Understanding the Factors Influencing Pedestrian Walking Speed over Elevated Facilities using Tree-Based Ensembles and Shapley Additive Explanations

Arunabha Banerjee (arunabhabanerjee77@gmail.com)
Indian Institute of Technology Guwahati https://orcid.org/0000-0002-6314-295X
Rahul Raoniar
IIT Guwahati: Indian Institute of Technology Guwahati
Akhilesh Kumar Maurya
IIT Guwahati: Indian Institute of Technology Guwahati

Research Article

Keywords: Tree Ensemble, Walking Speed, Machine Learning, Foot Over Bridge, Skywalk

Posted Date: August 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-373997/v1

Abstract

Accurate estimation of factors affecting pedestrian walking speed is of paramount importance for efficient operation and management of at-grade and grade-separated infrastructures (such as foot over bridges or skywalks). Understanding such factors helps in planning for better circulation of pedestrians within confined elevated passageways as well as evacuation preparedness during emergencies. The walking speed on elevated infrastructure generally depends on the microscopic factors (demographics characteristics), macroscopic factors (average flow and density), and geometric factors (obstruction, land use type, length, connectivity, and effective width). The wide variability of these factors and their impact on walking speed makes the speed prediction modeling complex. Therefore, accuracy of such models depends on accurate field data collection, identification of pertinent variables, and implementation of appropriate modeling approaches. With the increase in computational capabilities, tree-based ensembles have gained immense popularity due to their high prediction accuracy in comparison to traditional regression models. The tree-based ensembles provide better interpretable results without a huge data requirement and are able to capture the complex non-linear relationships. These properties make tree-based ensemble models better candidates for modeling pedestrian walking speed, however, exploration on the tree-based ensemble in pedestrian related research is limited. In the current study, an attempt is made to model and compare seven tree-based models (including ensembles) to suggest the best modeling approach to identify the dominating factors and accurate prediction of pedestrian walking speeds over elevated walkways. The result of the present study showed that Gradient Boosted Trees (MAE 9.27) and Light Gradient Boosted Trees (MAE: 9.96) were best in predicting walking speed over the skywalk and foot over bridge facilities, as these boosting based methods improved the weak trees (on the basis of accuracy) sequentially. The variable importance of final models was estimated using SHapley Additive exPlanations (SHAP) which revealed that walking speed was dependent on the average flow, average density, and length of the facility. Moreover, other features such as gender, age, height, and width of the facility also play a significant role in determining the pedestrian walking speeds. The identification of important variables not only provides better insight on factors that affect walking speed over elevated facilities but also provides a valuable source of information to researchers, planners, and policymakers for better designing, operation, and management of the elevated pedestrian infrastructures.

1. Introduction

Pedestrians are the most essential part of the transportation system. Every person at some stage of his or her journey is a pedestrian. While traveling to schools, colleges, offices, shopping areas, or recreational places, pedestrians require dedicated facilities to minimize their interaction with vehicular traffic and thus travel safely while maximizing their walking comfort. Apart from at-grade facilities (such as sidewalks, crosswalks, and walkways), there are grade-separated pedestrian facilities (such as overpasses and underpasses) which are built across vehicular traffic streams to avoid interaction between pedestrians and motorized traffic.

Pedestrians are the most complex part of the transportation system, as it is extremely difficult to predict their movement behavior. The fundamental factors such as pedestrian speed, flow, and density influence the overall dynamics of pedestrian movement. The pedestrians present in streams have varied demographic characteristics (like gender and age). Further pedestrian behavior differs in their group configuration (i.e. size of group), baggage conditions, and use of hand-held devices. All such behavioral differences among pedestrians significantly impact the overall flow and density of the stream. Therefore, developing a pedestrian speed prediction model while incorporating the impact of all the above-mentioned features becomes a cornerstone in better understanding and modeling pedestrian stream behavior, which leads to better design of such infrastructures, especially during emergency evacuations.

Previous researchers mainly focused on predicting stream speed characteristics based on the average flow or density, using the fundamental diagrams for sidewalk (Arasan et al., 1994; Cepolina et al., 2018; Hoogerdoorn and Daamen, 2005; Karatas et al., 2018; Laxman et al., 2010; Marisamynathan and Vedagiri, 2016; Morall et al., 1991; Navin and Wheeler, 1969; Oeding, 1963; Older, 1968; Polus et al., 1983; Sarsam, 2013), crosswalk (Asaithambi et al., 2016; Bargegol et al., 2014; Bowman and Vecellio, 1994; Gates et al., 2006; Marisamynathan and Vedagiri, 2018; Sahani and Bhuyan; 2017; Tarawneh, 2001; Wilson and Grayson, 1980) and stairway (Fruin, 1971; Lee and Lam, 2006; Liu et al., 2018; Weidmann, 1993; Zhang et al., 2011) facilities. The mean speed over sidewalks vary consistently, as low mean speeds (52 m/min) were reported by Nazir et al. (2012) and Poei et al. (1995) in Bangladesh and Indonesia; while Knoflacher (2006) and Finnis and Walton (2008) reported higher walking speeds of 97m/min and 88m/min in Austria and New Zealand, respectively. In India, the speed over sidewalks/ walkways was reported to range between 65 m/min (Sukhadia et al., 2016) to 84 m/min (Laxman et al, 2010). The variation in walking speed was mainly due to the physique (height), attractions (presence of vendors), and culture (attire and privacy). The pedestrian studies related to crosswalk speed observed that the crossing speed of Asian pedestrians was higher (78–91 m/min) in comparison to American/ Canadian pedestrians (80 m/min), due to the higher risk-taking tendency while crossing.

Recently, the focus has shifted from macroscopic to microscopic modeling in order to predict pedestrian dynamics using different machine learning (ML) and data-based algorithms, as these tools have higher prediction accuracy, faster training speeds, and can handle missing values better. Table 2 shows the various studies conducted for speed prediction using different modeling approaches for field or experimental studies.

Table 2
Studies in Traffic and Transportation planning related to pedestrian speed prediction modeling approach

Author	Objective of study	Facility & Type of study	Sample size & Locations covered	Variables used	Model approach
Al-Azzawi and Raeside (2007), UK	Predicting walking speed	F: Sidewalk, TS: Videography	SS: 7535, LC: 6	Flow, speed, density, land use, pedestrian characteristics	MVR
Vathsangam (2010), USA [70]	Walking speed estimation with on-body gyroscopes and accelerometers	TS: Experimental	SS: 8	Age, height, weight, gender	GRP, BLR, LSR
Chang et al. (2011), Taiwan	Prediction model development for crosswalks	F: Intersections, TS: Videography	SS: 5235, LC: 8	Demographics, geometry, signal type, and phase length	-
Rengarasu et al. (2012), Sri Lanka	Walking speed estimation	TS: Videography	SS: 50, LC: 1	Gender	HTBRM such as CART and CHAID
Park et al. (2012), USA	Walking speed estimation using hand-held devices	F: Long corridor, TS: Experimental	SS: 14	Device in hand, at ear, in trouser pocket, in bag	SVM
Rastogi et al. (2013), India	Estimating flow characteristics	F: Sidewalk, TS: Videography	SS: 4784, LC: 19	Width, flow direction	LR
Matsubayashi and Shiraishi (2016), Japan	Estimate walking speed by using magnetic signatures	F: Indoor passage, TS: Experimental	SS: 400, LC: 10	Length of section, walking time	SPRING
Byun et al. (2019), South Korea	Walking speed estimation using single inertial measurement unit	F: straight corridors, TS: Experimental	SS: 785	Demographics, anthropometrics	LR
Shrestha and Won (2018), USA	Develop smart phone-based approach using DeepWalking for walking speed prediction	F: Treadmill, TS: Experimental	SS: 10	Gender, age, physical condition	DCNN
Herrera-Angulo and Zenteno-Bolanos (2018), Peru	Pedestrian speed estimation using low cost microwave presence detector	TS: Experimental	SS: 100	-	ML techniques
Guo et al. (2019), China	Estimating pedestrian walking speed based on pose awareness solution using smart phones	F: Indoor, TS: Experimental	SS: 101	Gender, texting, swinging, pocket and calling mode	Bayes, k-nearest neighbour, DT, NN, SVM, RF
Kawaguchi et al. (2019), Japan	Walking speed estimation for smartphone PDR	TS: Experimental	SS: 5, LC: 79 routes	Position, gait, routes, walking time, route length	Dual CNN-LSTM
Bansal et al. (2019), India	Capture variation in pedestrians speed	F: Signalized intersection, TS: Videography	LC: 16	Age, gender, group size, luggage, geometry	SLR
Tordeux et al. (2020), Germany	Pedestrian speed prediction in complex architectures	F: Corridor, bottleneck, TS: Experimental	SS: 230	Geometry, flow, density	ANN

Note: F- Type of facility, TS- Type of study, SS- Sample size, LC- Locations covered, MVR- Multivariate Regression, GRP- Gaussian Process Based Regression, BLR- Bayesian Linear Regression, LSR- Least Squares Regression, HTBRM- Hierarchical Tree Based Regression Model, CART- Classification & Regression Trees, CHAID- Chi-square Automatic Interaction Detector, SVM- Support Vector Machine, LR- Linear Regression, DCNN- Deep Convolution Neural Network, ML- Machine Learning, DT- Decision Tree, NN- Neural Network, SVM- Support Vector Machine, RF- Random Forest, LSTM – Long Short Term Memory, SLR- Stepwise Linear Regression, ANN- Artificial Neural Network

As per Table 2, linear or multivariate regression models were used by some authors to predict the pedestrian speed on sidewalk or crosswalk facilities. However, majority of studies, which used ML tools to predict the pedestrian walking speed, were based on controlled or experimental setups.

Past studies on overpass facilities mostly compared the preference of pedestrian's choice for the grade-separated facility over at-grade facility (Abojaradeh, 2013; Anciaes and Jones, 2018; Hasan and Napiah, 2018; Oeding, 1963; Räsänen et al., 2007). Some recent studies also tried to measure illegal at-grade road crossing speeds (Demiroz et al., 2015; Truong et al., 2019). However, there is a paucity of studies highlighting the impact of different demographic and geometric features on pedestrian walking behavior (especially walking speed) over elevated facilities using field data collection techniques.

