

RESEARCH ARTICLE

Data generative machine learning model for the assessment of outdoor thermal and wind comfort in a northern urban environment



Nasim Eslamirad ^{a,*}, Francesco De Luca ^b,
Kimmo Sakari Lylykangas ^b, Sadok Ben Yahia ^c

^a *FinEst Centre for Smart Cities, Tallinn University of Technology, Tallinn 10319, Estonia*

^b *Department of Civil Engineering and Architecture, Tallinn University of Technology, Tallinn 10319, Estonia*

^c *Department of Software Science, Tallinn University of Technology, Tallinn 10319, Estonia*

Received 13 October 2022; received in revised form 26 November 2022; accepted 10 December 2022

KEYWORDS

Urban climate;
Outdoor thermal and
wind comfort;
Predictive model;
Data generative
model;
Machine learning
approach

Abstract Predicting comfort levels in cities is challenging due to the many metric assessment. To overcome these challenges, much research is being done in the computing community to develop methods capable of generating outdoor comfort data. Machine Learning (ML) provides many opportunities to discover patterns in large datasets such as urban data. This paper proposes a data-driven approach to build a predictive and data-generative model to assess outdoor thermal comfort. The model benefits from the results of a study, which analyses Computational Fluid Dynamics (CFD) urban simulation to determine the thermal and wind comfort in Tallinn, Estonia. The ML model was built based on classification, and it uses an opaque ML model. The results were evaluated by applying different metrics and show us that the approach allows the implementation of a data-generative ML model to generate reliable data on outdoor comfort that can be used by urban stakeholders, planners, and researchers.

© 2022 Higher Education Press Limited Company. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The world is experiencing rapid urbanization (Steemers and Yannas, 2000; Luo et al., 2017; Li et al., 2016). Thus, users

would determine the outdoor place's vibrancy, local economy, sustainability (Teoh et al., 2020), quality of life, and vitality of the city. Moreover, proper outdoor conditions invite people to be active in urban spaces (Gehl, 2011), and

* Corresponding author.

E-mail address: nasim.eslamirad@taltech.ee (N. Eslamirad).

Peer review under responsibility of Southeast University.

urban planners are challenged to find solutions that ensure comfort in cities (Steemers and Yannas, 2000; De Luca et al., 2018, 2021). Urban planners are constantly striving to improve the quality of life of urban dwellers (Luo et al., 2017) through various design concepts and solutions for liveable and pleasant public spaces (Kim et al., 2018). Studies on building configurations can guide decision-making to improve both existing urban areas and new developments to increase the level of outdoor thermal comfort. Recently, several researchers have investigated the factors affecting outdoor thermal comfort and the method for assessing thermal comfort (Wang et al., 2020; Lai et al., 2014; De Freitas and Grigorieva, 2015; Jamei and Rajagopalan, 2019; Chen and Ng, 2012; Lawson, 1978; Cohen et al., 2013).

Although urban planners and researchers are very interested in assessing the quality of urban open spaces (Nikolopoulou and Lykoudis, 2006), there is still a shortcoming in the existing outdoor thermal comfort studies, such as the low efficiency of simulations and incapability to apply the result of outdoor thermal comfort assessment in the initial steps of the design process. Therefore, there is unexplored potential in investigating optimal building morphologies for outdoor comfort to support the implementation of conscious design and planning practices (Marshall and Çalişkan, 2011). The ML approach is a powerful method for capturing non-linear relationships between independent and dependent variables and could also be used to map subjective responses to the physical environment (Wang et al., 2020).

Although there have been few studies on outdoor comfort based on Artificial Intelligence (AI) and ML methods, in this paper, we propose a data-driven approach to predict the outdoor comfort level in urban areas based on the results of Computational Fluid Dynamics (CFD) simulation on the sidewalk around buildings in Tallinn by applying various environmental and building feature, and weather data. The current study focuses on building features, the main ones are morphology, area, and height in Tallinn, Estonia (Lat. 59 26'N Lon. 24 45'E), which is characterized by a humid continental climate with warm summers, according to the Köppen-Geiger classification Dfb (Peel et al., 2007).

The present study is novel in two ways: firstly, it aims to create a model to assess outdoor thermal comfort, while few studies consider outdoor thermal comfort in cold climatic conditions. Secondly, the study's methodology can be generalized and applied in other locations with different urban and climatic features since the study is based on using urban data to build predictive and data-generative ML. The input data for the ML process of the current study comes from the modeling and simulation process conducted with three-dimensional building models using the ENVI-met simulation software. ENVI-met calculates the changes in a microclimate around urban structures and open spaces based on the principles of fluid mechanics, thermodynamics, radiation exchanges between urban surfaces, and temperature calculation, as well as the laws of atmospheric physics (Eslamirad et al., 2021; Park et al., 2014; Eslamirad et al., 2022).

The remainder of this paper is organized as follows: the related works are summarized in Section 2. Then, Section 3 thoroughly describes the proposed study's purpose and

methodology. In Section 4, we thoroughly described the machine learning steps and the building of the ML model of the study. Section 5 looks at the results and achievements of the study and finally sketches the study's limitations and future issues. Section 6 focuses on the discussion in which we analyzed, explained, and interpreted the study results. The last section, Section 7, recalls the conclusion and take-home study outcomes.

2. Related works

Urban planning and design influence microclimate factors, which determine people's outdoor thermal comfort (Park et al., 2014). Over the past 30 years, several studies have focused on wind effects in the urban canopy layer by applying field measurements, the morphometric (geometric) method, wind tunnel experiments, and CFD simulations at the district, neighborhood, and building scales (Taleghani et al., 2015; Marshall and Çalişkan, 2011; Javanroodi et al., 2022). Therefore, researchers have developed techniques to assess the environment (Zhang et al., 2020). Taleghani et al. (2015) compared microclimatic conditions allowed by three building forms during June in the Netherlands and showed that courtyards provide the most comfortable microclimate compared to the other forms. Mittal et al. (2018) investigated methods for evaluating wind comfort at the pedestrian level by comparing different comfort criteria that significantly changed near high rise buildings. Indeed, Kim et al. (2018) considered both wind comfort and thermal comfort to assess pedestrian comfort levels and pointed out that the wind speed in the winter season could cause discomfort for pedestrians. In another study, wind speeds and direction were estimated in the ENVI-met environment based on the measured wind data and used to evaluate outdoor thermal comfort (Park et al., 2014). Pancholy et al. (2021) studied pedestrian comfort in the street canyon between parallel buildings and investigated canyon geometry on pedestrian comfort. In another study to help urban planners to measure heat stress in Melbourne, ENVI-met was used for modeling and simulation (Jamei and Rajagopalan, 2019). In 1996, John Gero described AI as a new design process (Tamke et al., 2018). Similarly, ML opens new perspectives in planning (Ramsgaard Thomson et al., 2020). ML tools deal with endowing programs with the ability to learn and adapt (Osisanwo et al., 2017). In the past years, data-driven approaches and ML have succeeded to predict the environmental performance of design solutions in the early design stages (Huang et al., 2022).

Thomsen et al. mentioned that information is advanced due to its predictive power, and urban developments can be modeled through self-organizing systems. They proposed Neural-Network Steered Robotic Fabrication method explores novel path planning approaches for robotic fabrication and it uses Machine Learning (Ramsgaard Thomson et al., 2020).

AI and ML-based methods are fully explored in urban planning. ML planners can evaluate large amounts of data, find patterns or anomalies, and ultimately use them to make better-informed decisions (Tamke et al., 2018). Okhoya (2015) states that ML can be used for a familiar

architectural task. In addition, ML can potentially be used by the engineering community to speed up complex CFD simulations. (Sebestyen and Tyc, 2020) trained an ML model comparing datasets of different complexities and energy rating prediction performance.

