Encoding	n	Attributes	Name
P0022101	2	A young man of Asian ethnicity and a good eye condition, having a N-chronotype and a high photosensitivity, who typically doesn't consume caffeine	Mahendra
P0102101	5	A young woman of European ethnicity and a good eye condition, having a N-chronotype and a high photosensitivity, who typically doesn't consume caffeine	Aurora
P0102211	24	A young woman of European ethnicity and a good eye condition, having a E-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Julia
P0102111	1	A young woman of European ethnicity and a good eye condition, having a N-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Cora
P0121220	1	A young woman of Asian ethnicity and a moderate eye condition, having a E-chronotype and a low photosensitivity, who typically consumes caffeine	Dongmei
P1002010	1	A middle-aged man of European ethnicity and a good eye condition, having an M-chronotype and a high photosensitivity, who typically consumes caffeine	Johannes
P1102111	1	A middle-aged woman of European ethnicity and a good eye condition, having a N-chronotype and a moderate photosensitivity, who doesn't consume caffeine	Elizabeth

Table 7. Metrics of accuracy for each of the eight supervised machine learning models developed in the study

Rank	Model	MAE (‘*ClusAvg’*)	R^2 (‘*ClusAvg’*)	MAE (‘*ClusRep’*)	R^2 (‘*ClusRep’*)
1–2	XGBoost	8.560 ± 0.534	0.631 ± 0.084	10.231 ± 0.461	0.716 ± 0.017
1–2	CatBoost	8.305 ± 0.497	0.707 ± 0.034	10.497 ± 0.336	0.705 ± 0.012
3	LightGBM	9.972 ± 0.696	0.608 ± 0.063	12.316 ± 0.392	0.620 ± 0.022
4	Random Forest	10.243 ± 0.740	0.579 ± 0.066	13.015 ± 0.538	0.570 ± 0.029
5	Gradient Boost	10.561 ± 0.792	0.573 ± 0.090	14.918 ± 0.475	0.470 ± 0.021
6	Bayesian Ridge	13.313 ± 0.655	0.394 ± 0.129	18.905 ± 0.821	0.180 ± 0.036
7–8	ElasticNet	14.847 ± 0.959	0.300 ± 0.036	19.981 ± 0.685	0.120 ± 0.011
7–8	Support Vector	16.386 ± 1.696	0.127 ± 0.037	19.431 ± 0.811	0.160 ± 0.019

Table 8. Predictions for the reference set of attributes, and those evaluated after altering one predictor each instance.

Persona	Step	Aurora	Cora	Dongmei	Johannes	Mahendra
VAS	0	49.73	49.29	54.3	55.74	49.95
Action: changing ‘caffeine consumption’ from reference (10.5) down to 0.5						
VAS	1	54.89	54.45	59.4	61.45	55.12
Action: *[Previous attributes]* + Changing ‘fatigue’ from reference (3) to 6						
VAS	2	38.52	37.36	42.85	44.48	42.93
Action: *[Previous attributes]* + Changing ‘Tense’ from reference (2) to 5						
VAS	3	36.15	35.45	40.91	42.67	37.42
Action: *[Previous attributes]* + Changing ‘Sad’ from reference (2) to 4						
VAS	4	35.61	34.99	40.41	42.14	36.88
Action: *[Previous attributes]* + Changing Scale from ‘Disc’ to 'Tme'						
VAS	5	40.96	39.54	40.85	42.78	39.8

Table 9. The state of lighting variables optimized to match a target VAS on the discomfort scale at a given time of day, for a defined set of personal and contextual factors.

#	Variable	Optimized state	Possible range (from dataset)
1.	Luminance of source	9084.25	1627 to 21700
2.	Tcp of source	4420.75	2609 to 6304
3.	ΔUV	0.01	-0.0001 to 0.0141
4.	Vertical Illuminance	388.12	336 to 485
5.	CIE 1931 x coord	0.39	0.3159 to 0.4711
6.	CIE 1931 y coord	0.38	0.3545 to 0.4116
7.	CIE 1976 u coord	0.23	0.1908 to 0.2699
8.	CIE 1976 v coord	0.51	0.4818 to 0.5292
9.	s-cone EDI	213.77	85.57 to 400.46
10.	m-cone EDI	347.57	262.93 to 486.99
11.	l-cone EDI	391.03	333.96 to 494.30
12.	Rhodopic EDI	298.97	174.65 to 492.19
13.	Melanopic EDI	281.92	145.86 to 491.43
14.	Circadian Lighting	432.65	291.33 to 655.21
15.	Circadian Stimulus	0.38	0.31 to 0.46