1.1. The Study Motivation And Objective

Speed plays a very significant role in the better circulation of pedestrians (with comfortable levels of service) as well as evacuation preparedness during emergency situations. Moreover, as the speed depends on different individual and geometric factors, it affects the travel time accessibility to elevated facilities as well. As per Table 2, machine-learning tools were mainly used to predict the speed under controlled conditions; while the conventional regression modeling approach was used for estimating speed over sidewalks/ crosswalks.

However, in the present study, an attempt is made to predict the factors impacting pedestrian walking speed over elevated pedestrian walkways (foot over bridges and skywalks) using individual characteristics (gender, age, luggage condition, mobile use), group characteristics (average flow and density) and geometric conditions (the type of obstruction present, land use type, type of connectivity, length and width of the facility), through different tree-based machine learning algorithms. The outcomes of the present study can provide researchers, planners, designers, and policymakers with ample justification to account for the factors affecting the walking speed over elevated walkways, and thus come up with better designed user-friendly infrastructures in the future.

2. Methodology

2.1. Selection Of Survey Locations And Collection Of Data

In order to capture the factors affecting pedestrian walking speed, a videography data collection technique was used. Firstly, the different elevated pedestrian facilities (FOBs and skywalks) were visited across six cities (*NCR: National Capital Region, Bengaluru, Kolkata, Gangtok, Guwahati and Mumbai*) covering different geographic locations of India. In total 13 FOB locations and 7 skywalk locations were fixed for final data collection. The data was collected using high definition video camera fixed over a high vantage tripod stand. The duration of data collection over the mid-block section of the elevated facilities was approximately 3 hours, during either morning peak hour (7.30-10.30am) or evening peak hour (4.30-7.30pm). The trap length across different locations varied between 10-15m for both the elevated facilities. Figure 1 shows the position of the camera along with the trap length and effective width for an elevated walkway situated in Gangtok. The details of the locations from where the data were collected across different Indian cities are provided in Table 3.

Table 3
Site characteristics for elevated facilities

Details of section	City	Type of Facility	Land use type	Length of facility (m)	Effective width (m)	Trap length (m)	Obstruction	Total Sample	*Sample size (%)*				
									Gender		**Age**		
									Male	Female	≤ 10	11–20	21–40
Anand Vihar	NCR	FOB	PTT	61.5	4.7	10.0	Beggars	631	76.86	23.14	4.12	19.65	48.34
ITO			Commercial	39.2	1.6	10.5	Beggars	622	78.78	21.22	0.32	1.13	66.40
Maharani Bagh			Residential	41.3	1.6	8.5	Both	527	77.04	22.96	1.71	6.64	61.86
Vaishali			PTT	88.1	2.5	8.3	Beggars	572	77.45	22.55	1.40	7.17	73.43
Marathalli	Bengaluru		Commercial	32.1	2.4	10.2	None	598	79.43	20.57	1.34	11.87	72.91
Tin Factory			Commercial	49.8	1.6	8.4	Vendors	726	72.31	27.69	1.24	1.93	75.76
Lake Town	Kolkata		Residential	42.2	1.9	9.0	Vendors	610	64.75	35.25	2.62	13.44	50.00
Ultadanga			Commercial	42.4	1.9	12.0	Beggars	691	77.57	22.43	1.88	9.70	63.97
Telephone Exchange	Gangtok		Commercial	18.8	1.5	10.0	None	695	56.26	43.74	2.45	3.31	70.22
M.G. Marg			Commercial	18.5	1.4	10.0	None	809	67.49	32.51	1.24	0.99	69.34
STNM Hospital			Institutional	42.6	1.5	10.0	None	809	48.83	51.17	0.62	1.11	77.87

Table 3
Site characteristics for elevated facilities *(continued)*

Details of section	City	Type of Facility	Land use type	Length of facility (m)	Effective width (m)	Trap length (m)	Obstruction	Total Sample	*Sample size (%)*				
									Gender		**Age**		
									Male	Female	≤ 10	11–20	21–40
Maligaon	Guwahati	FOB	Commercial	34.0	1.6	10.2	None	651	68.36	31.64	5.53	10.45	49.92
IIT Bombay	Mumbai		Educational	43.3	2.6	10.0	None	883	75.54	26.46	1.02	12.12	76.33
Andheri		Skywalk	Commercial	581	2.5	10.0	Vendors	781	71.83	28.17	-	9.86	74.90
Bandra			Commercial	494	2.7	15.0	None	706	75.78	24.22	2.83	9.49	53.26
Ghatkopar			Residential	315	2.7	16.0	None	773	74.00	26.00	0.26	12.94	69.47
Goregaon			Commercial	625	3.4	15.0	Vendors	798	79.07	20.93	0.88	5.76	78.20
Kalyan			Shopping	1287	2.0	15.0	Vendors	669	80.87	19.13	1.79	6.28	67.71
Santa Cruz			Shopping	438	2.6	10.0	Vendors	896	69.53	30.47	1.00	12.39	66.52
Vile Parle			Shopping	460	3.0	15.0	Vendors	719	69.40	30.60	3.48	23.37	43.39

From Table 3, it can be observed that the total length of the elevated walkways varied between 18.5-88.1m (FOBs) and 315-1287m (skywalks). The effective width varied between 1.4-4.7m (for FOBs) and 2-3.4m (for skywalks). The effective width was calculated after deducting the shy away or buffer distance from the actual width. The average sample size per location varied between 679 (for FOBs) and 763 (for skywalks). The dominant categories across both the facilities were male pedestrians (≥ 70%), 21–40 years' age group (≥ 65%), with luggage (≥ 52%), and without mobile phones (≥ 85%). From Table 3 it is also observed that across many FOB locations, the beggars and vendors were prevalent; while across some skywalks, the vendors were present. Also, the different land-use types considered in the study ranged from public transport terminal (PTT) to commercial, residential, institutional, educational, and shopping locations.

2.2. Data extraction

Collected data were processed in the lab using the manual data extraction technique. As the aim of the study was to identify the factors affecting the pedestrian speed, thus individual parameters (such as age, gender, luggage condition, mobile use) along with other factors (such as obstruction, land use type, time of data collection, average flow, average density, length of facility, connectivity, effective width) were considered while extracting the video data. Table 4 shows the description of the factors extracted and considered for the modeling of pedestrian behavior.

Table 4
Description of factors considered for data extraction

Parameters	Description
Gender	*0: Male, 1: Female*
Age (in years)	*0: ≤10, 1: 11–20, 2: 21–40, 3: 41–60, 4: ≥60*
Luggage condition	*0: Without, 1: With*
Mobile use	*0: Without, 1: With*
Obstruction	*0: None, 1: Beggars, 2: Vendors, 3: Both*
Type of elevated facility	*0: FOB, 1: Skywalk*
Land use type	*0: Commercial, 1: Educational, 2. Institutional, 3. Public Transport Terminal (PTT), 4. Residential, 5. Shopping*
Time of data collection	*0: Morning, 1: Afternoon, 2: Evening*
Length of facility	*The distance between the entry and exit points of the facility*
Width of the facility	*The breadth of the facility, either in terms of actual width (with buffer space) or effective width (deducting buffer space)*
Average flow (ped/min/m)	*Calculated as the number of pedestrians crossing the entry section of the study area per minute by the effective width of the trap*
Average density (ped/m^2)	*Calculated by counting the number of pedestrians within a trap area at every 20 second interval. In one minute, three density reading were taken, and then average value of density was calculated per minute*

3. Data Analysis

The demographic characteristics and speed distribution data were obtained by performing exploratory data analysis on extracted videography data containing 7522 (FOB) and 5325 (skywalk) samples respectively. The final analysis was carried out comparing the speed prediction between the two types of elevated walkways (FOBs and skywalks) under different land-use types (commercial, educational, institutional, public transport terminal, residential, and shopping).

3.1 Demographic characteristics

The demographic characteristics such as gender, age, luggage condition, and mobile use are relevant in understanding the existing usage pattern of the elevated walkways. Table 5 presents the demographic characteristics of the pedestrians based on gender, age, and luggage condition for FOB and skywalk facilities. The table also shows the results of the statistical tests (t-test and ANOVA single factor test) between the different pedestrian demographic characteristics. The statistical tests were conducted to check whether a significant difference exists between the different pedestrian categories. The t-test is performed to compare if a significant difference exists between two sub-categories (e.g. *gender: male/ female, luggage: with/ without, and mobile: with/ without*), while the ANOVA test is performed to compare between two or more pedestrian categories (e.g.: *age: <10/ 11–20/ 21–40/ 41–60/ >60 years*). Higher values of t-statistical value in comparison to t-critical value (for t-test) and higher F-statistical value in comparison to F-critical value (for ANOVA test), signifies that significant difference exists between the different pedestrian demographic characteristics (at 5% significance level).