In addition, in a new study, a hybrid framework to perform a rapid wind assessment evaluation at the pedestrian level was applied by He et al. (2021). As the authors mentioned there is great potential in the integration of multiple computational tools such as parametric design and machine learning technique in sustainable and environmental research (Javanroodi et al., 2022) deployed a data-driven hybrid model to predict extreme microclimate conditions based on deep neural network (NN) machine learning architecture that was trained using the results of CFD simulation. Another study proposed an automated design process by applying a generative adversarial network (GNA) machine learning approach to achieve the real-time optimization of urban morphology. According to the study, the proposed method has an advantage in computational time comparing than traditional methods of outdoor environment optimization when the number of optimized samples is higher than 174 (Huang et al., 2022). Moreover, the ML approach could also be used to map subjective responses to the physical environment and to identify hidden trends (Wang et al., 2020).

This study analyzed the relationship between factors of the urban environment affecting outdoor thermal and wind comfort. The data of buildings features such as the morphological classes of the buildings, height, density of each building at the site, and orientation were collected from the cadastral map of Tallinn (Tallinn.ee) and used as variables. Then, defined samples of buildings that reflect actual buildings features in Tallinn were used for modeling, simulation, and ML approach through various interactive tools and metrics. This method helped to determine the effective design parameters to improve the outdoor thermal comfort of pedestrians and, consequently, outdoor quality of life even at the early design stage, considering patterns and structures of the urban data without the necessity of complicated and computer-intensive modeling and simulation processes.

3. Materials and methods

This study developed a predictive and data-generative model based on the ML approach to assessing outdoor comfort levels in Tallinn (cf., Fig. 1). The proposed methodology makes it possible to generate a high-accuracy learning algorithm.

3.1. The climatic condition of Tallinn in the studied date and time

The present work is based on the results of previously published research by the authors about the assessment of thermal and wind comfort around buildings in Tallinn (Eslamirad et al., 2021), conducted using the reference weather data of the springtime on 21 March, 10 a.m. in Tallinn (Table 1).

March is the month with the highest wind speeds and one of the coldest of the year (Fig. 2). The day and time were

chosen because they represent a worst-case scenario for thermal and wind comfort in spring during the morning when people are likely to spend some time in outdoor spaces. The rationale was that if people in this scenario are comfortable on the walkways of a particular building morphology, they are likely to be comfortable most of the time in spring.

3.2. Modeling and simulation process

Thanks to advanced computational techniques, the use of CFD is improving microclimate studies (Javanroodi et al., 2022). CFD simulation procedure consists of dominating equations, computational domain, boundary conditions, experimental validation, and mesh convergence analysis (He et al., 2021).

To define the geometric models, 11 morphological classifications were considered based on the most common morphology and shape of the buildings in the city plan of Tallinn: "Aggregated Blocks", "Articulated 1", "Articulated 2", "Block", "C-shape", "Court", "Free Form", "L-shape 1", "L-shape 2", "Linear" and "Series" (Table 2, Fig. 3). Then, the simplified geometries of the morphological classes were modeled in the geometric modeling process.

In the data shuffling process, the samples are defined based on their typical morphology and categorized following existing literature (De Luca et al., 2018). The approach followed in the study is to define building models based on six variables: morphological classes (11 types); several floors (1–19); the number of buildings in a site; density of the built area, floor area ratio per site (500–9500 m²); height (3–57 m); and orientation (0, 20, 40, 60, 80°), as Table 3 shows.

The results of CFD simulation in the ENVI-met environment for thermal and wind comfort were used together to define the target used in the ML process (Eslamirad et al., 2021). Furthermore, Table 4 and Table 5, respectively, show more details about personal parameters and surface features applied in the CFD simulation and PET calculation.

As a result, the area considered for the thermal comfort assessment (green margin in Fig. 4) has a depth of 4 m for the entire building perimeter and a height of 1.1 m.

3.3. Thermal and wind comfort analyses

The data used in the study is related to the results of another published research by the authors (Eslamirad et al., 2021) that assessed outdoor thermal and wind comfort by applying the PET, Physiological Equivalent Temperature (°C) (Table 4) as the thermal comfort index and the LDDC variant of the Lawson comfort criteria (Lawson, 1978; Lawson and Penwarden, 1975; Lawson and Bristol, 1990) through the Wind Comfort Level (WCL) method (De Luca, 2019). Wind comfort was then determined using the LDDC Lawson criteria for pedestrian comfort (Table 7).

The ML method is based on the principle of learning algorithms from the discovered patterns, structures, and correlation of the labelled features and targets in the Python environment to predict outdoor comfort data in an urban context. The raw data set consisted of the attributes as independent variables in the numerical and categorical

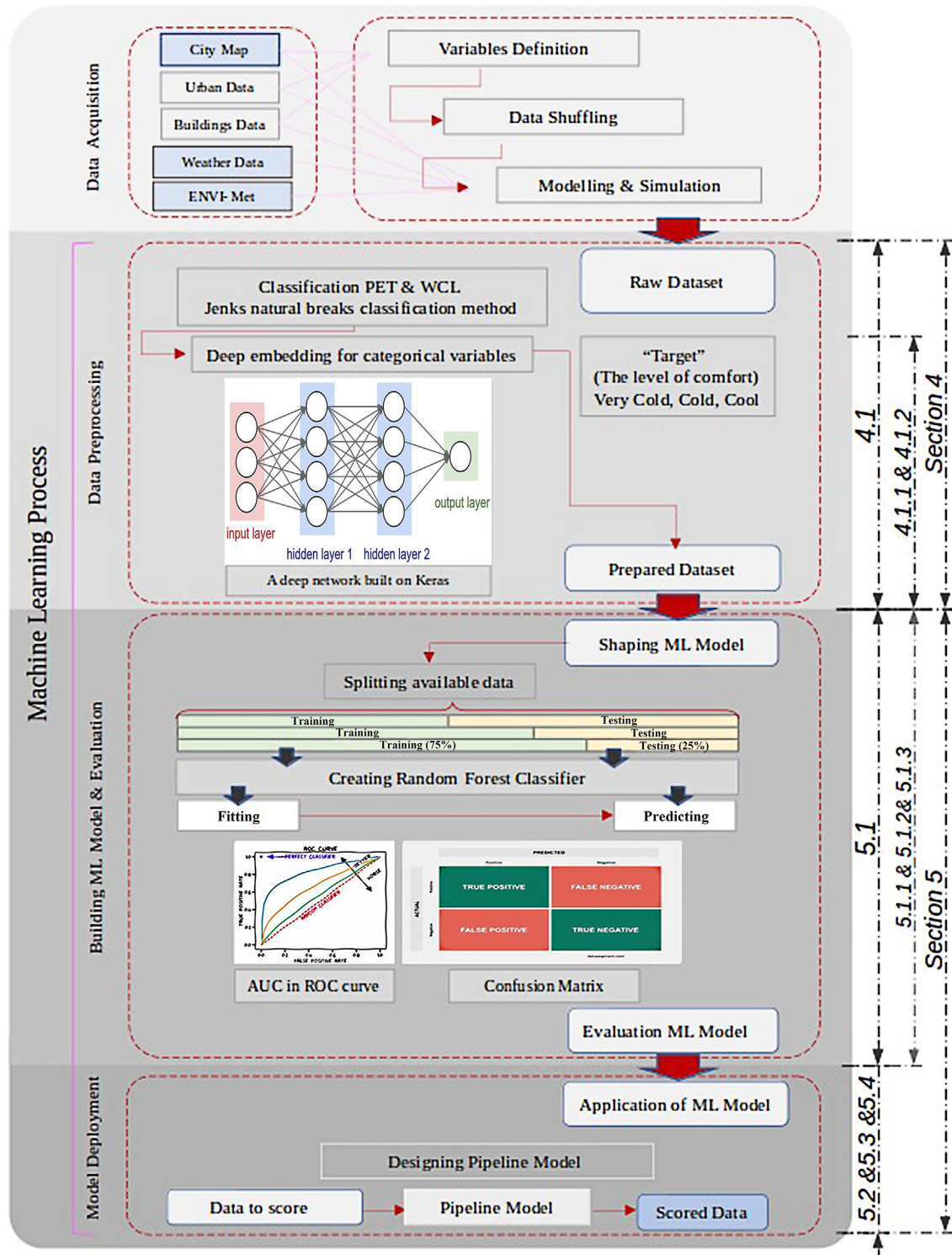


Fig. 1 The workflow developed for the study and the manuscript sections where the workflow steps are presented.

features, the PET thermal comfort metric, and the Wind Comfort Level (WCL).

