The Effects of Urban Morphology and Environmental Quality on Health Factors

Pietro Boiardi, Matteo Caviglia, Mauro Pellonara
CS-433 Machine Learning, Project II, EPFL, Fall 2025

I. Abstract

Urban morphology and environmental quality (EQ) are important determinants of population health, yet their combined effects are often challenging to identify and further translate into practical urban policies. In this work, we propose a machine-learning–based framework to systematically explore potential associations between neighbourhood-scale morphological characteristics and a set of environmental quality indicators across the City of Geneva. In addition, the framework investigates how these urban and environmental features relate to a series of health risk endpoints derived from the Specchio population health dataset collected in *Geneva* [1]. Supervised models predict cardiovascular, respiratory, mental health, and sleep-related risks using neighborhood-level descriptors, while unsupervised clustering identifies latent urban typologies. As synthetic data was used for the residents' health risk endpoints in this study, the framework is designed to seamlessly accommodate the forthcoming real clinical data. It supports interpretable risk profiling, enabling the data-driven identification of vulnerable urban neighborhoods typologies and the development of policy-compliant interventions.

II. Introduction

Over the past decade, a primary struggle in urban planning has been aligning strict energy transition legislation with the preservation of environmental quality and resident well-being. European Union directives are pushing towards a complete transition to "nearly zero energy buildings", which has subsequently sparked the creation of various national guidelines [2][3]. However, these structural improvements often conflict with living standards, as energy-focused designs and densification can unintentionally worsen negative factors such as urban heat island effects, noise exposure and high air pollution levels [4][5][6].A growing body of evidence indicates that these morphological characteristics are associated with elevated public health risks, including respiratory disorders, sleep disturbances, and adverse mental health outcomes.[7][8]. In this context, Switzerland has launched SWICE, an initiative that aims to place greater emphasis on human well-being and the improvement of living standards while supporting the ongoing energy transition. Within this scope, a recent work sponsored by the EPFL *Laboratory of Integrated Performance In Design* (LIPID) by Lyu et al. employed a multidomain approach and identified 11 distinct neighbourhood typologies through K-means clustering on PCA-reduced urban morphology integrated data from the city of *Geneva* [9]. Building on this preliminary work, the current project is intended to set up machine learning models that can explain the relationship from neighborhood-level environmental and morphological footprints to aggregated health risk indicators, using *Geneva* as the main case study. It is emphasized that the focus is not on individual clinical diagnosis, but rather on modeling population-level risk scores to assess data-driven vulnerabilities.

The scope of the project is twofold, comprising of an initial unsupervised analysis and a subsequent predictive phase.

First, an unsupervised analysis is conducted to investigate the latent structure within the feature space of both Morphological and EQ data. By decoupling the inherent environmental patterns from specific outcomes, this phase establishes a structural baseline for the study.

Secondly, we include the **Health** Data in the analysis. Those available features are aggregated into four interpretable risk scores, while the rest of the dataset is considered as predictor. The ultimate goal of this initiative is to apply our framework on real clinical data from the *Geneva University Hospital*, however the current proof-of-concept only operates on a synthetic, yet structurally identical dataset. This was due to delays on the Laboratory's side in the ethical approval and data transfer process, which did not reach us before the deadline. Thus, our objective for this part has been to enhance the robustness, adaptability and logical correctness of the framework to later extend it effortlessly to the real data as soon as it becomes available.

III. Methods

In this section the methodological workflow will be outlined, being structured as follows:

A. **Exploratory Data Analysis (EDA):** A preliminary assessment of the distribution and structure of the EQ and Morphology datasets.
B. **Data Preprocessing:** The initial procedures and feature engineering techniques applied to prepare the data for computational analysis.
C. **Unsupervised Learning:** The application of dimensionality reduction and clustering algorithms to reveal latent dependencies and typologies within the available data.
D. **Synthetic Data Integration:** The handling of synthetic **Health** indicators to introduce a predictive analysis.
E. **Predictive Modeling:** The development of an automated machine learning workflow to identify, train, and evaluate the most robust supervised models.