Table 5
Demographic characteristics of the pedestrians for both elevated facilities

Pedestrian demographics	Category	Facility	Sample size (%)	Mean speed (m/min)	Standard deviation	Variance	t-test		ANOVA test		Significance
							t-stat	t-crit	F-stat	F-crit	
Gender	*Male*	FOB	69.24	73.18	13.10	182.51	22.71	1.96	-	-	Sig.
	Female		30.76	66.55	12.15	146.05					
	Male	Skywalk	76.98	81.51	12.80	159.98	18.07	1.96			Sig.
	Female		23.02	74.31	12.24	141.94					
Age (years)	*< 10*	FOB	1.89	63.09	11.16	175.65	-	-	121.23	2.37	Sig.
	11–20		8.62	70.56	12.65	168.84					
	21–40		64.44	73.33	12.98	177.29					
	41–60		21.07	67.93	12.50	155.99					
	≥ 60		3.97	61.28	12.41	155.09					
	< 10	Skywalk	0.97	72.27	11.60	154.43	-	-	33.51	2.37	Sig.
	11–20		10.82	77.55	13.37	154.31					
	21–40		66.44	80.02	12.75	160.09					
	41–60		17.85	78.08	12.73	158.99					
	≥ 60		3.93	70.3	13.64	187.21					
Luggage	*With*	FOB	51.52	71.04	12.80	163.84	3.96	1.96	-	-	Sig.
	Without		48.48	71.93	13.59	184.66					
	With	Skywalk	77.91	77.95	12.71	163.07	6.48	1.96			
	Without		22.09	81.37	13.79	162.86					
Mobile	*With*	FOB	7.81	69.58	13.28	166.66	3.16	1.96	-	-	Sig.
	Without		92.19	71.34	13.15	180.15					
	With	Skywalk	14.93	77.45	13.30	190.72	5.24	1.96			
	Without		85.07	80.16	12.89	158.19					

From Table 5 it is observed that the majority of the pedestrians using the FOB and skywalk facilities were male pedestrians (73–81%) in the age group of 21–40 years (73–78%) and with luggage (71–78%). The male pedestrians were observed to walk at higher average speeds in comparison to the female pedestrians over both the elevated walkways by 6-7m/min. The pedestrians in the age group of 21–40 years walked fastest in comparison to the other age categories. The pedestrians without luggage had significantly higher average walking speeds in comparison to the pedestrians with luggage for both the facilities. Further, the proportion of pedestrians using a mobile phone while walking alone over skywalk facilities was double in comparison to FOB facilities. The reason for the higher proportion of mobile users over skywalks could be due to the long traveling length on skywalk facilities and thus the pedestrians may use the mobile phones to overcome boredom. The average walking speeds for mobile users were 3m/min slower than the non-mobile users. The results of the statistical tests (t-test and ANOVA test) showed that for different demography categories, significant difference exists between gender (*male/ female*), luggage (*with/ without*), mobile (*with/ without*) and age *(< 10/ 11–20/ 21–40/ 41–60/ >60* years).

3.2. Speed distribution

Probability density functions were also used to understand the speed variation among the different categories of pedestrians (based on gender, age, and luggage condition) for FOBs and skywalks (refer to Fig. 2). The x-axis represents the walking speed (m/min) while the y-axis represents the probability or relative frequency.

Figure 2(a) shows that the male pedestrians walked faster than the female pedestrians for both the facilities by 6m/min. The male pedestrians using skywalks were observed to have higher mean speed than the male pedestrians using FOBs by 7m/min. This increase in speed is observed as skywalks offer a wide path for pedestrians which encourages them to walk at a higher speed than FOB facilities. Further, as the walkways are much lengthier than FOBs, pedestrians try to travel faster to cover the longer length quickly.

The age-wise speed distribution (refer to Fig. 2(b)) shows that the child (< 10 years) and elderly (> 60 years) pedestrians over FOB facilities were the slowest pedestrians. The young adult pedestrians (21–40 years) were observed to have the highest walking speed across both the elevated facilities. In comparison to the FOB pedestrians (of age 21–40 years), skywalk pedestrians users (in the age category of 21–40 years) had higher walking speeds by 7m/min due to the greater available walkway widths and thus had the freedom to choose higher walking speeds to cover longer distances.

From Fig. 2(c), it was observed that the pedestrians with and without luggage had higher walking speeds on skywalks in comparison to FOBs by 6–10 m/min. The speeds for pedestrians with and without luggage over FOBs were quite similar, while over skywalk the pedestrians without luggage walked at higher speeds (4m/min) in comparison to the pedestrians with luggage. The reason for similar walking speeds over FOB facilities was due to the fact that as the traveling length was smaller, the luggage did not have much impact on their speed. However, over skywalk facilities when pedestrians had to travel longer distances, carrying luggage played a crucial role and significantly reduced the walking speed.

Figure 2(d) shows that for both skywalks and FOBs, the pedestrians without mobile usage had a higher walking speed than the ones with mobile by 3m/min. Also, similar to other distribution functions, the pedestrians with/ without mobile over skywalks had higher speeds in comparison to FOBs. The main reason for higher speed over the skywalks could be the available higher walkable width and the longer length which pedestrians had to travel over skywalk facilities.

4. Modelling Approaches

In the present study, an effort was made to understand the best-suited model in terms of prediction accuracy of walking speed over elevated walkway facilities. The tree-based algorithms have several advantages over other machine learning algorithms, described in Table 6. The different tree-based modeling approaches used in the current study were Light Gradient Boosting Machine (LGBM), Gradient Boosting Regressor (GBR), Adapting Boosting Regressor (Ada Boost), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Extra Tree Regressor (ETR) and Decision Tree (DT).

Table 6
Description of different tree-based modeling approaches

Modeling approach & Developer	Description	Features	Important parameters	Advantages	Disadvantages
Light Gradient Boosting Machine, Ke et al. (2017)	It is a gradient boosting framework based on a decision tree algorithm, used for regression or classification	It splits the tree leaf wise	num_leaves, min_data_in_leaf, max_depth	• Faster training speed • Higher efficiency • Lower memory usage	• Performance of small dataset is poor
Gradient Boosting Machine, Friedman et al. (2000)	It is an ML technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees	GBM sequentially builds regression trees on all the features of the dataset in a fully distributed way	ntrees, max_depth, learning rate	• High prediction accuracy • High flexibility towards loss	• Can cause overfitting and cross-validation must be used to neutralize • Computationally expensive
AdaBoost Regressor, Freund and Schapire (1997)	It is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset	• It works by putting more weight on difficult to classify instances • It can be used for both classification and regression problems.	base_estimator, n_estimators, learning_rate, loss, random_state	• Very useful for weak classifiers and cascading. • AdaBoost has a high degree of precision;	• Data imbalance leads to a decrease in classification accuracy • Training is time-consuming
Extreme Gradient Boosting, Chen and Guestrin (2016)	XGBoost uses second-order gradients, i.e. second partial derivatives of the loss function which makes faster optimization	• It provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way.	min_child_weight, max_depth, max_leaf_node, gamma, subsample, comsample_bylevel, lambda, alpha, scale_pos_weight	• Better regularization which reduces overfit • Parallel processing, • Higher flexibility • Handles missing values better	• More likely to overfit than bagging (random forest) if the model is not stopped early, • Training time is higher in comparison to Light GBM

Table 6
Description of different tree-based modeling approaches *(continued)*

Modeling approach & Developer	Description	Features	Important parameters	Advantages	Disadvantages
Random Forest, Breiman (2017)	It uses a bootstrapping analogy similar to bagging but adds additional randomness to models at each split by randomly selecting input variables/ features.	In comparison to GBM, where one tree is built at a time, in RF each tree is trained independently using random sample data	mtries, ntrees, depth	• Fast training speed, • Simple to implement	• High computation power required • Complex as it creates a lot of trees, • Longer training period
Extra Trees Regressor, Geurts et al. (2006)	In Extra Trees, the features and splits are selected at random.	• Creates a large number of unpruned decision trees from the training dataset • Splits are selected at random for each feature in the Extra Trees Classifier,	max_features, max_depth, min_samples_split, min_samples_leaf, min_weight_fraction_leaf, max_leaf_nodes, bootstrap, oob_score, n_jobs, random_state, verbose, warm_start	• Provide the best trade-off between bias and variance, • It produces trees with low variance	—
Decision Tree/ CART (Classification & Regression Trees), Breiman (1984)	It is a non-parametric supervised learning method that is used for classification and regression.	• Classification trees: models where the target variable can take a discrete set of values • Regression trees: decision trees where the target variable can take continuous values	max_depth, min_samples_split, min_samples_leaf, max features	• Can solve both classification and regression problems, • Could handle missing data • Easy to explain	• Unstable, • Overfitting • Suffer from high variance

5. Study Methodology

Algorithm 1 shows the step-by-step methodology of speed prediction for elevated walkways. The study methodology involved literature survey, preliminary site inspection, videography data collection and extraction, followed by speed prediction modeling, and finally extracting the important features for a policy decision.

As explained earlier in Table 3, the data was collected from 13 FOBs and 7 skywalks across different Indian cities. This data must be processed carefully before using them for training speed prediction models. Initially, the data columns were normalized using min-max scalar. Further, one hot encoding was applied to the categorical columns. The final prepared dataset was randomly divided into 80% (*FOB: 11332, Skywalk: 4273*) for training and 20% (*FOB: 2833, Skywalk: 1069*) for testing of the developed model respectively. Different modeling algorithms offer different hyper-parameters. Thus, initially, all the selected algorithms were trained using 10-fold cross-validation (CV) with 10 random hyperparameter space on 80% train data. The models were ranked in decreasing order based on Mean Absolute Error (MAE) evaluation metric. The MAE metric was selected for model performance evaluation due to its less sensitivity to outliers. Once the top algorithms were identified, they were further tuned with 100 random hyperparameter space using a 10-fold CV to get more reliable estimates. The tuned model was then finally tested on the remaining 20% test dataset.

Algorithm #1: Estimation of factors for walking speed prediction over FOBs and Skywalks

Input

Demographics, existing condition, usability dependent, land use type, facility type, mid-block walkway length and width, steps, and stair characteristics.

Output

Predicting the factors influencing the walking speed of pedestrians for FOB and skywalk facilities

// Pre-Processing Stage

1. For *a column in violation dataset*

2. Call handle missing values
3. Call normalize
4. End For

// Model Building and Ranking

5. ***For*** *i in range (1: total samples)*

6. Split the dataset into 80% training and 20% testing

7. Split 80% dataset according to 10-fold CV training and validation dataset

8. ***End*** for

9. ***For*** *i in range (1: N model algorithms)*

10. ***For*** *each 10-fold CV training part*

11. train model with 10 random hyperparameter search space

12. ***End*** for

13. ***Call evaluate MAE on the validation set for Model Ranking***

14. ***End*** *for*

// Hyperparameter Tuning

15. ***For*** *best model's each 10-CV training part*

16 . ***For*** *i in range (1 to 100 random hypermeters combination)*

17. with each i train model

18. ***Call evaluate MAE on the validation set***

19. ***End*** for

// Evaluation stage

20. Evaluate the best model on the remaining 20% test dataset

21. Print test MAE score

22. Save the best model

23. ***Compute Shaply Values and rank variables as per their importance***

6. Speed Prediction Model Development

In the present study, different tree-based modeling approaches (GBM, LGBM, XGBoost, Adaboost, RF, ETR, and DT) were explored to predict the walking speed determinants over elevated pedestrian facilities (*regression: continuous outcome*) using PyCaret 2.0 (Ali, 2020) machine learning library (through open-source programming language Python version 3.6). The speed models were trained for two separate elevated pedestrian facilities i.e., Foot Over Bridges (FOBs) and Skywalks.