The first step of the ML process is to import the data into the Python environment as Comma Separated Values (CSV) format file and create the data frame. To define the target in the data set, which indicates the outdoor comfort level of each sample based on thermal and wind comfort, the

Jenks optimizations method was applied. Based on the ranges of PET and WCL, the target comfort values were split into three groups: Very Cold (Target 1), Cold (Target 2), and Cool (Target 3).

The next step is pre-processing, which here the process benefits from the libraries in the Python environment to prepare the dataset for processing through the ML approach.

Table 1 Weather data.

Weather data, 21 March, 10 a.m.	
Wind speed (m/s)	8.40
Wind direction (degree)	202
Dry Bulb Temperature (°C)	6.30
Dew Point Temperature (°C)	2
Humidity (%)	74
Direct Normal Radiation (Wh/m ²)	63

First, the entire data is divided into a training set and a test set to build the ML model. The next is applying the Standard Scaler function, so the variables of the training dataset are in the same range. Then, the Random Forest Classifier trained on the training data, adjusted, and called predict to obtain predictions on the targets. Finally, the classification accuracy is evaluated to check the accuracy of the learning algorithm in the following criteria: Accuracy value, Confusion Matrix, and Area Under the Curve (AUC) in the Receiver Characteristic (ROC) curve.

The last step is building the Pipeline model based on the ML learning algorithm, which comes from the Random Forest Classifier. Then, the Pipeline is trained with the new data from the other dataset. Finally, the prediction function is called only for the pipeline object based on the Random Forest Classifier to make predictions and generate new data for the target column. For example, this study evaluates the comfort level of pedestrians on the sidewalks surrounding the buildings in Tallinn. By applying the ML approach and using urban data, building characteristics, weather data, and thermal and wind comfort, the ML model will be able to generate new trustworthy results about the level of outdoor comfort in the pedestrian level in the city.

4. Machine learning process

The ML process of the study consists of two steps (Fig. 5). First, we trained a model to reach the learning algorithm. The next is using the learning algorithm in Pipeline to predict and score new input data.

The first step consists of four sub-steps to build, evaluate, and check the learning algorithm of ML before building the data generative Pipeline model.

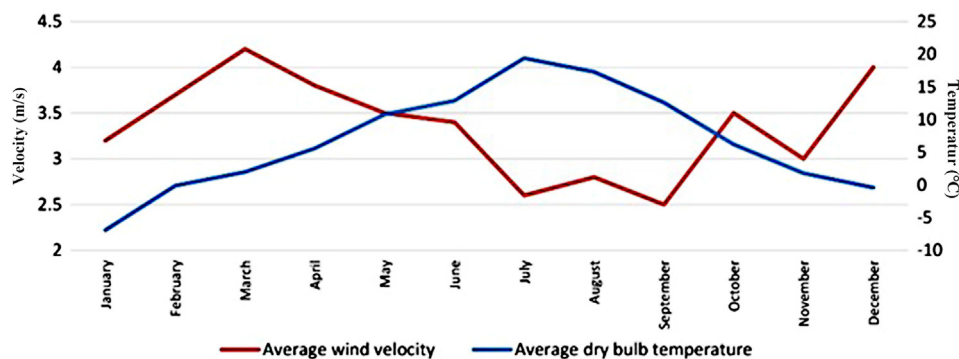
Table 2 The definition and height of the morphological classification of the study.

Morphological class	The explanation of the morphological classes	Height (m)
Block	Compact square or rectangle	3–57
Aggregated blocks	The aggregation of a few or several blocks	3–57
Linear	A long rectangle	3–27
Series	The aggregation of small blocks in the main rectangle	3–27
L-shape	Two rectangles or some small aggregations	3–27
C-shape	Three rectangles with small aggregations	3–27
Articulated	Different rectangles and blocks without a central main part	3–12
Court	A block shape with an open area in the centre	3–18
Free-form	Non-rectangular, curved, aggregated shapes	3–12
Cluster	Separated or articulated buildings	3–12

1. Pre-processing and data exploration
2. Classification and building the ML model
3. Checking the accuracy of the classified data and the ML model

4.1. Data exploration and pre-processing

The first step in designing an ML dataset is finding the most relevant data through pre-processing. Thus, a raw tabular data set is created and characterized by the type of features, their classes, and their range. Because the ML model algorithm requires not only precision, accuracy, and minimum error to make an accurate prediction but also works better with features with higher correlation rates, the dataset needs to be evaluated in the pre-processing step to find and omit the data that are out of range. Therefore, while the feature engineering process is performing, the

**Fig. 2** Wind velocity and monthly temperature averages of Tallinn.

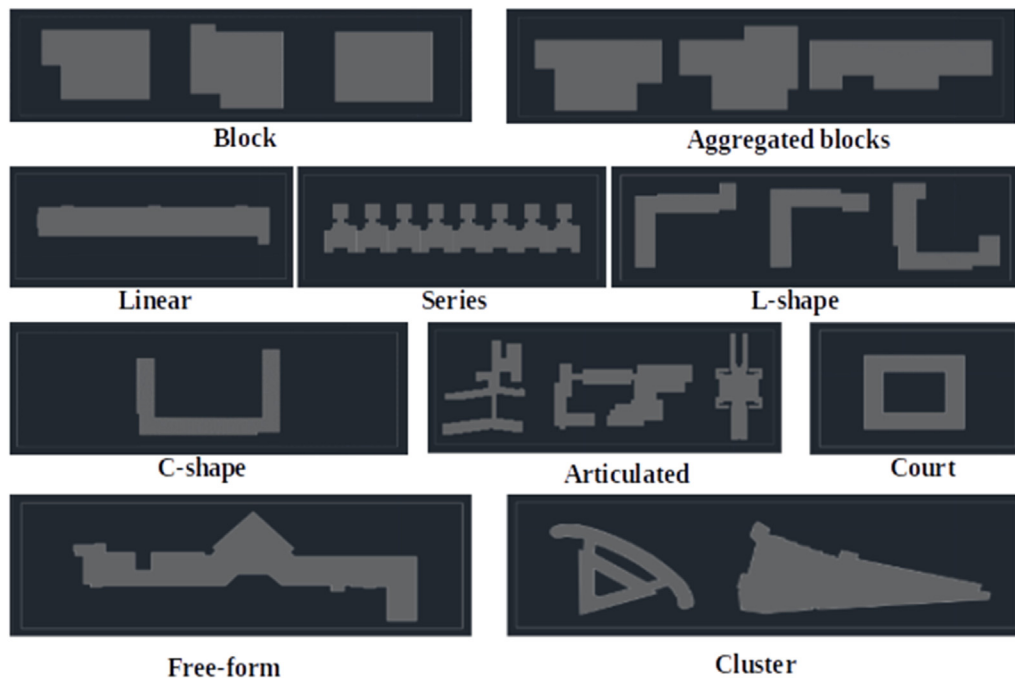


Fig. 3 The morphological classification and models of the study.

Table 3 The definition of samples.

Morphological class	Number of floors	Number of buildings	Density (m ²)	Tallness (m)	Orientation
Block	1–19	1–6	500–9500	3–57	0, 20, 40, 60, 80
Aggregated block	1–19	1–6	500–9000	3–57	0, 20, 40, 60, 80
Linear	1–9	1–6	1000–9000	3–27	0, 20, 40, 60, 80
Series	1–9	1–4	1000–9000	3–27	0, 20, 40, 60, 80
L shape 1	1–9	1–4	1000–9000	3–27	0, 20, 40, 60, 80
L shape 2	1–9	1–4	1000–9000	3–27	0, 20, 40, 60, 80
C shape	1–9	1–6	1000–9000	3–27	0, 20, 40, 60, 80
Articulated 1	1–4	1–6	2000–8000	3–12	0, 20, 40, 60, 80
Articulated 2	1–4	1–6	2000–8000	3–12	0, 20, 40, 60, 80
Court	1–6	1–6	1500–9000	3–18	0, 20, 40, 60, 80
Free form	1–4	1–6	2000–8000	3–12	0, 20, 40, 60, 80

Table 4 Human parameters in PET calculation.