A. Exploratory Data Analysis

Initially, we restricted our analysis to the available *Geneva* data, divided into two views: **Morphology** (Y) and **Environmental Quality** (X). Specific feature details are provided in TablesII and III. The dataset's rows correspond to neighborhoods and *NaN* values are not encountered in this dataset, except for a singular row, that was promptly removed and not considered in subsequent steps of the analysis. We assessed the distributions by analyzing histograms and the skewness metrics. The results indicated that the features in X are predominantly normally distributed, while features in Y show significant skewness. This variation is expected, as specific morphological elements, such as *railway_length* and *water_cover*, are absent in many neighborhoods. When looking at the internal correlations, it is evident that most morphological features (Y) don't interact much with each other, except for some spurious cases. However, a small specific group of features in the EQ dataset is highly correlated. This strong redundancy suggests that X can likely be simplified into a smaller number of key components.

B. Preprocessing

Following the initial data exploration, a brief preprocessing pipeline was applied to optimize the features for subsequent steps. To align with epidemiological standards and literature [10], the age variables were discretized into seven bins, ranging from "Early Childhood" (0–6 years) to "Older Adults" (70+ years). This binning procedure helps in improving the interpretability of age-related risks. Categorical variables were processed according to their specific, hand-labeled type: ordinal features were mapped to integer sequences to preserve their hierarchical structure, nominal features were One-Hot Encoded to prevent unwanted ordinal relationships.

C. Unsupervised learning

This framework employs a sequence of unsupervised algorithms to describe the interaction between the EQ and Morphological features.

Principal Component Analysis (PCA) is applied independently to both subsets to evaluate their effective dimensionality and complexity. PCA projects the original data onto a new coordinate system defined by the axes of maximum variance. This step identifies the minimal subspace required to capture the majority of the information in both environmental and morphological domains.

To quantify the linear coupling between the two domains, **Canonical Correlation Analysis** (CCA) is used. Unlike PCA, which maximizes variance within a single dataset, CCA looks for linear combinations of variables from both datasets that are maximally correlated. The objective is to find the projection vectors that maximize the Pearson correlation coefficient. This method isolates specific environmental patterns that strongly vary with specific morphological configurations, and viceversa.

As a next step, Agglomerative Hierarchical Clustering is utilized to identify distinct neighborhood typologies. This bottom-up approach constructs a hierarchy of clusters based on similarity. This simplifies the selection of the optimal number of natural groupings by analyzing the dendrogram structure, ensuring that the identified typologies exhibit high internal consistency and clear separation from one another.

To visualize the latent space of the data and verify the coherence of these clusters, Manifold Learning techniques (such as t-SNE) are employed. This non-linear dimensionality reduction method projects the data into a lower-dimensional space while maintaining local neighborhoods.

D. Synthetic Data Integration

To make up for the unavailability of real medical records, we used a synthetic **Health** dataset (Z) provided by the EPFL LIPID facility. This dataset was created to have the same column structure of the real data: (1) **Sociodemographic variables**, including potential confounders controls such as age, income, and education level; (2) **Respiratory disorders**; (3) **Cardiovascular risks**; (4) **Sleep disorders**; (5) **Mental Health risks**.

Preliminary exploratory analysis (aligned with what is expected) revealed a significant class imbalance: patients with respiratory (2) and cardiovascular (3) risks represent a small minority of the data, whereas sleep (4) and mental health (5) indicators show a broader distribution across the population. To facilitate predictions, these features were processed to define four distinct target variables representing overall **risk**. Each target was normalized to the unit interval $[0, 1]$, where 0 indicates null risk and 1 represents maximum risk. The aggregation strategies were domain-specific and defined as follows:

1) **Respiratory Risk:** Modeled as a binary variable. If any single "at risk" condition is present, the aggregate risk is set to True (1), False (0) otherwise.
2) **Cardiovascular Risk:** Follows the same binary logic as respiratory risk. The presence of any cardiovascular condition results in a positive risk classification (1).
3) **Sleep Disorder Risk:** A continuous score normalized to $[0, 1]$ combining three weighted components that account both for behavioral risk and clinical floor: Gaussian deviation from recommended sleep duration (see Table V), subjective insufficiency, and a seasonal clinical indicator. For more details, see appendix.
4) **Mental Health Risk:** Derived from the GHQ-12 questionnaire scores. The raw values were mapped to the $[0, 1]$ interval using a custom sigmoid function, based on the literature available [11]. For more details, see appendix.