6.1. Model Training And Hyperparameters Tuning

In order to train the speed models, PyCaret 2.0 machine learning library was utilized. The total samples (*FOB: 14165, Skywalk: 5342*) were randomly split into 80% train (*FOB: 11332, Skywalk: 4273*) and 20% test dataset (*FOB: 2833, Skywalk: 1069*). Comparing multiple models and tuning all types of hyperparameters could be time-consuming; thus, initially, a 10-fold CV was performed with default hyper-parameters to get the idea about the overall best-performing model. The 10-fold CV models were trained with different tree-based models including ensembles. The models include Light Gradient Boosting Machine (LGBM), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), Adaptive Boosting Regressor (AdaBoost), Random Forest (RF), Extra Tree Regressor (ETR), and Decision Tree (DT). The models were trained, and the average performance of the CV was reported using various regression metrics such as MAE, MSE, RMSE, RMSLE, MAPE, and Training Time (TT), refer to Tables 7 and 8. After training, the models were sorted based on the MAE criteria as this evaluation metric is robust. The 10-fold CV result of the FOB speed prediction model (refer to Table 7) revealed that LGBM topped in the overall performance (MAE: 9.520). Similarly, models were trained using 10-fold CV for skywalks; where GBR was observed to be the best performing model with an MAE of 9.232 (refer to Table 8).

Table 7
FOB 10-fold CV model comparison summary

MODEL	MAE	MSE	RMSE	RMSLE	MAPE	TT (Sec)
Light Gradient Boosting Machine (LGBM)	**9.520**	156.111	12.493	0.174	0.141	0.165
Gradient Boosting Machine (GBM)	9.967	159.707	12.636	0.176	0.143	0.442
Extreme Gradient Boosting (XGBoost)	10.116	165.312	12.855	0.178	0.145	0.292
AdaBoost Regressor (Ada Boost)	10.229	167.105	12.925	0.180	0.147	0.126
Random Forest (RF)	10.480	177.035	13.303	0.184	0.149	0.415
Extra Trees Regressor (ETR)	10.530	178.070	13.365	0.185	0.150	0.278
Decision Tree (DT)	10.531	178.717	13.366	0.185	0.150	0.019

Table 8
Skywalk 10-fold CV model comparison summary

MODEL	MAE	MSE	RMSE	RMSLE	MAPE	TT (Sec)
Gradient Boosting Machine (GBM)	**9.232**	136.888	11.695	0.153	0.123	0.217
Light Gradient Boosting Machine (LGBM)	9.261	136.361	11.801	0.154	0.123	0.138
AdaBoost Regressor (Ada Boost)	9.846	154.428	12.418	0.161	0.130	0.074
Extreme Gradient Boosting (XGBoost)	9.995	162.098	12.727	0.165	0.132	0.243
Random Forest (RF)	10.262	169.975	13.034	0.169	0.136	0.402
Extra Trees Regressor (ETR)	10.945	165.413	13.974	0.181	0.144	0.273
Decision Tree (DT)	11.483	216.114	14.696	0.191	0.151	0.014

6.2. Model Hyper-Parameters Optimization

To obtain the best performing model and to reduce overfitting, a random hyper-parameter search was performed. Random search is faster and computationally less expensive compared to complete grid search [10]. For the FOB speed model (i.e., LGBM), the hyper-parameters were the number of leaves, maximum tree depth, learning rate, number of estimators, minimum split gain, regression alpha, and lambda. Similarly, for the skywalk model (i.e., GBM) the hyper-parameters were loss, the number of estimators, learning rate, subsample, criterion, minimum samples split, minimum samples leaf, maximum depth, and features. The different hyperparameters, their ranges, and definitions are presented in Table 9.

Table 9. Hyper parameters used in Light Gradient Boosting Machine (LGBM) and Gradient Boosting Machine (GBM) and their unique

parameters

Model	Hyper-parameters	Unique parameter definition
Light Gradient Boosting Machine (LGBM)	num_leaves: [10,20,30,40,50,60,70,80,90,100,150,200]	• num_leaves: it is the main parameter that controls the complexity of the tree-based models.
	max_depth: [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110]	• max_depth: it defines how long a tree will be allowed to grow, i.e., the maximum number of children which can grow out from the tree until the tree is cut off.
	learning_rate: [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]	• learning_rate: it is the process of adding weighting factor to new trees in the model to slow down the leaning.
	n_estimators: [10, 30, 50, 70, 90, 100, 120, 150, 170, 200]	• n_estimators: the parameter represents the number of trees that need to be built before majority voting or an average of predictions.
	min_split_gain: [0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]	• min_split_gain: it is the minimum loss reduction requires in order to make a further partition on the leaf node of the tree.
	reg_alpha: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]	• reg_alpha and reg_lambda: regularization terms based on weights.
	reg_lambda: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]	• loss: it is a function that defines the mean squared error (MSE), which can be calculated by using gradient descent and updating the predictions based on the learning rate.
Gradient Boosting Machine(GBM)	loss: ['ls', 'lad', 'huber', 'quantile']	• subsample: the parameter controls the proportion of random samples for each tree. The lower value of subsample prevents overfitting.
	n_estimators: np.arange(10,200,5)	• criterion: it is the parameter that measures the impurity of the split. In the case of regression, it is represented by “friedman_mse”, “mse”, or “mae.
	learning_rate: np.arange(0,1,0.01)	• min_samples_split: it represents the minimum number of data points or samples placed in a node before splitting operation.
	subsample: [0.1,0.3,0.5,0.7,0.9,1]	
	criterion: ['friedman_mse', 'mse', 'mae']	
	min_samples_split: [2, 4, 5, 7, 9, 10]	• min_samples_leaf: it represents the minimum number of samples that are required in the leaf node.
	min_samples_leaf: [1, 2, 3, 4, 5, 7]	
	max_depth: [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110]	• max_features: while splitting a node, it is the size of the random subset of features to be considered in the model.
	max_features: ['auto', 'sqrt', 'log2']	

By default, PyCaret performs 10 random iterations over search space. Thus, to get highly optimized FOB and skywalk models, the random iteration for hyperparameter was set to 500. The training involved 10-fold cross-validation to get a better estimate (average) of the model performance. The MAE was selected as the model evaluation metric. The FOB 10-fold CV hyper-parameter optimization results revealed an average MAE of 9.839 with a standard deviation of 0.182 (refer to Table 10). In case of the skywalk model, the hyperparameter tuning showed an MAE of 9.223 with a standard deviation of 0.119 (refer to Table 11). The optimized model hyperparameters for both FOB and skywalk speed models are presented in Table 12.

Table 10
Summary of tuned 10-fold CV FOB Model Performance

CV Folds	MAE	MSE	RMSE	RMSLE	MAPE
1	9.912	157.868	12.564	0.177	0.114
2	9.615	152.079	12.332	0.167	0.134
3	10.037	161.588	12.711	0.173	0.141
4	9.587	148.379	12.181	0.174	0.143
5	9.965	161.440	12.705	0.175	0.141
6	9.768	159.022	12.610	0.179	0.142
7	10.003	156.811	12.522	0.172	0.142
8	9.549	148.692	12.193	0.170	0.137
9	9.986	156.961	12.528	0.173	0.143
10	9.973	160.706	12.677	0.180	0.147
Mean	**9.839**	156.355	12.502	0.174	0.141
SD	**0.181**	4.720	0.189	0.003	0.003

Table 11
Summary of tuned 10-fold CV Skywalk Model Performance

CV Folds	MAE	MSE	RMSE	RMSLE	MAPE
1	9.074	132.700	11.519	0.146	0.118
2	9.135	133.192	11.540	0.153	0.123
3	9.361	147.865	12.160	0.171	0.134
4	8.866	123.222	11.100	0.138	0.113
5	9.476	143.502	11.979	0.149	0.120
6	9.103	133.738	11.564	0.148	0.118
7	9.365	139.856	11.826	0.153	0.123
8	9.448	138.764	11.779	0.160	0.129
9	9.011	131.864	11.483	0.154	0.123
10	9.391	141.629	11.900	0.157	0.125
Mean	**9.223**	136.633	11.685	0.153	0.123
SD	**0.119**	6.719	0.288	0.008	0.005

Table 12
Tuned best performing models hyperparameters for FOB and skywalk facilities

Models	Model Hyperparameter
Light Gradient Boosting Machine (for FOB)	num_leaves: 20, max_depth: 20, learning_rate: 0.1, n_estimators: 90, min_split_gain: 0.2, reg_alpha: 0.1, reg_lambda: 0.1
Gradient Boosting Machine (for Skywalk)	loss: huber, n_estimators: 175, learning_rate: 0.01, subsample: 0.3, criterion: friedman_mse, min_samples_split: 10, min_samples_leaf: 1, max_depth: 10, max_features: auto

Further, to obtain the model performance on the unseen (test) dataset, the final models were tested on the remaining 20% (*FOB: 2833, Skywalk: 1069*) test dataset. Table 13 shows the model performance summary on the test data set. The performance summary revealed that the overall optimized FOB speed prediction model (i.e., using LGBM) performed well on the unseen/test dataset *(MAE: 9.960)*. Similarly, the skywalk speed prediction model (i.e., using GBM) performance on unseen/test dataset provided an overall good performance (*MAE: 9.273*).