Personal human parameters	
Age of person (year)	35
Gender	Male
Weight (kg)	75
Height (m)	1.75
Static Clothing Insulation (clo)	0.90
Work Metabolism (W)	80
Calculate from walking speed (m/s)	1.21

Table 5 The surface parameter in PET calculation.

Surface feature of buildings	
Roughness	0.010

raw input data was modified and cleaned of outlier data (that falls outside the range) and features with the lowest correlation.

The next step before building the ML model is to categorize the target columns, which indicate the comfort level of each sample based on thermal and wind comfort. Here one of the best range finder algorithms, the Jenks optimization method, is the data clustering method used to determine the best arrangement of values in different classes. Additionally, the Jenks optimization method was applied to classify PET and WCL into three classes. Table 6 shows the defined Targets based on the Jenks optimization method.

- PET class 1: 0.17 °C–2.77 °C, WCL class 1: 50.00%–72.97%
- PET class 2: 2.77 °C–4.5 °C, WCL class 2: 72.97%–87.77%



Fig. 4 The supposed sidewalk area in the simulation process.

- PET class 3: 4.5 °C–11.7 °C, WCL class 1: 87.77%–100%

In the feature selection, two variables, “Orientation” and “Density” were removed from the dataset as they had the lowest correlation with the variable “Target”, which defines the outdoor comfort level on the sidewalks in Tallinn. Fig. 6 gives general information about samples in 4 different Area, Number of buildings(s), Height, and Target in Design Explorer tool. Fig. 6 samples show by filtering according to the area value; different colours represent specific areas 500 m², 1000 m², 1500 m², and 2000 m².

4.1.1. Transformation of the categorical variable

Before the data is transferred to the ML algorithm, the categorical variables of the morphological classes of the buildings are transformed into numerical features that are easier to identify in the ML process.

The transformation has a significant impact on the performance of the model. In this study, embedded layers for Deep Learning were created using Keras, which provides a complete framework. The input is the variable of each morphological class of buildings. This step results in a numerical representation of each morphological class in Tallinn in the 11 dimensions of the input set.

According to (Cerdeira et al., 2018) practical recommendations for encoding categorical features are an efficient choice for capturing resemblance. In this study, the output is a scaled column of the target (the degree of convenience). As categorical features differ from numerical, an operation to define a feature matrix is needed. Machine

learning models that need vector data are applied after a categorical variable encoding to replace the tuple element with feature vectors. Each category is mapped to a particular vector, and the properties of the vector are adapted or learned when training a neural network. The vector space provides a projection of categories so that close correlated categories are naturally grouped. Here, the model has been trained for 50 iterations or epochs. This approach allows us to capture the relationships between categories and define three embedding variables that determine each class of building morphology in the dataset. The final result of the embedding process for each morphological class looks as follows, where each row in three elements represents one class of building morphology (Table 7).

The definition of “Target” classes is based on the classification of PET and WCL, applying the Jenks optimization method.

4.2. ML model building and classification

The learning algorithm finds patterns in the training data that map the attributes of the input data to the target and the results of an ML model that captures these patterns. First, the machine is trained with training data, and then it is checked to see how well it has learned by providing it with a test set. The test set has the same structure as the training set but does not contain any answers (Ramsgaard Thomson et al., 2020).

In this study, the Random Forest algorithm was used, one of the most popular algorithms for regression problems that

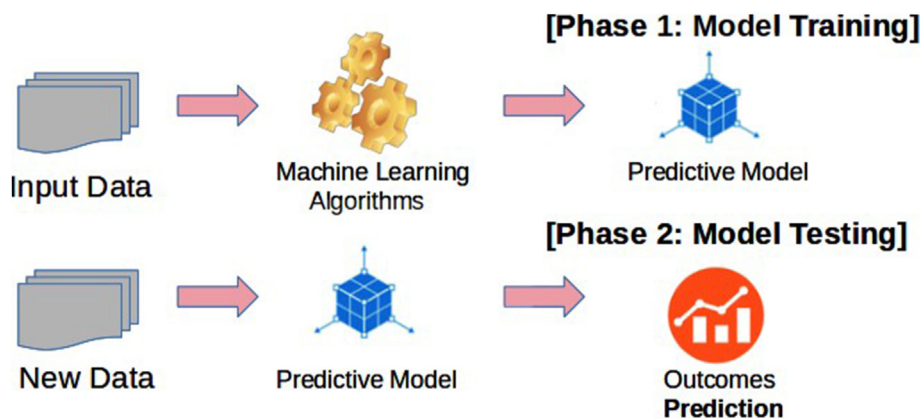


Fig. 5 Machine Learning training and scoring process.

Table 6 The range of thermal index predicted thermal perception by human beings and physiological stress on human beings (Cohen et al., 2013).

PET (°C)	Thermal Perception	Grade of physiological stress
4	Very cold	Extreme cold stress
8	Cold	Strong cold stress
13	Cool	Moderate cold stress
18	Slightly cool	Slight cold stress
23	Comfortable	No thermal stress
29	Slightly warm	Slight heat stress
35	Warm	Moderate heat stress
41	Hot	Strong heat stress
	Very hot	Extreme heat stress

Table 7 LDDC Lawson comfort criteria (Lawson and Penwarden, 1975).

LDDC Lawson comfort criteria		
Category	Comfort activity	Threshold Wind Velocity (m/s)
1	Sitting	4
2	Standing	6
3	Walking	8
4	Business walking/cycling	10
5	Uncomfortable for all activities	>10

work with both categorical and numerical input variables well. Leo Breiman proposed Random Forest in the 2000s to create a predictor ensemble with a set of decision trees that grow in randomly selected data subspaces (Wang et al., 2020).

Like the decision tree, the idea of the Random Forest Regressor comes from the ordinary tree structure, which consists of a root and nodes (the positions where branches

divide), branches, and leaves (Jehad et al., 2012) represented by the segments connecting the nodes.

This study uses the Random Forest Classifier from the Scikit Learn library. From the Sklearn model selection, the train test split function was applied to split the dataset into the train and test subsets. In this study, the classification performance was tested based on the different sizes of training and test datasets. Finally, the best classification performance was obtained with a training size of 75% and a test size of 25%. In addition, we used the stratification parameters in the splitting step to ensure that the training and test datasets contained examples of each target class in the same proportions as in the original dataset.

The next step was normalized and rescaled the training dataset in Python using the Standard Scaler function, so the variables were put in the same range.

Finally, the forest model was externalized by fitting the training subset data and calling predict to obtain predictions. In this way, we fitted and evaluated the model on separate chunks of the dataset by applying the Random Forest Classifier.

5. Results

5.1. Evaluation of classification accuracy

Three methods were applied to assess the accuracy of the classification (Table 8). This subsection describes the application of the methods to assess classification accuracy and the interpretation of the assessment results mentioned in the Results section.

5.1.1. Accuracy score

The classification accuracy score is the first method to determine the accuracy of the built learning model, which measures how many labels got right out of the total number of predictions. This metric evaluation benefits Scikit Learn by using the true labels from the test set and the predicted labels for the test set. In this study, the classification accuracy is 0.69.

5.1.2. Confusion matrix

A confusion matrix, also known as an error matrix expresses how many of a classifier's predictions were correct and, if they were incorrect, where the classifier was confused. It allows visualization of an algorithm's performance and easy identification of confusion between classes, e.g., when one class is frequently mislabelled (website of [geeksforgeeks](https://www.geeksforgeeks.org/)).