Because our primary goal is to build a robust predictive framework, we managed the imbalances to prevent the models from becoming biased toward the majority cases. Classification tasks utilized methods like SMOTE and class weighting to amplify the signal of minority groups, while regression tasks employed target stratification to ensure the models learned effectively from the full range of values.

Finally, we included a comprehensive feature selection pipeline to automatically remove irrelevant features from X and Y to manage data redundancy. We start by screening

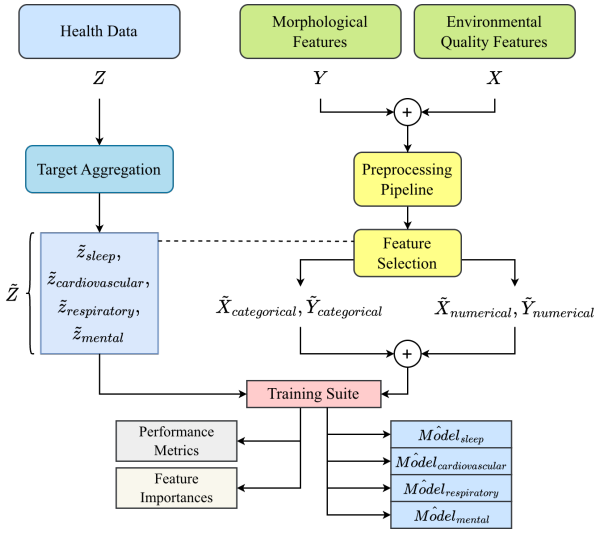

Fig. 1: Visual schema of the proposed Predictive workflow.

for potential predictive power onto the newly introduced target variable Z using statistical tests specifically aligned with the data types: *ANOVA* and *Chi-square* are employed to evaluate categorical variables, while *correlation* and *regression metrics* are applied to continuous ones. To avoid discarding useful information, especially considering the limited amount of available features, we use permissive thresholds during this screening, ensuring that even weak signals that might contribute to the prediction are retained.

E. Predictive Modeling

A suitable modeling suite is set up to dispatch on the target type (regression or classification depending on the health aggregated endpoint) and trains a collection of ML algorithms with nested 5-fold cross-validation and grid search. The suite automatically handles progressive exploration and refinement of hyperparameters and compares a rich combination of ML models with each other, returning the most effective one. The dataset is then divided in train/test split, with a ratio of $80\%/20\%$. For a high-level visual representation of the proposed workflow, look at Figure 1.

a) Regression Pipeline: For continuous outcomes, five regressors are evaluated (see Table VII for further details). The pipeline undergoes the CV targeting for the best Mean Square Error (MSE). Stratified splitting is carried forward, via quantile bins to preserve proportions in train/test folds in regression-ready datasets. To avoid overfitting the grid, a two-stage computationally advantageous refinement strategy is employed: after identifying the best hyperparameters of regularization strength, a second narrower and finer grid is searched in the surrounding of the previous CV best. This balances exploitation and exploration, allowing fine-tuning without excessive effort. The best estimator from each model is refitted on the full training set, and coefficients or feature importances are collected. Finally, the model performing best among all the optimal ones is picked as the one with the lowest test-set Root Mean Square Error (RMSE). Refitted on the complete dataset, the estimator is ready for inference.

b) Classification Pipeline: Similar criteria apply when six classifiers are trained (see Table VI for details). Class imbalance is addressed via scaling loss contributions inversely to class frequencies. The scoring metric used during CV is Recall, while the two-stage grid search refinement mirrors the regression case. The metric guiding the choice of the best-performing classification model among the optimal ones is the F1-score. After training, an additional threshold sweep step is applied to the best model, and maximization of the F1-score drives the automated choice of the adopted threshold. Coefficients (normalized) of logistic models and importances are extracted for interpretability and diagnostics to check main drivers.

c) Risk Inference and Typology Aggregation: Lastly, risk inference and typology aggregation are performed. With trained models in hand, the cleaned morphology dataset is cross-joined with all combinations of sociodemographic variables (see Table IV). Each best model predicts a risk score for every neighbourhood-demographic combination, originating a dense risk matrix. Therefore, per-neighbourhood averaged profiles are achieved to attain a spatial risk mapping, and per-typology summaries are returned when risks are grouped by urban typology. Z-scores are calculated to flag outlier typologies for elevated or protective profiles. Importantly, predictions are averaged across residents and sociodemographic variables are used as controls to verify that the observed risk patterns are not demographic-imbalance-driven artifacts. The separation between training-time adjustment and inference-time aggregation is key to avoiding discriminatory and unfair use.