Table 13
Summary of FOB and skywalk model performance estimated on the test dataset

Model	MAE	MSE	RMSE	RMSLE	MAPE
Light Gradient Boosting Machine (LGBM)	9.960	156.394	12.505	0.176	0.144
Gradient Boosting Machine (GBM)	9.273	132.358	11.504	0.147	0.122

7. Applications Of Tree-based Machine Learning Techniques In Other Areas Of Transportation Engineering And Its Comparison With The Current Study

There are different studies based on application of advanced soft computing techniques in the transportation engineering domain (refer to Table 14), but very few of them are related to pedestrian-based research. Results of the present study highlighted that boosting-based model could be one of the best choices for predicting pedestrian walking speed over FOBs and skywalks.

Table 14
Application of Tree-Based ML techniques in the Transportation Sector

Study type	Author	Soft computing technique used	Evaluation metrics	Important conclusions
Travel mode choice	**Ermagun et al. (2015), USA**	Nested logit, RF	Accuracy	RF significantly outperformed nested logit
	Sekhar and Madhu (2016), India	RFDT, MNL	Kappa statistic, MAE, RMSE, RAE	RFDT model had higher accuracy
	Cheng et al. (2019), China	RF, SVM, AdaBoost, MNL	MAPE, accuracy	RF and SVM outperformed AdaBoost and MNL
	Ha et al. (2019), Indonesia	GBM, DNN	Relative importance plot	GBM outperformed DNN
Travel time prediction	**Zhang and Haghani (2015), USA**	RT, GBM, RF	MAPE	GBM performed better in comparison to RT and RF
	Cheng et al. (2019), China	SVM, GBDT, BPNN	MAD, MAPE, RE	On basis of MAPE, GBDT outperformed other methods
Traffic prediction	**Yang et al. (2017), China**	GBM, SVM, and BPNN	MAPE, MAE	GBM performed better than SVM and BPNN
	Alajali et al. (2018), Australia	XGB, GBRT, RF	MSE, MAE	Similar MSE values were obtained using the three techniques
Lane changing maneuvers	**Mousa et al. (2018), USA**	DT, RF, GBM, and XGB	AUC	XGB outperformed other techniques
Driver's stop-or-run behavior	**Ding et al. (2016), China**	GBM	R^2	GBM better handled different predictor variables and fit complex non-linear relationships
Bike-sharing	**Regue and Recker (2014), USA**	GBM, Neural Network, Linear Regression	RMSE	GBM performed well
Flow prediction	**Ling et al. (2018), China**	Historical average (HA), multilayer perception neural network (MLP), SVM, GBRT	RMSE, MAPE	SVR and MLR performed best
Crash prediction	**Hossain and Muromachi (2011), Japan**	CART, RF	OOB error, misclassification rate	RF was able to predict better
	Pande et al. (2011), USA	MLPNN, RF	Impurity	Performance of RF was better than MLNPP

Note: *RF- Random Forest, RFDT- Random Forest Decision Tree, MNL- Multi Nominal Logistic, SVM- Support Vector Machine, GBM- Gradient Boosting Machine, DNN- Deep Neural Network, RT- Regression Tree, GBDT- Gradient Boosting Decision Tree, BPNN- Back-Propagation Neural Networks, XGB- Extreme Gradient Boosting, GBRT- Gradient Boosting Regression Tree, RF- Random Forest, CART- Classification & Regression Trees, MPLNN- Minimum Parameter Learning of Neural Network, MAE- Mean Absolute Error, RMSE- Root Mean Square Error, RAE- Relative Absolute Error, MAPE- Mean Absolute Percentage Error, MAD- Mean Absolute Deviation, RE- Relative Error, MSE- Mean Squared Error, AUC- Area Under Curve, OOB- Out of Bag*

The current LGBM and GBM models showed overall good prediction accuracy ($MAE \leq 10\%$). As per past studies, better prediction accuracy was obtained using different boosting-based algorithms for different study domains. For example, the study of Ha et al. (2019) showed that GBM could achieve an astonishing 95.1% accuracy in travel behavior prediction. Ding et al. (2016) studied the stop or run behavior of drivers in China and reported that GBM could be useful when data had complex non-linear patterns. The other studies proved the efficiency of boosting-based algorithms in different study domains, as mentioned in Table 14.

Similar to other domains, in pedestrian-based researches where accurate pedestrian macroscopic behavior (speed and flow) prediction is required, these algorithms could provide an accurate solution. They would help in smooth management of busy facilities such as bus, train or airport terminals. In this regard, the current study results tried to fill this gap and showed the effectiveness of such algorithms in pedestrian-based research, which could act as a better alternative when model quality (or accuracy) is the main goal.

8. Variable Importance Analysis

As discussed in Table 13, the LGBM (FOB) and GBM (skywalk) models were found to perform best on the test dataset as well. The main advantage of a tree-based regressor is that it provides the global importance scores of each feature which explains the contribution of different predictors in the model. Still, these high-end black-box models lack interpretability as they do not provide the direction of impact, i.e., whether the model variables have a positive or negative influence. Thus, to trust a black-box model, the understanding of inner workings is essential. Lundberg and Lee (2017) proposed the SHAP (SHapley Additive exPlanations) values method which is fast and offers a high level of interpretability for a model. In the present study, the "shap" python library was utilized to

interpret the existing trained models. The SHapley values were estimated for tuned LGBM (i.e., for FOBs) and GBM (i.e., for skywalks) on the test data and were plotted using a summary plot (refer to Figs. 3 and 4).

The summary plot not only provides the variable importance in descending order but also illustrates the positive or negative relationship with the outcome variable. The y-axis shows different variables (top five predictors) while the x-axis shows SHAP values ranging from -ve to + ve. The feature value is illustrated with blue and red color gradients. The red color indicates a high feature value, while blue indicates a low feature value.

The feature importance plot of the FOB model (refer to Fig. 3) illustrates crucial findings for the top five factors which influence walking speed over FOBs. As per Fig. 3, the total length of the facility, average density, average flow, facility height, and mid-block width are the top five parameters that impact the pedestrian walking speed over the FOB facilities. The most important feature is the length of the facility which determines the walking speed. In FOBs, after climbing the stairs (in most of the FOBs considered, stairways were the only form of vertical connectivity), pedestrians feel tired. Due to this tiredness, the pedestrian speed was initially observed to be a little slower. However, with the increase in length of the FOB as the pedestrians approach the middle section, this impact on pedestrian speed towards the middle portion of the FOB (where the data was collected) does not show much variability in speed. From Fig. 3 it is observed that density values present a wider range and have a negative relationship with walking speed. As the average density increases, the space for faster movement reduces and thus pedestrians' walking speed reduces. The impact of width and height of the facility on the pedestrian speed is not clear, and this necessitates data requirement over a wider range of facilities to establish a concrete relationship. The impact of the flow parameter reflected here (as well as in Fig. 4 for skywalk facilities) is somewhat unclear or contradictory. Such behavior of flow is reflected as few sites for both elevated facilities (FOBs as well as skywalks) were in congested conditions (i.e. speed increases with increase in flow under congested regime), as opposed to most of the other sites which were in free flow condition with lower densities. Observation of pedestrian speed data over a wide range of densities in most sites might resolve this ambiguity.

The feature importance plot of the skywalk model illustrated that the top five parameters impacting the walking speed over skywalks were the average flow, average density, gender (male/ female), age (< 10, 11–20, 21–40, 41–60, and > 60 years), and length of the facility (refer to Fig. 4). Males were found to walk faster compared to female pedestrians. Moreover, the proportion of male pedestrians leads to higher stream speed. Similarly, pedestrians belonging to age group of 21–40 (young adults) walked faster than any other age group. The old (> 60 years) and young (< 10 years) pedestrians are observed to negatively impact the relative stream speed. Thus with a higher proportion of old and young pedestrians, the overall stream speed would be significantly reduced. The total length of the facility although is found significant, however, its direction of influence on the walking speed is not clear.

9. Conclusions, Limitations, And Future Recommendations

The current study focuses on the accurate prediction of factors impacting pedestrian walking speed over elevated facilities. The observed data of pedestrian behavior were collected using the videography survey method over 13 Foot Over Bridges (FOBs) and 7 skywalk locations across different land-use types in India. The different factors considered for speed prediction modeling were microscopic factors (demographics characteristics), macroscopic factors (average flow and density), and geometric factors (obstruction, land use type, length, connectivity, and effective width). In total 7,522 observations of FOB pedestrian data and 5,342 observations of skywalk pedestrian data were utilized to model the walking speed of pedestrians. This study conducted a comparative analysis of different tree-based models like Gradient Boosting Machine (GBM), Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (Adaboost), Random Forest (RF), Extra Tree Regressor (ETR) and Decision Tree (DT) to obtain a model that predicts pedestrian walking speed accurately.

The major findings of the current study are as follows:

i) Demographic characteristics showed that majority of the pedestrians using the elevated facilities were male pedestrians (FOBs: 69.24%; skywalks: 76.98%) of age 21–40 years (FOBs: 64.44%; skywalks: 66.44%).

ii) The majority of the pedestrians on the skywalk were observed to carry luggage (77.91%) in comparison to pedestrians on FOBs (51.52%). Further, a small proportion of pedestrians (FOBs: 7.81%; skywalks: 14.93%) were observed using a mobile phone while walking on both facilities.
iii) Light Gradient Boosting Machine (with MAE: 9.96) and Gradient Boosting Machine (with MAE: 9.27) provide the best prediction accuracy of pedestrian walking speed models over FOB and skywalk facilities respectively.
iv) Variable importance for both elevated facilities revealed that average flow and average density were extremely important to predict the walking speed.
v) FOBs variable importance plot revealed that the length of the facility influences the walking speed positively while the reduction in available walkway width reduces the pedestrian walking speed.
vi) Skywalk variable importance plot revealed that pedestrian demographics (gender and age) were important predictors for walking speed. Male pedestrians walked faster than female pedestrians, and a higher proportion of male pedestrians led to higher stream speeds. The young adults (aged between 21–40 years) had the highest walking speed among all age groups. Similarly, locations with a higher proportion of young (< 10 years) and old (> 60 years) pedestrians significantly reduced the overall walking speed of the facilities.