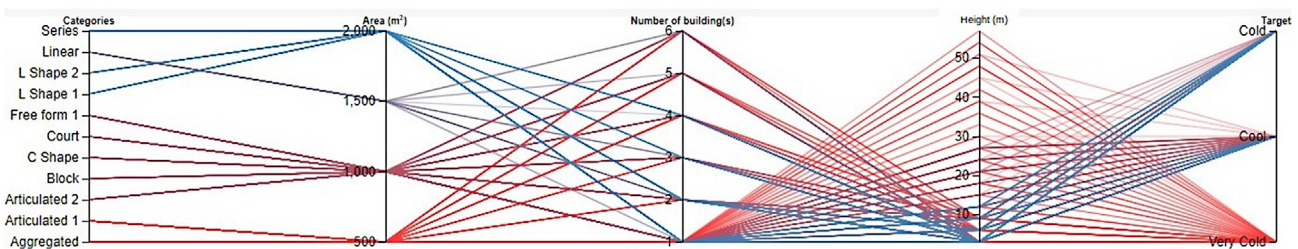
**Fig. 6** The initial data before the categorical data transformation and scaling.

Table 8 The definition of “Target” classes is based on the classification of PET and WCL, applying the Jenks optimization method.

Level of outdoor comfortability		
PET	WCL	“Target”
0.17 °C–2.77 °C	50.00%–72.97%	Target 1 Very Cold
	72.97%–87.77%	
	87.77%–100%	
2.77 °C–4.5 °C	50.00%–72.97%	Target 2 Cold
	72.97%–87.77%	
	87.77%–100%	
4.5 °C–11.7 °C	50.00%–72.97%	Target 3 Cool
	72.97%–87.77%	
	87.77%–100%	

org). The columns of a Confusion Matrix represent the results of the prediction classes, and the rows represent the results of the actual classes. It enumerates all possible cases of a classification problem (Zhang et al., 2020).

A confusion matrix of size $n \times n$ associated with a classifier shows the predicted and actual classification, where n is the number of different classes. Table 9 shows a confusion matrix for $n = 2$, whose entries have the following meanings.

- a is the number of correct negative predictions;
- b is the number of incorrect positive predictions;
- c is the number of incorrect negative predictions;
- d is the number of correct positive predictions.

The prediction accuracy and classification error can be obtained from this matrix (Visa et al., 2011).

$$\text{Accuracy} = (a + d) / (a + b + c + d).$$

$$\text{Error} = (b + c) / (a + b + c + d).$$

In the confusion matrices, the rows represent the true labels, and the columns the predicted labels. The values on the diagonal indicate the number of cases in which the predicted label matches the true label. The values in the other cells represent cases where the classifier mislabelled an observation. The column indicates what the classifier predicted, and the row indicates the correct label.

The Python library Sklearn was applied to calculate the confusion matrix and the results of the classification report. Based on the dataset, 26% of the data in the Target column are Target 1, while Target 2 and 3 account for 36% and 37.5%, respectively.

Fig. 7 shows the confusion matrix plot using labels and the colour scale, with the normalized values of all three targets. The Confusion Matrix shows Target 1 was labelled 86% correctly, i.e., 74 out of 86, while Target 2 (86 out of 120), meaning Cold label, was marked 72%, and Target 3 (68 out of 124) as Cool label was correctly labelled only 55%. Fig. 7 also shows that 14% (12 out of 86) of Target 1 instances were incorrectly labelled as Target 3, and 28% of Target 2 instances (34 out of 120) were considered Target 3 when classified. For Target 3, only 0.48% (less than 1 out of 124) were misclassified as Target 1, while 40% of instances in Target 3 were misclassified as Target 2 (50 out of 124).

5.1.3. AUC in ROC

We also used Area Under the Curve (AUC) in the Receiver Operator Characteristic (ROC) curve, also known as area under the curve, the evaluation metric used to calculate the performance of a binary classifier to find out how accurate the classification is. It is a curve that plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds, separating the “signal” from the “noise”. The AUC value measures a classifier’s ability to discriminate between classes and is used as a summary of the ROC curve. The higher the AUC value, the better the model’s performance in distinguishing between the positive and negative classes. Here, the AUC value in ROC indicates 0.78.

Meaning AUC in ROC shows the classification performances very well.

5.2. Designing pipeline skeleton

The next step is to create a Pipeline since we have done the necessary pre-processing and ML processing. The Pipeline creation process starts with classification and then assessment of the classification (Yahia and Johansson, 2011). Since classification and evaluation of the accuracy of the ML model were completed in the previous steps, next is fitting the algorithm to the new training data and predicting the target values in the three different groups: Very Cold,

Table 9 The embedding categorical morphological class in three sizes.

Name of the Morphological class		Embedding_1	Embedding_2	Embedding_3
1	Aggregated blocks	−0.07856157	0.04777364	0.12004933
2	Articulated 1	−0.09694797	0.12489636	0.140933
3	Articulated 2	0.09129719	−0.09825113	−0.11955576
4	Block	0.07574975	−0.12907775	−0.11605914
5	C Shape	0.05438302	−0.08725278	0.01638534
6	Court	0.03764237	−0.10152158	−0.06313951
7	Free form	0.05259462	−0.07555065	−0.10173931
8	L Shape 1	0.00210845	−0.02964082	−0.02180827
9	L Shape 2	0.1176304	−0.06369058	−0.04275405
10	Linear	0.06750278	−0.02232246	−0.01423943
11	Series	0.07436664	−0.0262874	−0.03133094

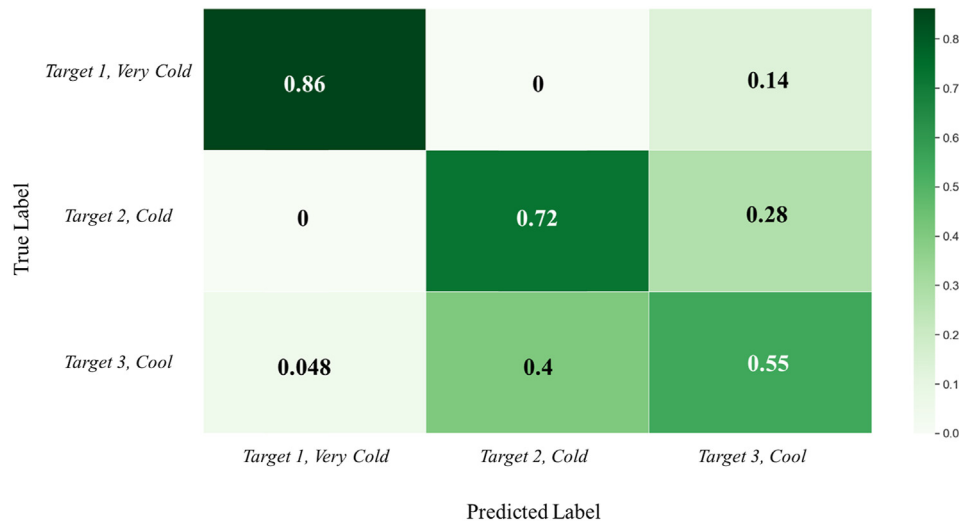


Fig. 7 The representation of the confusion matrix in this study.

Cold, and Cool. The new dataset has the same building features of Table 3, but the Target column is missing. By applying the built learning algorithm and fitting the new training data, Targets will predict through the ML Pipeline. Here the dataset is created with the samples representing the different models in the initial dataset. Therefore, the prediction function is only called for the Pipeline object to make predictions about the targets in the new test data. In this step, the new test data must first be entered into the Python environment as a CSV file using the Pandas library. Then, about 10,500 new target variables were predicted based on the learning algorithm and the model capability of ML Pipeline. The target variables correspond to the ML model in three classes representing the comfort quality of the sidewalks around each building in Tallinn in the ranges from Target 1 to Target 3, i.e., the sidewalks with different comfort levels around the building models.

The Pipeline ML model reads the data from the new dataset as training data and calls the learning algorithm based on the Random Forest Classifier. Finally, the highly accurate target values were saved in an Excel CSV file and added as the last column of the test data with about 10,500 building models showing the comfort value of the sidewalks of the building models in Tallinn based on the predictive and data-generative ML model.

5.3. The evaluation of the model

In classification problems, good classification accuracy is the primary concern (Visa et al., 2011). The accuracy of the ML model was assessed in three ways (Table 10). The first

assessment of the model's performance as an accuracy score yields 69%, which means that most of the labels in the test data set were detected very well in all three targets (Table 11).