IV. Results and Discussion

A. Unsupervised Results

The unsupervised results allow us to provide some insights between Environmental Quality (X) and Urban Morphology (Y) in the dataset.

The **Principal Component Analysis (PCA)** results showcase how a high degree of redundancy is present in the data. For the Environmental Quality dataset, just 4 principal components are able to explain 95.52% of the total variance, while for the morphological data we need to reach 7 components to capture 96.17% of the variance. This step of dimensionality reduction agrees with the initial observation about collinearity in the EQ data split and it suggests that the data's underlying patterns can be summarized by a limited set of latent factors.

The CCA analysis revealed significant links connecting the two splits of the dataset, with the first canonical dimension achieving a correlation coefficient of 0.64, followed by a second dimension of 0.41. These coefficients highlight a robust intrinsic relation between the morphological structures and environmental features, further confirming the lab's findings [9]. Furthermore, the CCA illustrated a significant relationship associating environmental quality indicators and urban morphology across *Geneva* neighborhoods (see Figure 2).

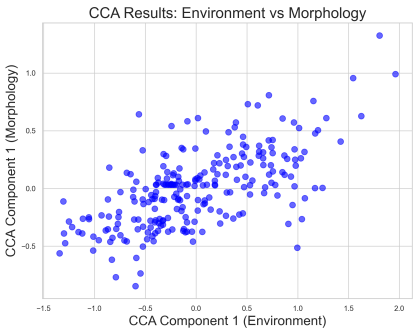

Fig. 2: Environment vs Morphology data with CCA Analysis.

The first pair of canonical variables exhibited a clear positive correlation, showing the presence of a dominant joint environmental-morphological axis shared across spatial units. On the environmental side (see Figure 3), the first canonical component was primarily driven by land surface temperature, elevated air pollutants and larger noise exposures, while solar irradiation loaded negatively. The pattern highlights the consolidated profile typical of urban *heat islands* effect and traffic-related pollution. The corresponding morphological canonical component was shaped by high building cover fraction, increased street intersections, high frontal area index and negative pervious surface fractions loadings, hence outlining a compact and infrastructure-heavy urban class. The strong alignment underscores that urban form is indeed a key structural determinant of environmental exposure, and the findings encapsulated by CCA provide empirical support for the neighborhood-level modeling and risk stratification strategy currently adopted.

B. Predictive Results

Given the temporary usage of a synthetic dataset for the sociodemographic and health data of *Geneva*'s inhabitants for the development of the workflow, sound conclusions cannot be drawn from the classification and regression performances. Although the pipeline is ready to be fed with real-world data, interpretation of the results at the moment should require caution and could be highly misleading. The current health dataset, provided by EPFL's LIPID supervisors to simulate real data source architecture, assigns health related outcomes in an almost random fashion, hence preventing the models from learning patterns in morpho-environmental conditions that allow inference.

In this instance, our main contribution is the ML pipeline, tailored to be directly applicable to real clinical data. All preprocessing, target construction, and modeling steps are modular, data-driven, robust to missing values and distributional changes. The system automatically adapts to binary or continuous targets and performs feature screening and selection, handling multicollinearity and yielding separate results depending on health aggregated risk score and the predictive task.

Considering the nature of the targets, the ridge logistic regression achieved a F1-score of 0.2606 in the respiratory risk prediction task, which holds good promise for future transition to *Geneva*'s real clinical data. Moreover, an attempt for an ablation study, shown in I, was conducted to evaluate incremental improvements brought by successive evolutions of the ML pipeline. In particular, the marginal additive effects of one-hot-encoding (OHE), feature selection (FS), class balancing (CB), refined grid search for hyperparameter tuning (HPT) were assessed. To be coherent with the model selection criteria, incremental improvements are evaluated in terms F1-score for classification and RMSE for regression.