The identification of important variables not only provides better insight on factors that affect walking speed over elevated facilities but also provides a valuable source of information to researchers, planners, and policymakers for better design, operate and manage elevated pedestrian infrastructures.

Similar to other studies this study also has some limitations. Some of the significant challenges were: duration of data collection (restricted to a single day observation for 3 hours) and the number of locations covered.

Future studies could be extended by including a larger spectrum of facilities and covering more location types (e.g. recreational) across major Indian cities. Further, studies can be carried out including both observational and survey data.

10. Abbreviations

Table 1: Acronyms/ abbreviations

Acronyms	Full form	Acronyms	Full form
AdaBoost	Adaptive Boosting Regressor	LSR	Least Squares Regression
ANN	Artificial Neural Network	MAD	Mean Absolute Deviation
AUC	Area Under Curve	MAE	Mean Absolute Error
BLR	Bayesian Linear Regression	MAPE	Mean Absolute Percentage Error
BPNN	Back-Propagation Neural Networks	ML	Machine Learning
CART	Classification & Regression Trees	MNL	Multi Nominal Logistic
CHAID	Chi-square Automatic Interaction Detector	MPLNN	Minimum Parameter Learning of Neural Network
CV	Cross-Validation	MSE	Mean Squared Error
DCNN	Deep Convolution Neural Network	MVR	Multivariate Regression
DNN	Deep Neural Network	OOB	Out of Bag
DT	Decision Tree	RAE	Relative Absolute Error
ETR	Extra Tree Regressor	RE	Relative Error
FOB	Foot Over Bridge	RF	Random Forest
GBDT	Gradient Boosting Decision Tree	RFDT	Random Forest Decision Tree
GBM	Gradient Boosting Machine	RMSE	Root Mean Square Error
GBR	Gradient Boosting Regressor	RMSLE	Root Mean Squared Log Error
GBRT	Gradient Boosting Regression Tree	RT	Regression Tree
GRP	Gaussian Process Based Regression	SHAP	SHapley Additive exPlanations
HTBRM	Hierarchical Tree-Based Regression Model	SLR	Stepwise Linear Regression
LGBM	Light Gradient Boosting Machine	SVM	Support Vector Machine
LR	Linear Regression	XGBoost/XGB	Extreme Gradient Boosting

11. Declarations

Funding

Not applicable.

Conflict of interest/ Competing interests

The authors declare that there is no actual or potential conflict of interest in relation to this article.

Availability of data and material

Not applicable.

Code availability

Not applicable.

Authors' contributions

Arunabha Banerjee: Data Collection, Conceptualization, Methodology, Writing Original Draft, Review & Editing.

Rahul Raoniar: Conceptualization, Methodology, Formal Analysis, Investigation, Visualization, Review & Editing.

Akhilesh Kumar Maurya*:* Conceptualization, Review & Editing, Supervision, Project administration.

12. References

1. Abojaradeh M (2013) Evaluation of pedestrian bridges and pedestrian safety in Jordan. Civil and Environmental Research, 3(1): 66-79.
2. Alajali W, Zhou W, Wen S, Wang Y (2018) Intersection traffic prediction using decision tree models. Symmetry, 10(9): 386. https://doi.org/10.3390/sym10090386.
3. Ali M (2020) PyCaret: An open source, low-code machine learning library in Python. https://doi.org/10.1101/2020.06.26.174524.
4. Al-Azzawi M, Raeside R (2007) Modeling pedestrian walking speeds on sidewalks. Journal of Urban Planning and Development, 133(3): 211-219. https://doi.org/10.1061/(ASCE)0733-9488(2007)133:3(211).
5. Anciaes PR, Jones P (2018) Estimating preferences for different types of pedestrian crossing facilities. Transportation Research Part F: Traffic Psychology and Behaviour, 52: 222-237. https://doi.org/10.1016/j.trf.2017.11.025.
6. Arasan VT, Rengaraju VR, Rao KK (1994) Characteristics of trips by foot and bicycle modes in Indian city. Journal of Transportation Engineering, 120(2): 283-294. https://ascelibrary.org/doi/abs/10.1061/(ASCE)0733-947X(1994)120:2(283)
7. Asaithambi G, Kuttan MO, Chandra S (2016) Pedestrian road crossing behavior under mixed traffic conditions: A comparative study of an intersection before and after implementing control measures. Transportation in Developing Economies, 2(2): 14. https://doi.org/10.1007/s40890-016-0018-5.
8. Bansal A, Goyal T, Sharma U (2019) Modelling the Pedestrian Speed at Signalised Intersection Crosswalks for Heterogeneous Traffic Conditions. Promet-Traffic & Transportation, 31(6): 681-692. https://doi.org/10.7307/ptt.v31i6.3299.
9. Bargegol I, Gilani VNM, Farghedayn S (2014) Analysis of the effect of vehicles conflict on pedestrian's crossing speed in signalized and un-signalized intersection. Advances in Environmental Biology, pp.502-510.
10. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13(1): 281-305.
11. Bowman BL, Vecellio RL (1994) Pedestrian walking speeds and conflicts at urban median locations. Transportation Research Record 1438, Transportation Research Board, Washington, DC, pp.67–73. http://onlinepubs.trb.org/Onlinepubs/trr/1994/1438/1438-009.pdf.
12. Breiman L, Friedman J, Charles JS, Richard AO (1984) Classification and regression trees. CRC press.
13. Breiman L (2001) Random forests. Machine learning, 45(1): 5-32. https://doi.org/10.1023/3A1010933404324.
14. Byun S, Lee HJ, Han JW, Kim JS, Choi E, Kim KW (2019) Walking-speed estimation using a single inertial measurement unit for the older adults. PLoS One, 14(12): p.e0227075. https://doi.org/10.1371/journal.pone.0227075.
15. Cepolina EM, Menichini F, Rojas PG (2018) Level of service of pedestrian facilities: Modelling human comfort perception in the evaluation of pedestrian behaviour patterns. Transportation Research Part F: Traffic Psychology and Behaviour, 58: 365-381. https://doi.org/10.1016/j.trf.2018.06.028.
16. Chang CY, Woo TH, Wang SF (2011) Analysis of pedestrian walking speeds at crosswalks in Taiwan. Journal of the Eastern Asia Society for Transportation Studies, 9: 1186-1200. https://doi.org/10.11175/easts.9.1186.
17. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. https://doi.org/10.1145/2939672.2939785.
18. Cheng J, Li G, Chen X (2019a) Research on travel time prediction model of freeway based on gradient boosting decision tree. IEEE Access, 7: 7466-7480. https://doi.org/10.1109/ACCESS.2018.2886549.
19. Cheng L, Chen X, De Vos J, Lai X, Witlox F (2019b) Applying a random forest method approach to model travel mode choice behavior. Travel Behaviour and Society, 14:1–10. https://doi.org/10.1016/j.tbs.2018.09.002
20. Demiroz YI, Onelcin P, Alver Y (2015) Illegal road crossing behavior of pedestrians at overpass locations: factors affecting gap acceptance, crossing times and overpass use. Accident Analysis & Prevention, 80: 220-228. https://doi.org/10.1016/j.aap.2015.04.018.
21. Ding C, Wang D, Ma X, Li H (2016) Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. Sustainability, 8(11): 1100. https://doi.org/10.1016/j.trc.2016.09.016.
22. Ermagun A, Rashidi TH, Lari ZA (2015) Mode choice for school trips: long-term planning and impact of modal specification on policy assessments. Transportation Research Record, 2513(1): pp.97-105. https://doi.org/10.3141/2513-12.
23. Finnis KK, Walton D (2008) Field observations to determine the influence of population size, location and individual factors on pedestrian walking speeds. Ergonomics, 51(6): 827-842. https://doi.org/10.1080/00140130701812147.
24. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1): 119-139. https://doi.org/10.1006/jcss.1997.1504.
25. Friedman J, Hastie T, Tibshirani R (2000) Special invited paper. Additive logistic regression: A statistical view of boosting. Annals of Statistics, pp.337-374. 10.1214/aos/1016218223.
26. Fruin JJ (1971) Pedestrian planning and design. Metropolitan association of urban designers and environmental planners, New York, pp.2-6. https://www.elevatorbooks.com/shop/construction-design/pedestrian-planning-and-design/.
27. Gates TJ, Noyce DA, Bill AR, Van Ee N (2006) Recommended walking speeds for timing of pedestrian clearance intervals based on characteristics of the pedestrian population. Transportation Research Record, 1982(1): 38-47. https://doi.org/10.1177%2F0361198106198200106.
28. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Machine Learning, 63(1): 3-42. https://doi.org/10.1007/s10994-006-6226-1.
29. Guo G, Chen R, Ye F, Chen L, Pan Y, Liu M, Cao Z (2019) A pose awareness solution for estimating pedestrian walking speed. Remote Sensing, 11(1): 55. http://dx.doi.org/10.3390/rs11010055.