However, since accuracy is the perfect measure of the classifier's performance when the classes are unbalanced (like the dataset in this study), the other metrics should be used to evaluate the model's performance.

The Confusion Matrix results show that the classifier and ML model perfectly understood Target 1 as a Very Cold level and Target 2 as defining the Cold level. But for Target 3, which denotes the Cool level of comfort (Cool), the model was confused and misclassified 50% of the labels as Target 2 as Cold. The results show that the model ML recognized the difference between Very Cold and other targets. Therefore, although the classification is not good enough to label Target 3, it is good enough to distinguish Target 1 (no comfort quality) from the others that correspond to comfort level.

The report of the classification contains the following values:

Precision is the number of correctly identified class members divided by the number of model predictions for that class.

Recall as the number of members of a class correctly identified by the classifier divided by the total number of members of that class.

F1 is a quick method to determine whether the classifier identifies the members of a class well.

Although the model has a high accuracy rate in all three metrics precision, Recall, and F1 score for Target 1, it mixes labels of Target 2 and Target 3. The classification report shows that the model ML and classification were

Table 10 The metrics for evaluation of classification.

The metric of evaluation	The value of the accuracy
Classification accuracy score	0.69
Confusion Matrix (Macro average)	0.71
AUC in ROC	0.78

Table 11 The confusion matrix for two-class classification problems (Visa et al., 2011).

	Predicted Negative	Predicted Positive
Actual Negative	<i>a</i>	<i>b</i>
Actual Positive	<i>c</i>	<i>d</i>

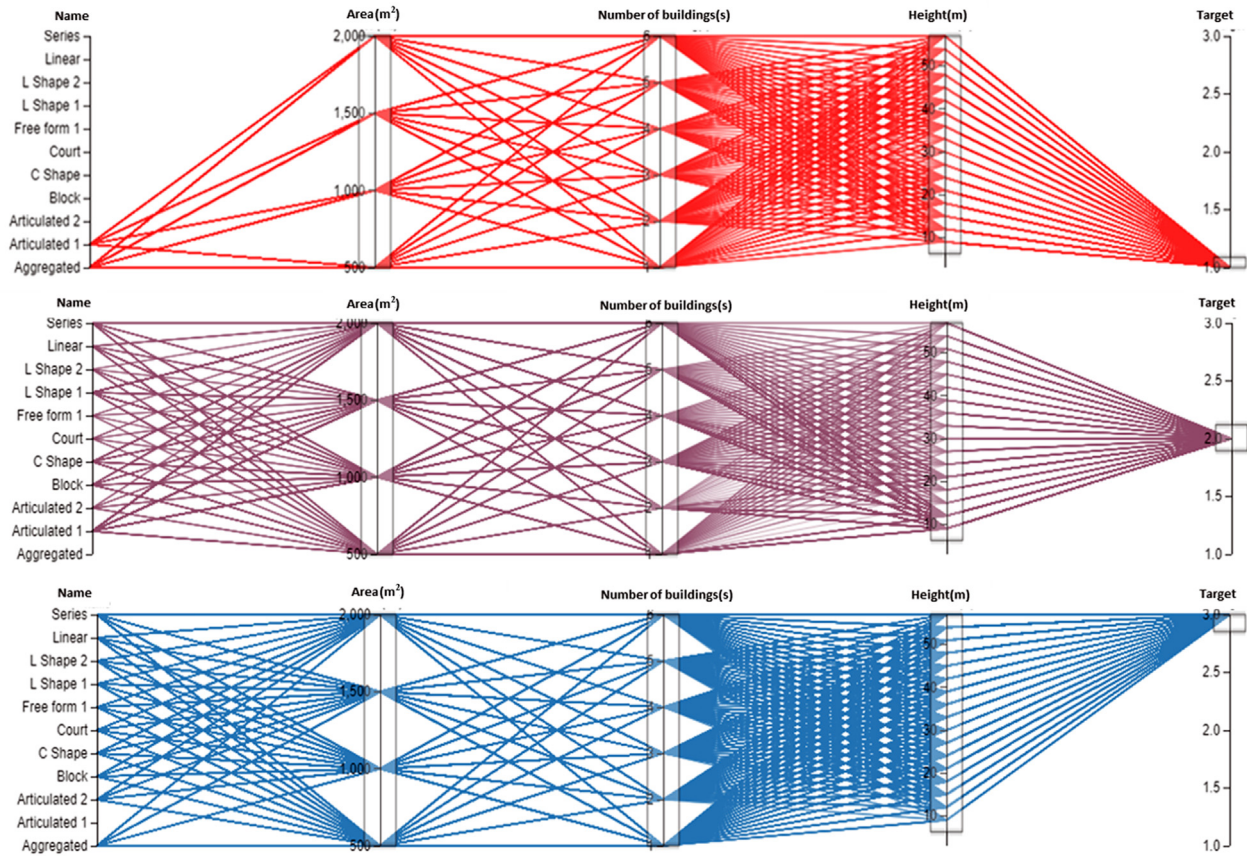


Fig. 8 Using ML Pipeline, the generated data over Targets 1, 2, and 3.

able to classify the classes of the target very well, as at least 60% of the instances were classified correctly. In this study, the results of applying AUC in ROC using the Scikit-Learn libraries show 0.78. Therefore, the maximum value of the AUC in the ROC metric is 1. Then in the case of $0.5 < \text{AUC} < 1$, there is a high probability that the classifier can distinguish the positive class values from the negative class values, and the classifier can detect more numbers of True positives and True negatives than False negatives and False positives. Thus, the value of 0.78 in AUC in the ROC metric tells us that the ML model and classification work well, and the algorithm is reliable to apply in the next step.

5.4. The evaluation of the generated data

Fig. 8 shows the results of the generated target in the new dataset introduced into the ML model. The target column was filtered by "Target 1", "Target 2" and "Target 3". The generated data includes features of 500 initial geometric models.

In this study, according to the diagrams in Fig. 9, side-walks surrounded by buildings of the morphological classes "Articulated" and "Aggregated" have the lowest level of thermal and wind comfort, i.e., the first level (Target 1, Very Cold). The remaining morphological classes provide the medium or highest level of comfort, i.e., the lowest

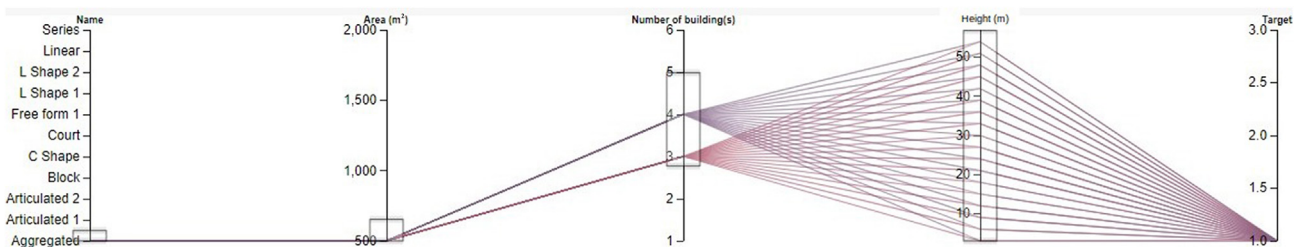


Fig. 9 The samples in the "Aggregated" morphological class cause the lowest.

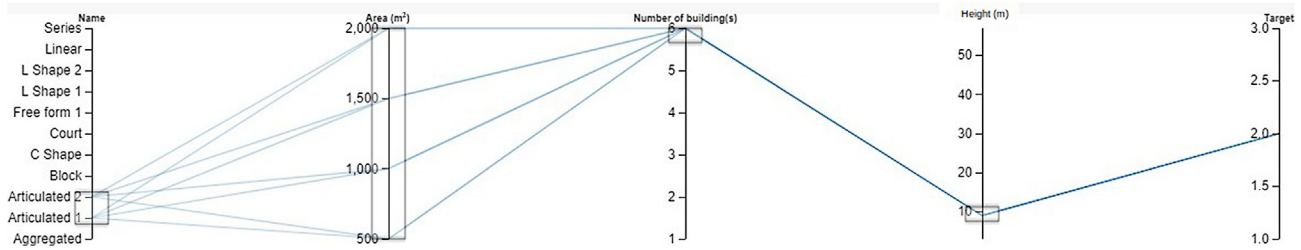


Fig. 10 The outdoor comfort level in the "Aggregated" morphological class.