TABLE I: Ablation Study Results on Synthetic Data

Task / Metric	Base	OHE	+FS	+CB+FS	+CB+FS+HPT
CVD Risk (F1)	0.1798	0.1798	0.1924	0.1818	**0.1867**
Resp. Risk (F1)	0.2581	0.2581	0.2587	0.2606	**0.2606**
Sleep Risk (RMSE)	0.2433	0.2433	0.2432	0.2432	**0.2432**
Mental Risk (RMSE)	0.3218	0.3218	0.3217	0.3217	**0.3217**

The ablation study demonstrates a consistent trend of marginal improvement in classification performance. However, since the synthetic targets were defined solely by specifying independent distributions rather than through correlations, these observations (including the mixed results from class balancing) reflect the model fitting on statistical noise rather than genuine data features. Similarly, the flat regression performance confirms the low-signal carried by the synthetic data. However, none of the steps led to stability collapse; this confirms that the pipeline is robust and possibly capable of extracting effective gains once the real clinical data is introduced.

V. Conclusion

This project presents a flexible, transferable and reproducible machine learning framework designed to quantify relationships between urban morphology, environmental quality and aggregated public health risk indicators at the neighborhood scale. Through the integration of domain-informed health target scores, rigorous feature screening, supervised modeling and complementary unsupervised clustering, the workflow is capable of enabling policy-ready risk assessment across urban typologies. From a reproducibility perspective, all steps are deterministic, well-documented and parametrized. This largely facilitates replication across cities or datasets and therefore the versatility positions the proposed pipeline as a reusable decision-support tool for sensitive urban planning with the integration of future clinical cohorts.

The model could also serve the purpose of scenario simulation and counterfactual analysis to evaluate how predicted risk would change upon modification of input features (e.g. increasing pervious surface coverage). Although the current study relies on a synthetic health dataset and therefore does not permit definitive epidemiological conclusions, it serves as a robust proof of concept. Once applied to such data, the model has the potential to identify urban typologies with elevated health risks, and hence to contribute to design of more resilient cities.

VI. Ethical Risk Assessment

This project will soon transition from synthetic data to real clinical data. Consequently, the introduced risks are much

more sensitive and they were explicitly considered from the early design phase, guided by the Digital Ethics Canvas.

A first key ethical risk concerns **privacy** and potential **re-identification**. The workflow is designed to integrate sensitive health and sociodemographic attributes, which could expose individuals if improperly handled. To mitigate this risk, strict anonymization protocols are required, informed consent regarding data usage must be ensured, and no personally identifiable information (PII) should be used, stored, or inferred. The pipeline operates at the neighborhood level, avoiding individual predictions, and age variables are binned to further reduce re-identification risk. Sociodemographic variables are included cautiously as confounding controls. In future iterations, their use should rely on coarse-grained representations (e.g., binned values or neighborhood-level proportions), while model performance should also be evaluated across demographic *strata* to monitor potential bias.

A second ethical concern involves the possible misinterpretation or **misuse** of neighborhood-level health risk predictions. If probabilistic outputs were treated as deterministic or causal statements, this could result in stigmatization of specific areas, unfair resource allocation, or unjustified policy interventions. These outcomes raise fairness and welfare-related risks. Although the potential severity of such impacts could be high, their likelihood is assessed as moderate to low, as model outputs are explicitly framed as population-level vulnerability indicators rather than causal claims.

It is also essential to acknowledge potential biases likely to be present in this modeling approach. First, the study is subject to a **Geographic Bias**. In fact, the findings are characteristic of the dynamics of the **City of Geneva** and may not generalize to other urban contexts. Second, a **selection bias** is likely present in the hospital-sourced data. As the dataset only includes patients who could access and afford treatment at the specific facility, individuals who cannot afford care or who are treated by other hospitals are not represented, making the work domain-limited.

VII. Source Code and Data Availability

The original acquired morphological data used in this work is publicly accessible at the following link and can be found in the open-source repository containing all the code used in this study.