30. Ha TV, Asada T, Arimura M (2019) The Application of Gradient Boost Machine and Local Interpretable Modelagnostic Explanations in Examining the Travel Multi-mode Choice, the Case Study of Jakarta City, Indonesia. Journal of the Eastern Asia Society for Transportation Studies, 13: 503-522. https://doi.org/10.11175/easts.13.503.
31. Hasan R, Napiah M (2018) The perception of Malaysian pedestrians toward the use of footbridges. Traffic injury prevention, 19(3): 292-297. https://doi.org/10.1080/15389588.2017.1373768.
32. Herrera-Angulo A, Zenteno-Bolanos E (2018) A Low-Cost Microwave System for Pedestrian Speed Estimation. In 2018 IEEE MTT-S Latin America Microwave Conference (LAMC 2018), pp. 1-3. https://doi.org/10.1109/LAMC.2018.8699061.
33. Hoogendoorn SP, Daamen W (2005) Pedestrian behavior at bottlenecks. Transportation science, 39(2): 147-159. https://doi.org/10.1287/trsc.1040.0102.
34. Hossain M, Muromachi Y (2011) Understanding crash mechanisms and selecting interventions to mitigate real-time hazards on urban expressways. Transportation Research Record, 2213(1): 53-62. https://doi.org/10.3141%2F2213-08.
35. Karatas P, Tuydes-Yaman H (2018) Variability in sidewalk pedestrian level of service measures and rating. Journal of Urban Planning and Development, 144(4): 04018042. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000483.
36. Kawaguchi N, Nozaki J, Yoshida T, Hiroi K, Yonezawa T and Kaji K (2019) End-to-end walking speed estimation method for smartphone PDR using DualCNN-LSTM. In IPIN (Short Papers/Work-in-Progress Papers), pp. 463-470.
37. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, pp. 3146-3154.
38. Knoflacher H, New York Pedestrian Study, New York City (2006) Pedestrian Level of Service, Phase I, Dept. of City Planning, Transportation Division, New York.
39. Laxman KK, Rastogi R, Chandra S (2010) Pedestrian flow characteristics in mixed traffic conditions. Journal of Urban Planning and Development, 136(1): 23-33. https://doi.org/10.1061/(ASCE)0733-9488(2010)136:1(23).
40. Lee JY, Lam WH (2006) Variation of walking speeds on a unidirectional walkway and on a bidirectional stairway. Transportation Research Record, 1982(1): 122-131. https://doi.org/10.1177%2F0361198106198200116.
41. Ling X, Huang Z, Wang C, Zhang F, Wang P (2018) Predicting subway passenger flows under different traffic conditions. PLoS One, 13(8): e0202707. https://doi.org/10.1371/journal.pone.0202707.
42. Liu MW, Wang SM, Oeda Y, Sumi TN (2019) Simulating uni-and bi-directional pedestrian movement on stairs by considering specifications of personal space. Accident Analysis & Prevention, 122: 350-364. https://doi.org/10.1016/j.aap.2017.11.012.
43. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pp. 4765-4774.
44. Marisamynathan S, Lakshmi S (2016). Performance analysis of signalized intersection at metropolitan area. Journal of Advanced Research in Applied Sciences and Engineering Technology, 2(1): 19-29.
45. Marisamynathan S, Vedagiri P (2018) Modeling pedestrian crossing behavior and safety at signalized intersections. Transportation Research Record, 2672(31): 76-86. https://doi.org/10.1177%2F0361198118759075.
46. Matsubayashi M, Shiraishi Y (2016) A method for estimating walking speed by using magnetic signature to grasp people flow in indoor passages. In Adjunct Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing Networking and Services, pp. 94-99. https://doi.org/10.1145/3004010.3004033.
47. Morrall JF, Ratnayake LL, Seneviratne PN (1991) Comparison of central business district pedestrian characteristics in Canada and Sri Lanka. Transportation Research Record, 1294. http://onlinepubs.trb.org/Onlinepubs/trr/1991/1294/1294-010.pdf.
48. Mousa SR, Bakhit PR, Osman OA, Ishak S (2018) A Comparative Analysis of Tree-Based Ensemble Methods for Detecting Imminent Lane Change Maneuvers in Connected Vehicle Environments. Transportation Research Record, 2672(42): 268-279. https://doi.org/10.1177%2F0361198118780204.
49. Navin FP, Wheeler RJ (1969) Pedestrian flow characteristics. Traffic Engineering, Inst Traffic Engr, 39.
50. Nazir MI, Adhikary SK, Hossain QS, Ali SA (2012) Pedestrian flow characteristics in Khulna metropolitan city, Bangladesh. Journal of Engineering Science, 3(1): 25-31.
51. Oeding D (1963) Verkehrsbelastung und Dimensionierung von Gehwegen und anderen Anlagen des FuÞga«ngerverkehrs [Traffic volume and dimensioning of footways and other facilities of pedestrian traffic], StraÞenbau und StraÞenverkehrstechnik series number 22, Ministry of Traffic, Bonn.
52. Older SJ (1968) Movement of pedestrians on footways in shopping streets. Traffic engineering & control, 10(4).
53. Oviedo-Trespalacios O, Scott-Parker B (2017) Footbridge usage in high-traffic flow highways: The intersection of safety and security in pedestrian decision-making. Transportation research Part F: Traffic Psychology and Behaviour, 49: 177-187. https://doi.org/10.1016/j.trf.2017.06.010.
54. Pande A, Das A, Abdel-Aty M, Hassan, H (2011) Estimation of real-time crash risk: are all freeways created equal?. Transportation Research Record, 2237(1): 60-66. https://doi.org/10.3141%2F2237-07.
55. Park JG, Patel A, Curtis D, Teller S, Ledlie J (2012) Online pose classification and walking speed estimation using handheld devices. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 113-122. https://doi.org/10.1145/2370216.2370235.
56. Poei EP, Hanzawa Y, Koyama S (1995) Pedestrian flow analysis at Yogyakarta, Indonesia. Journal of Japan Society of Civil Engineers, 18(1): 131-134.
57. Polus A, Schofer JL, Ushpiz A (1983) Pedestrian flow and level of service. Journal of Transportation Engineering, 109(1): 46-56. https://doi.org/10.1061/(ASCE)0733-947X(1983)109:1(46).
58. Räsänen M, Lajunen T, Alticafarbay F, Aydin C (2007) Pedestrian self-reports of factors influencing the use of pedestrian bridges. Accident Analysis & Prevention, 39(5): 969-973. https://doi.org/10.1016/j.aap.2007.01.004.

59. Rastogi R, Thaniarasu I, Chandra S (2013) Pedestrian flow characteristics for different pedestrian facilities and situations. European Transport\Trasporti Europei, Issue 53, Paper n° 6, ISSN 1825-3997.
60. Regue R, Recker W (2014) Proactive vehicle routing with inferred demand to solve the bikesharing rebalancing problem. Transportation Research Part E: Logistics and Transportation Review, 72: 192-209. https://doi.org/10.1016/j.tre.2014.10.005.
61. Rengarasu TM, Jayawansa HN, Perera GPW (2012) Estimation of Pedestrian walking speeds at controlled cross walks in Sri Lanka- A pilot study. Presented at International Symposium on Advances in Civil and Environmental Engineering Practices for Sustainable Development (ACEPS 2012).
62. Sahani R, Bhuyan PK (2017) Pedestrian level of service criteria for urban off-street facilities in mid-sized cities. Transport, 32(2): 221-232. https://doi.org/10.3846/16484142.2014.944210.
63. Sarsam SI (2013) Assessing Pedestrian flow characteristics at Baghdad CBD area. Civil Engineering Department Geotechnical and Transportation Engineering, pp.120.
64. Sekhar C, Madhu E (2016) Multimodal Choice Modeling Using Random Forest Decision Trees. International Journal for Traffic & Transport Engineering, 6(3). http://dx.doi.org/10.7708/ijtte.2016.6(3).10.
65. Shrestha A, Won M (2018) Deepwalking: Enabling smartphone-based walking speed estimation using deep learning. In 2018 IEEE Global Communications Conference (GLOBECOM), pp. 1-6. https://doi.org/10.1109/GLOCOM.2018.8647857.
66. Sukhadia H, Dave SM, Shah J, Rathva D (2016) The effect of events on pedestrian behavior and its comparison with normal walking behavior in CBD area in Indian context. Transportation Research Procedia, 17: 653-663. https://doi.org/10.1016/j.trpro.2016.11.120.
67. Tarawneh MS (2001) Evaluation of pedestrian speed in Jordan with investigation of some contributing factors. Journal of Safety Research, 32(2): 229-236. https://doi.org/10.1016/S0022-4375(01)00046-9.
68. Tordeux A, Chraibi M, Seyfried A, Schadschneider A (2020) Prediction of pedestrian dynamics in complex architectures with artificial neural networks. Journal of Intelligent Transportation Systems, pp.1-13. https://doi.org/10.1080/15472450.2019.1621756.
69. Truong LT, Nguyen HT, Nguyen HD, Vu HV (2019) Pedestrian overpass use and its relationships with digital and social distractions, and overpass characteristics. Accident Analysis & Prevention, 131: 234-238. https://doi.org/10.1016/j.aap.2019.07.004.
70. Vathsangam H, Emken A, Spruijt-Metz D, Sukhatme GS (2010) Toward free-living walking speed estimation using Gaussian process-based regression with on-body accelerometers and gyroscopes. In 2010 4th International Conference on Pervasive Computing Technologies for Healthcare, IEEE, pp.1-8. https://doi.org/10.4108/ICST.PERVASIVEHEALTH2010.8786.
71. Weidmann U (1993) Transport technique of pedestrian. Schriftenreihe Ivt-Berichte, 90.
72. Wilson DG, Grayson GB (1980) Age-related differences in the road crossing behaviour of adult pedestrians. Report No. TRRL-LR-933, Transport Research Laboratory, UK.
73. Yang S, Wu J, Du Y, He Y, Chen X (2017) Ensemble learning for short-term traffic prediction based on gradient boosting machine. Journal of Sensors. https://doi.org/10.1155/2017/7074143.
74. Zhang G, Chen Y, Wu D, Li P (2011) Study on pedestrian traffic characteristics of transfer passageways in subway transfer stations. In ICCTP 2011: Towards Sustainable Transportation Systems, pp. 2768-2781. https://doi.org/10.1061/41186(421)276.
75. Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58: 308-324. https://doi.org/10.1061/41186(421)276.