Table 12 The report on the classification of the ML model.

Precision		Recall	F1-score
Target 1, Very Cold	0.93	0.86	0.89
Target 2, Cold	0.63	0.72	0.67
Target 3, Cool	0.61	0.55	0.57

level of discomfort, i.e., Cold or Cool, on the sidewalks around the buildings.

Furthermore, the results in Fig. 10 show that the buildings of the morphological class "Aggregated" with the smallest area (500 m²) and aggregated in groups of 3–5 in one place cause very cold discomfort for pedestrians on the sidewalk around the buildings. In addition, this morphological class leads to a higher level of outdoor comfort, Cold and Cool level when the variables are in the higher values.

The evaluation of two morphological classes with medium comfort levels shows that the "Articulated" class (1 and 2) causes medium, i.e., Cool, comfort level on the sidewalk of the studied cases when the number of buildings on the site is high (6). On the other hand, the height is the lowest (3 m, 6 m, or 9 m) (Fig. 10).

As Table 13 shows, most of the morphological classes in the studied cases in Tallinn provide the medium or highest comfort level (Table 12). Thus, the sidewalks around the buildings are suitable for creating a pleasant space like all kinds of activities in spring during the analyzed time and day. However, considering the effects of variables such as the area, the number of buildings on the site, and the height of the buildings on the site does not clearly indicate their independent relationships to the level of comfort around buildings. Therefore, assessing outdoor thermal

comfort depends on considering not just one of the variables but several.

6. Discussion

6.1. Analyzing data based on the ML model

Since the study aims to determine the level of outdoor thermal and wind level at the pedestrian level and understand the best morphological classes of the building that provide optimal thermal comfort, the predictive targets show us which morphological classes could offer the best comfort level to people.

Moreover, since the data were classified into three classes and the classification was assessed using the Confusion Matrix, the report shows us that the data with the labels of Target 2 and Target 3 were mixed in some cases. Therefore, ML could not identify them very well. This may be because the definition of these two classes stands for the Cold and Cool classes, while the definition of the first class, which stands for the Very Cold comfort level, is clearly different from the other comfort levels.

Therefore, evaluating the collected data on the comfort level in urban areas in the early stages of urban planning would be more helpful if the target classes were defined in binary classes (two groups), i.e., Comfort and Non-Comfort. In this way, the model of ML can better divide the samples into two groups, and the results are more accurate.

Optimizing outdoor thermal comfort at the urban scale helps to create a better environment where people engage in various activities, so architects and urban planners seek to consider the elements that contribute to maximizing comfort quality in urban areas. Collecting data on the condition of buildings in the urban area, assessing thermal

Table 13 The definition of "Target" is based on the classification of PET and WCL.

Models	Morphological class	Area (m ²)	Height (m)	Density (m ²)	Orientation (degree)	WS (m/s)	PET (C)	WCL	Comfort Level
M74	Block	500	45	7500	60	1.22	12.2	85.29	Cool
M86		500	54	9000	0	0.65	12.8	85.29	Cool
M226	Aggregated	500	36	6000	0	2.23	12	85.29	Cool
M236	blocks	500	42	7000	0	2.23	12	85.29	Cool
M240		500	42	7000	80	1.97	12.2	85.29	Cool
M242		500	45	7500	20	1.97	12.2	85.29	Cool
M246		500	48	8000	0	1.97	12.2	85.29	Cool
M256		500	54	9000	0	1.78	12.2	84.29	Cool

comfort and wind comfort around the buildings, and using this data to make predictions and generate comprehensive data on other buildings can be helpful steps in the pre-design and design processes in architecture and urban planning. The results of this study provide new information about the comfort level on the sidewalks around buildings in Tallinn, which can be helpful in the planning and preliminary design of buildings and urban areas. In this way, applying the generated data from the predictive ML Pipeline model can be considered a design tool to meet urban users' outdoor thermal and wind comfort needs.

6.2. Analyzing results

Since the study aims to evaluate outdoor comfort, it focuses on the sidewalk around the building to determine which morphological classification of buildings provides better thermal and wind comfort for people standing, sitting, and working around the buildings.

In this study, data from more about 500 models were derived and analyzed separately through a simulation process to obtain thermal and wind comfort in the proximity buildings to feed the ML model. Then, the data is evaluated with all variables that correlate with the concept of thermal and wind comfort conditions on the sidewalks around different buildings to create an ML Pipeline model and generate new data on the comfort quality. In the study, the correlations between climatic factors that determine outdoor thermal comfort and wind comfort was considered and applied in the ML process to predict outdoor comfort about morphological aspects and other characteristics of buildings. The results of the comfort level of new buildings show that the proposed framework has the potential to be applied in architectural and urban design studies to provide results based on both engineering and statistical modeling. This helped in finding effective design alternatives and reliable data, such as the morphological building class, the area, the number of buildings on the site and the height of the buildings, and their impact on the outdoor thermal and wind comfort and consequently, on the level of activities that people can perform in the city. The results show that, except for the "Articulated" and "Aggregated" morphological classes, other samples provide the comfort level of Cool or Cold and that the surrounding sidewalks of these patterns are better spaces for all levels of activities for people.

7. Conclusions

This article provides information on the correlation between climatic factors that are important for determining outdoor thermal and wind comfort around buildings in Tallinn, Estonia, and the ML approach to assess them. This study is subject to the following limitations.

The first limitation relates to the input data for the assessment of PET in the ENVI-Met software. Depending on the aim of the study, personal data, clothing and activity level of people, and climatic data can be changed. Moreover, though 500 building cases were generated to train and evaluate the ML model, the data was still relatively few.

The second limitation is considering weather data of a specific date and time during the CFD simulation process. Since outdoor thermal comfort depends on climatic conditions, the input data and, consequently, the thermal comfort and wind comfort will change at different seasons, dates, and times.

However, because of not considering the impact of urban elements, like green areas and vegetables, simulated models would be more or less different from realistic designs. In the same meaning, during the CFD simulation process, buildings' anthropogenic energy was not considered.

Nevertheless, in the previous study by the authors, the day and time were selected because it represents air temperature and wind conditions during winter and spring in Tallinn. The third limitation is the difficulty of considering the qualitative approach and people's thermal sensations. However, people's thermal senses can be considered when assessing the extent to which an area is perceived as comfortable or uncomfortable. To address this issue, the ML model will be based on microclimatic measurements, simulations, interviews, and surveys in the development of the work. The last limitation of the study is related to the fact that the prediction results of the machine learning method are less accurate than CFD simulation, and more on mathematics, and machine learning is based on empirical predictive models. The hybrid simulation is based on mathematics, and machine learning is based on empirical predictive models. The hybrid framework presented in this research suggests that machine learning can be applied in urban studies as a trustable assistance to CFD simulation.

For future work, the authors are considering a quantitative and a qualitative approach to assessing outdoor thermal comfort. This means that thermal comfort will be evaluated through simulation processes and people's subjective thermal sensations obtained through interviews and surveys. Thus, the ML algorithm will benefit from both aspects to generate new and more accurate data for outdoor thermal comfort assessments. In addition, future research aims to study larger pedestrian areas in urban environments during cold and hot seasons.

Declaration of interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CRedit authorship contribution statement

Nasim Eslamirad: Conceptualization, Methodology, Programming, Modeling and simulation, Machine learning model, Visualization, Writing- Original draft preparation. **Francesco De Luca:** Conceptualization, Methodology, Writing- Reviewing and Editing, Supervision. **Kimmo Sakari Lylykangas:** Conceptualization, Writing- Reviewing and Editing, Supervision. **Sadok Ben Yahia:** Conceptualization, Methodology, Writing- Reviewing and Editing, Machine learning model, Supervision.