References

[1] Helene Baysson et al. “Development of a web-platform for dynamic monitoring of population health in Geneva: Specchio project”. In: *European Journal of Public Health* 30.Supplement_5 (2020), ckaa165–440.

[2] Marina Economidou et al. “Review of 50 years of EU energy efficiency policies for buildings”. In: *Energy and buildings* 225 (2020), p. 110322.

[3] Effrosyni Giama, Elli Kyriaki, and Agis M Papadopoulos. “Energy policy and regulatory tools for sustainable buildings”. In: *IOP Conference Series: Earth and Environmental Science*. Vol. 410. 1. IOP Publishing. 2020, p. 012078.

[4] Marta Salgado et al. “Environmental determinants of population health in urban settings. A systematic review”. In: *BMC public health* 20.1 (2020), p. 853.

[5] Siavash Ghorbany et al. “Data driven assessment of built environment impacts on urban health across United States cities”. In: *Scientific Reports* 15.1 (2025), p. 19998.

[6] Ingrid Pelgrims et al. “Association between urban environment and mental health in Brussels, Belgium”. In: *BMC public health* 21.1 (2021), p. 635.

[7] James Woodcock et al. “Quantitative Health Impact Assessment of Environmental Exposures Linked to Urban Transport and Land Use in Europe: State of Research and Research Agenda”. In: *Current environmental health reports* 12.1 (2025), pp. 1–23.

[8] Lisa S Barsties et al. “Residents’ perceptions of urban densification and health–a systems dynamics approach”. In: *BMC Public Health* 25.1 (2025), p. 1473.

[9] Kun Lyu et al. “A multidomain approach to neighbourhood typology for urban environmental studies”. In: *Sustainable Cities and Society* (2025), p. 106378.

[10] Max Hirshkowitz et al. “National Sleep Foundation’s updated sleep duration recommendations”. In: *Sleep health* 1.4 (2015), pp. 233–243.

[11] David P Goldberg et al. “The validity of two versions of the GHQ in the WHO study of mental illness in general health care”. In: *Psychological medicine* 27.1 (1997), pp. 191–197.

APPENDIX

A. Relevant Tables

TABLE II: Environmental Quality (EQ) Features and Data Sources

Feature	Description	Unit	Data Source
lst_mean	Mean land surface temperature	°C	Landsat satellite imagery
solar_summ	Summer solar radiation exposure	$\mathrm{kWh\,m^{-2}}$	Urban solar radiation model
solar_wint	Winter solar radiation exposure	$\mathrm{kWh\,m^{-2}}$	Urban solar radiation model
pm10_mean	Mean PM10 concentration	$\mu\mathrm{g\,m^{-3}}$	NABEL air quality network
pm25_mean	Mean PM2.5 concentration	$\mu\mathrm{g\,m^{-3}}$	NABEL air quality network
no2_mean	Mean nitrogen dioxide (NO_2) concentration	$\mu\mathrm{g\,m^{-3}}$	NABEL air quality network
noiseday_m	Average daytime noise level	dB(A)	SONBASE noise database
noisenight	Average nighttime noise level	dB(A)	SONBASE noise database

TABLE III: Urban Morphology variables

Category	Parameter	Unit	Data Source
Urban block geometry			
Urban block geometry	Building height	m	LiDAR DEM / DTM
	Sky view factor	–	LiDAR DEM / DTM
	Frontal area index	–	SwissTLM3D, LiDAR
Canyon geometry			
Canyon geometry	Aspect ratio	–	SwissTLM3D
	Canyon length	m	SwissTLM3D, LiDAR
	Canyon intersection	count	SwissTLMRegio
	Street orientation	degrees	SwissTLMRegio
	Street width	m	SwissTLMRegio
	Street hierarchy	categorical	SwissTLMRegio
Covered land geometry			
Covered land geometry	Impervious surface fraction	–	Land Use Statistics NOLC04
	Building surface fraction	–	Land Use Statistics NOLC04
	Pervious surface fraction	–	Land Use Statistics NOLC04
	Water surface fraction	–	Land Use Statistics NOLC04