Figures

Figure 1

Geometric description along with camera position

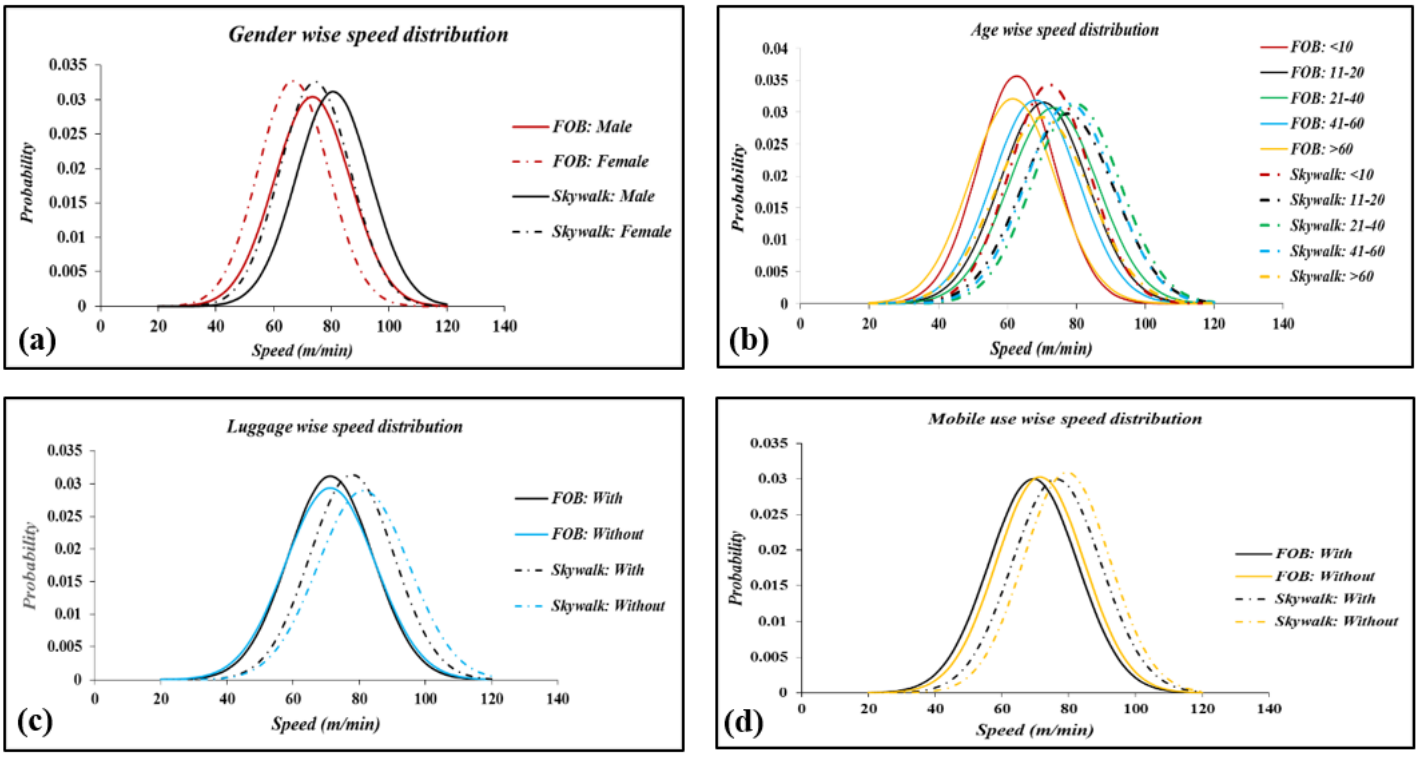

Figure 2

Variation of pedestrian speed with (a) Gender, (b) Age, (c) Luggage, and (d) Mobile use

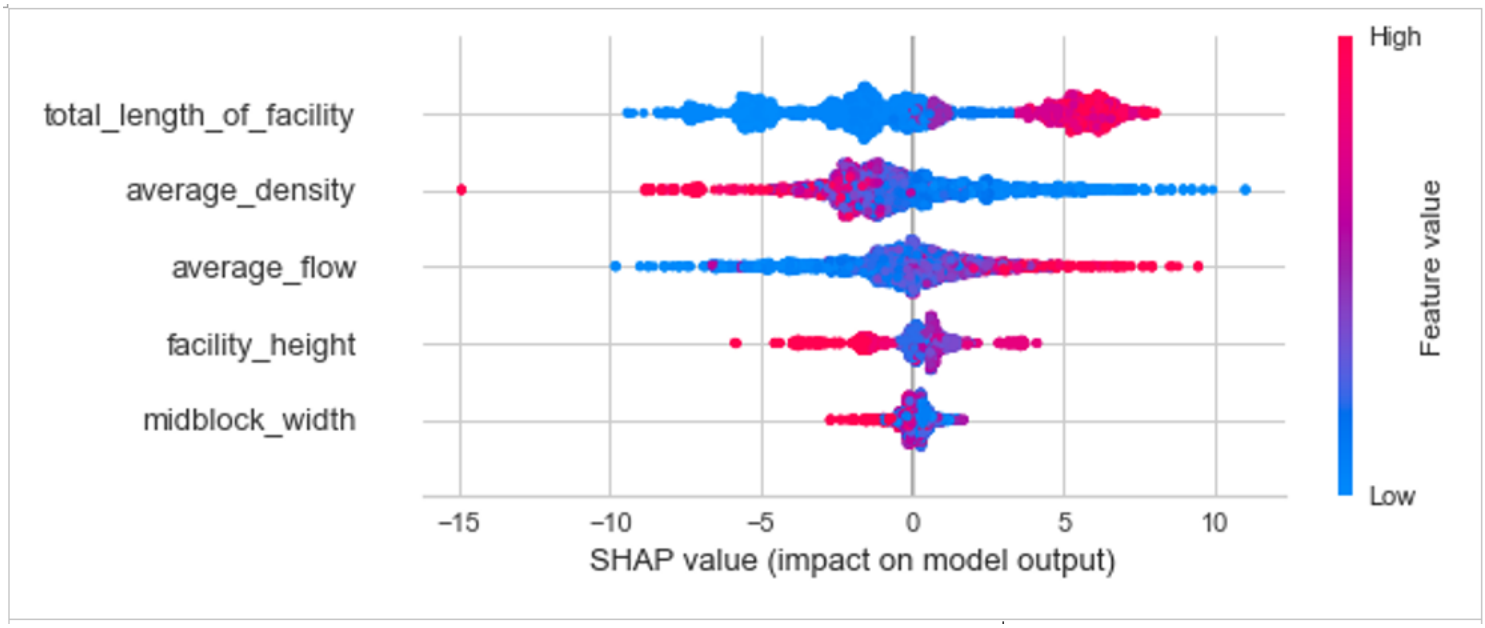

Figure 3

Variable importance plot for FOB walking speed model

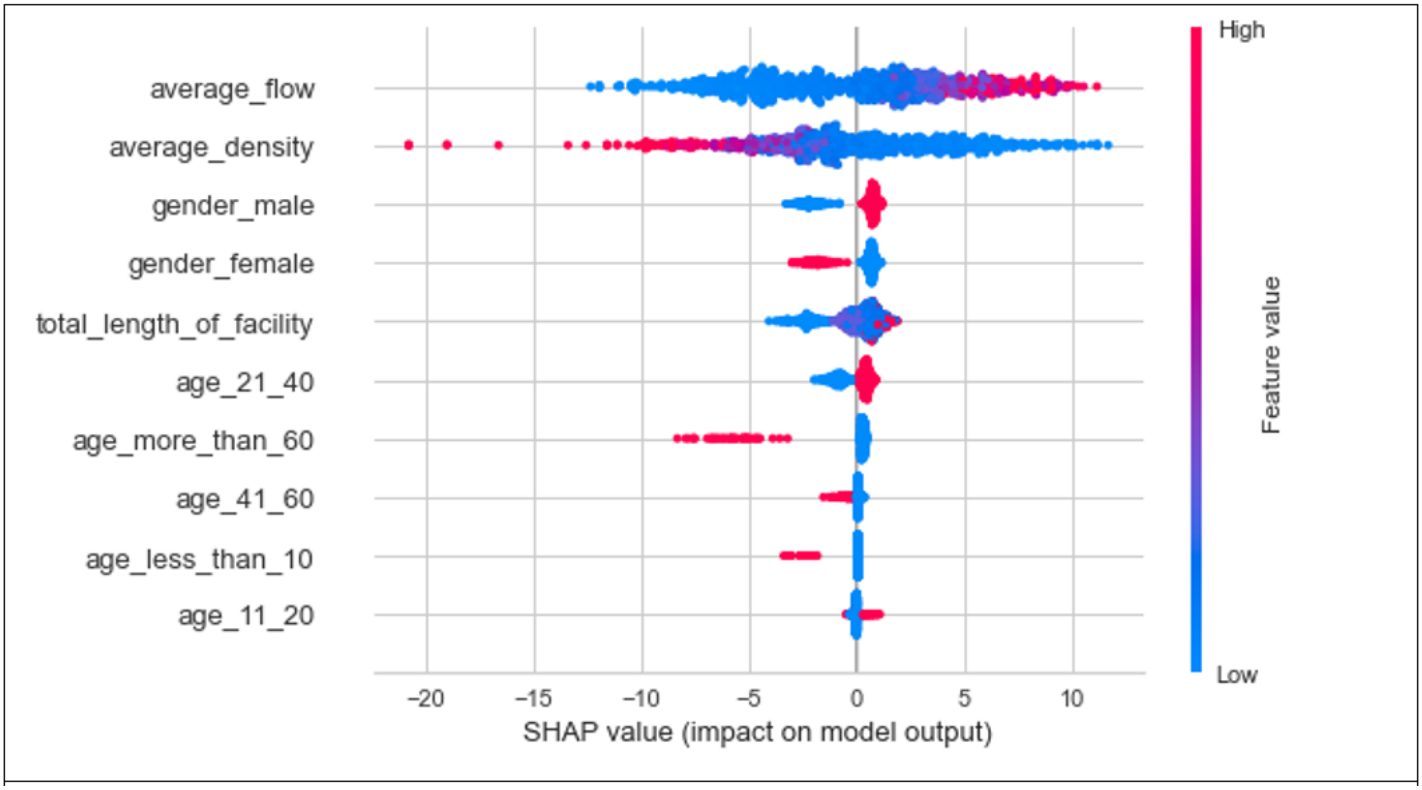

Figure 4

Variable importance plot for skywalk walking speed model