Acknowledgments

This work has been supported by the European Commission through the H2020 project Finest Twins (grant No. 856602).

References

- Cerda, P., Varoquaux, G., Kégl, B., 2018. Similarity encoding for learning with dirty categorical variables. *Mach. Learn.* 107 (8), 1477–1494.
- Chen, L., Ng, E., 2012. Outdoor thermal comfort and outdoor activities: a review of research in the past decade. *Cities* 29 (2), 118–125.
- Cohen, P., Potchter, O., Matzarakis, A., 2013. Human thermal perception of Coastal Mediterranean outdoor urban environments. *Appl. Geogr.* 37, 1–10.
- De Freitas, C.R., Grigorieva, E.A., 2015. A comprehensive catalogue and classification of human thermal climate indices. *Int. J. Biometeorol.* 59 (1), 109–120.
- De Luca, F., 2019, April. Sun and wind: integrated environmental performance analysis for building and pedestrian comfort. In: *Proceedings of the Symposium on Simulation For Architecture And Urban Design*, pp. 1–8.
- De Luca, F., Dogan, T., Kurnitski, J., 2018, June. Methodology for determining fenestration ranges for daylight and energy efficiency in Estonia. In: *Proceedings of the Symposium on Simulation For Architecture And Urban Design*, pp. 1–8.
- De Luca, F., Naboni, E., Lobaccaro, G., 2021. Tall buildings cluster form rationalization in a Nordic climate by factoring in indoor-outdoor comfort and energy. *Energy Build.* 238, 110831.
- Eslamirad, N., De Luca, F., Lylykangas, K.S., 2021, November. The role of building morphology on pedestrian level comfort in Northern climate. *J. Phys. Conf.* 2042 (1), 12053. IOP Publishing.
- Eslamirad, N., Sepúlveda, A., De Luca, F., Sakari Lylykangas, K., 2022. Evaluating outdoor thermal comfort using a mixed-method to improve the environmental quality of a university campus. *Energies* 15 (4), 1577.
- Gehl, J., 2011. Three Types of Outdoor Activities, "Life between Buildings," and "Outdoor Activities and the Quality of Outdoor Space": from Life between Buildings: Using Public Space. Routledge, The city reader, pp. 586–608, 1987.
- He, Y., Liu, X.H., Zhang, H.L., Zheng, W., Zhao, F.Y., Schnabel, M.A., Mei, Y., 2021. Hybrid framework for rapid evaluation of wind environment around buildings through parametric design, CFD simulation, image processing and machine learning. *Sustain. Cities Soc.* 73, 103092.
- Huang, C., Zhang, G., Yao, J., Wang, X., Calautit, J.K., Zhao, C., et al., 2022. Accelerated environmental performance-driven urban design with generative adversarial network. *Build. Environ.* 224, 109575.
- Jamei, E., Rajagopalan, P., 2019. Effect of street design on pedestrian thermal comfort. *Architect. Sci. Rev.* 62 (2), 92–111.
- Javanroodi, K., Nik, V.M., Giometto, M.G., Scartezzini, J.L., 2022. Combining computational fluid dynamics and neural networks to characterize microclimate extremes: learning the complex interactions between meso-climate and urban morphology. *Sci. Total Environ.* 829, 154223.
- Jehad, A., Khan, K., Ahmad, N., Maqsood, I., 2012. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)* 9.
- Kim, H., Lee, K., Kim, T., 2018. Investigation of pedestrian comfort with wind chill during winter. *Sustainability* 10 (1), 274.
- Lai, D., Guo, D., Hou, Y., Lin, C., Chen, Q., 2014. Studies of outdoor thermal comfort in northern China. *Build. Environ.* 77, 110–118.
- Lawson, T.V., 1978. The wind content of the built environment. *J. Wind. Eng. Ind. Aerod.* 3 (2–3), 93–105.
- Lawson, T.V., Bristol, 1990. In: The determination of the wind environment of a building complex before construction. Report Number TVL 9025. Department of Aerospace Engineering, University of Bristol.
- Lawson, T.V., Penwarden, A.D., 1975. The effects of wind on people in the vicinity of buildings. In: *Proceedings 4th international conference on wind effects on buildings and structures*. Cambridge University Press, pp. 605–622.
- Li, K., Zhang, Y., Zhao, L., 2016. Outdoor thermal comfort and activities in the urban residential community in a humid subtropical area of China. *Energy Build.* 133, 498–511.
- Luo, Y., He, J., Ni, Y., 2017. Analysis of urban ventilation potential using rule-based modeling. *Comput. Environ. Urban Syst.* 66, 13–22.
- Marshall, S., Çalışkan, O., 2011. A joint framework for urban morphology and design. *Build. Environ.* 37 (4), 409–426.
- Mittal, H., Sharma, A., Gairola, A., 2018. A review on the study of urban wind at the pedestrian level around buildings. *J. Build. Eng.* 18, 154–163.
- Nikolopoulou, M., Lykoudis, S., 2006. Thermal comfort in outdoor urban spaces: analysis across different European countries. *Build. Environ.* 41 (11), 1455–1470.
- Okhoya, V., 2015. Towards the Application of Machine Learning on Architectural Projects. <https://doi.org/10.13140/RG.2.1.1164.3048>.
- Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O., Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. *Int. J. Comput. Trends Technol.* 48 (3), 128–138.
- Pancholy, P.P., Clemens, K., Geoghegan, P., Jermy, M., Moyers-Gonzalez, M., Wilson, P.L., 2021. Numerical study of flow structure and pedestrian-level wind comfort inside urban street canyons. *J. Roy. Soc. N. Z.* 51 (2), 307–332.
- Park, S., Tuller, S.E., Jo, M., 2014. Application of Universal Thermal Climate Index (UTCI) for microclimatic analysis in urban thermal environments. *Landsc. Urban Plann.* 125, 146–155.
- Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* 11 (5), 1633–1644.
- Ramsgaard Thomson, M., Nicholas, P., Tamke, M., Gatz, S., Sinke, Y., Rossi, G., 2020. Towards machine learning for architectural fabrication in the age of industry 4.0. *Int. J. Architect. Comput.* 18 (4), 335–352.
- Sebestyen, A., Tyc, J., 2020. Machine learning methods in energy simulations for architects and designers. In: *Proceedings of the of the 38th eCAADe Conference*. Berlin, pp. 613–622.
- Steemers, K., Yannas, S., 2000. Architecture, City, Environment. In: *Proceedings of PLEA 2000*. Earthscan, Cambridge, United Kingdom.
- Taleghani, M., Kleerekoper, L., Tenpierik, M., Van Den Dobbelsteen, A., 2015. Outdoor thermal comfort within five different urban forms in The Netherlands. *Build. Environ.* 83, 65–78.
- Tamke, M., Nicholas, P., Zwierzycki, M., 2018. Machine learning for architectural design: practices and infrastructure. *Int. J. Architect. Comput.* 16 (2), 123–143.
- Teoh, M.Y., Shinozaki, M., Said, I., 2020. Redefining the notion of outdoor thermal comfort. *SEATUC journal of science and engineering* 1 (1), 52–61.
- Visa, S., Ramsay, B., Ralescu, A.L., Van Der Knaap, E., 2011. Confusion matrix-based feature selection. *MAICS* 710 (1), 120–127.
- Wang, Z., Wang, J., He, Y., Liu, Y., Lin, B., Hong, T., 2020. Dimension analysis of subjective thermal comfort metrics based

- on ASHRAE Global Thermal Comfort Database using machine learning. *J. Build. Eng.* 29, 101120.
- Yahia, M.W., Johansson, E., 2011. The influence of environment on people's thermal comfort in outdoor urban spaces in a hot dry Climate. The example of Damascus, Syria. In: 27th International Conference on Passive and Low Energy Architecture (PLEA). Presses universitaires de Louvain, pp. 589–594.
- Zhang, X., Weerasuriya, A.U., Zhang, X., Tse, K.T., Lu, B., Li, C.Y., Liu, C.H., 2020, December. Pedestrian wind comfort near a super-tall building with various configurations in an urban-like setting. *Build. Simulat.* 13 (6), 1385–1408.