TABLE IV: Sociodemographic Variables

Variable	Type	Categories / Values
sex	Categorical	M, F
income	Ordinal	<50k, 50–100k, 100–150k, >150k
education_level	Ordinal	Primary, Secondary, Tertiary
age_bin	Ordinal	Age-group categories

TABLE V: Recommended Sleep Duration by Age Group [10]

Age Group	Expected Sleep Hours
Early Childhood (0–6y)	12
Children (6–12y)	10
Teenagers (12–18y)	9
Young Adults (18–30y)	8
Adults (30–50y)	8
Middle-Aged Adults (50–70y)	8
Older Adults (70+y)	7

Sleep and circadian disorder risk is modeled as a continuous score $R_{\text{sleep}} \in [0, 1]$ combining behavioral and clinical components:

$$R_{\text{sleep}} = R_{\text{floor}} + (1 - R_{\text{floor}}) \left[\alpha \left(1 - \exp\left(-\frac{(H - H_{\text{exp}})^2}{2\sigma^2} \right) \right) + \beta D \right], \tag{1}$$

where H denotes observed sleep duration (hours), H_{exp} is the age-specific expected duration, D is the normalized sleep deprivation score, $\sigma = 2$ h, $\alpha = 0.4$, $\beta = 0.3$, and

$$R_{\text{floor}} = \begin{cases} 0.8, & \text{if a clinical sleep disorder is reported,} \\ 0, & \text{otherwise.} \end{cases}$$

This formulation yields a smooth risk continuum while enforcing a minimum baseline risk for clinically diagnosed sleep disorders.

Mental health risk is modeled as a probability-like continuous score derived from the GHQ-12 questionnaire:

$$R_{\text{mental}} = \frac{1}{1 + \exp(-k\,(GHQ - \tau))}, \tag{2}$$

where GHQ is the GHQ-12 score, $\tau = 4$ corresponds to the established clinical screening threshold, and $k = 0.8$ controls the slope of the transition. This mapping provides a calibrated and robust outcome for regression-based modeling.

B. Hyperparameter Optimization and Model Selection

1) Two-Stage Cross-Validated Grid Search: For each candidate model m, hyperparameter optimization is performed using a two-stage grid search procedure with cross-validation. Let $\Theta_m^{(0)}$ denote the initial hyperparameter grid associated with model m. The first stage selects the optimal configuration by maximizing the expected cross-validation performance:

$$\theta_m^* = \arg\max_{\theta \in \Theta_m^{(0)}} E_{k=1}^{K} \Big[\mathcal{M}\big(f_m(X_{\text{train}}^{(k)}; \theta), y_{\text{val}}^{(k)}\big) \Big], \tag{3}$$

where K is the number of folds, f_m is the model with hyperparameters θ, and $\mathcal{M}$ is the optimization metric (recall for classification and negative mean squared error or R^2 for regression).

A refined grid $\Theta_m^{(1)}$ is then constructed by locally perturbing each selected hyperparameter:

$$\Theta_m^{(1)} = \{\theta_i : \theta_i \in \{0.8\theta_i^*,\, 0.9\theta_i^*,\, \theta_i^*,\, 1.1\theta_i^*,\, 1.2\theta_i^*\}\}\,. \tag{4}$$

A second grid search is performed over $\Theta_m^{(1)}$ using the same cross-validation strategy. The resulting estimator is evaluated on a held-out test set and retained as the final trained model.

2) Classification Models: Binary health targets (cardiovascular and respiratory risk) are modeled using six classifiers. Table VI summarizes the initial and refined hyperparameter grids (regularization strength for L1 and L2), otherwise nearest neighbors number (k-NN), maximum depth and estimators number (for random forest)

TABLE VI: Classification models and hyperparameter grids

Model	**Initial Grid**	**Refined Grid**
Logistic Regression (L1/L2)	$C \in \{0.01, 0.1, 1, 10\}$	$C^* \times \{0.8, 0.9, 1, 1.1, 1.2\}$
Random Forest	$n \in \{100, 300\}$, $d \in \{\text{None}, 10, 20\}$	$n^* \pm \{10, 20\}$, $d^* \pm \{1, 2\}$
SVM (Linear)	$C \in \{0.01, 0.1, 1, 10\}$	$C^* \times \{0.8, 0.9, 1, 1.1, 1.2\}$
SVM (RBF)	$C \in \{0.1, 1, 10\}$, $\gamma \in \{\text{scale}, \text{auto}\}$	Local refinement of C
k-NN	$k \in \{3, 5, 7\}$	$k^* \pm 2$

Model selection is driven by recall, prioritizing sensitivity to high-risk cases. Class imbalance is addressed through class weighting or resampling strategies when required.

3) Regression Models: Continuous health risk scores (sleep and mental health) are constrained to the $[0, 1]$ interval. All regressors are wrapped in a transformed-target formulation:

$$\tilde{y} = \log\left(\frac{y}{1 - y}\right), \qquad \hat{y} = \sigma(\tilde{y}), \tag{5}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function. This guarantees valid probabilistic predictions while enabling standard regression techniques.

TABLE VII: Regression models and hyperparameter grids

Model	**Initial Grid**	**Refined Grid**
Linear Regression (Ridge / Lasso)	$\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$	$\alpha^* \times \{0.8, 0.9, 1, 1.1, 1.2\}$
Kernel Ridge	$\alpha \in \{0.1, 1, 10\}$	Local refinement of α
Random Forest	$n \in \{100, 300\}$, $d \in \{\text{None}, 10, 20\}$	$n^* \pm \{10, 20\}$, $d^* \pm \{1, 2\}$
SVR (RBF)	$C \in \{0.1, 1, 10\}$, $\gamma \in \{\text{scale}, \text{auto}\}$	Local refinement of C, γ
k-NN	$k \in \{3, 5, 7\}$	$k^* \pm 2$

Regression performance is evaluated using RMSE, MAE, and R^2. Residuals are retained for post-hoc diagnostics and spatial analysis.

4) Rationale: This two-stage optimization strategy enables efficient global exploration followed by localized refinement, ensuring stable hyperparameter selection while limiting computational cost. The approach is consistent across health outcomes and supports downstream neighborhood-level inference.

C. Additional Figures

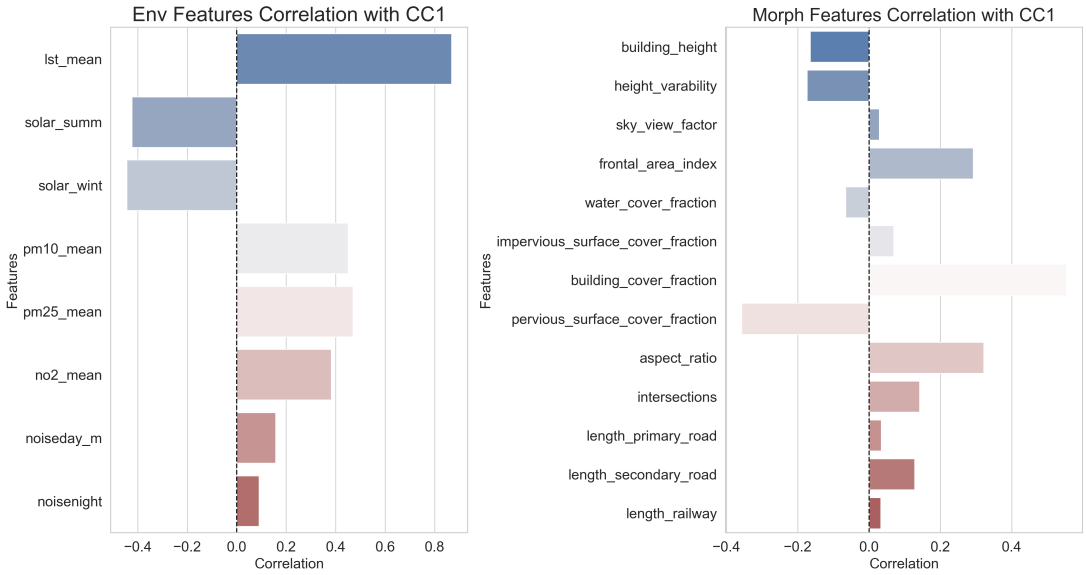

Fig. 3: Feature Importances of the Environmental Quality (left) and the Morphological (right) data, according to the CCA.