Journal of Building Engineering 70 (2023) 106365

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

Interpretable machine learning models for the estimation of seismic drifts in CLT buildings

Eknara Junda [a], Christian Málaga-Chuquitaype [a,*], Ketsarin Chawgien [b]

[a] Department of Civil and Environmental Engineering, Imperial College, London, SW7 2AZ, UK
[b] Program of Intelligent Control Innovation, Nakhon Pathom Rajabhat University, Nakhon Pathom, 73000, Thailand

A B S T R A C T

An accurate estimation of drift demands is crucial for designing and assessing structures under seismic loads. Given the novelty of massive timber buildings, predictive models for the estimation of drifts in mid-to high-rise CLT structures are lacking, particularly in the form of simple models suitable for preliminary design evaluations or regional seismic assessments. In this paper, we present and compare several Machine Learning (ML) models for the estimation of peak inter-storey and roof drifts in multi-storey Cross-Laminated Timber (CLT) walled structures. The ML techniques used include: Multiple Linear Regression, Regression Trees, Random Forest, K-nearest Neighbor, and Support Vector Regression. To this end, 69 structures spanning mid-rise to tall timber buildings are subjected to a large collection of acceleration records and used to create the training and testing datasets. Different structural configurations and behaviour factors, related to the assumed energy dissipation capacity of the buildings, are considered. A diversity of feature selection techniques informs our choice of parameters to the reduced input space leading to a set of six most efficient features: the spectral acceleration at the building's fundamental period ($S_a(T_1)$), the Peak Ground Velocity (PGV), tuning ratio (T_1/T_m), behaviour factor (q), wall height (H_w), and the wall subdivision ratio (W_r). After verifying the high accuracy of our model predictions, the SHapley Additive exPlanation method (SHAP) is used to gain insight into the influence of key input features on the ML model outputs. Finally, our ML drift estimations are compared against previous proposals and design code assumptions, and the potential causes of disagreement are discussed.

1. Introduction

Cross-Laminated Timber (CLT) is a sustainable engineered material that is gaining popularity in the building sector where it is used in load-bearing elements such as walls or slabs. Its comparative advantages include a low carbon footprint, high prefabrication potential, reduced construction times, high tolerances, and a high strength-to-weight ratio. These advantages have made CLT an attractive option for mid-rise and tall timber buildings. Besides, since CLT panels are formed from glued timber sections laid up in layers at right angles from each other, they have characteristically high in-plane strength and stiffness and oftentimes constitute the lateral force-resisting system of their host structures [1]. The design of CLT structures is also being included in the revision process of the Eurocodes [2,3].

The numerous sustainability and structural advantages of CLT have led to a worldwide expansion of the system's use, including in earthquake-prone regions. However, the seismic response of CLT buildings is fundamentally different from that of other, more common, structural systems such as moment frames or reinforced concrete walls. For example, while the deformations of a moment-resisting frame are generally governed by the flexural behaviour of its elements, a CLT building deforms based on a combination of rocking and sliding mechanisms [4]. Moreover, as CLT panels are stiff but fragile, ductile components such as steel connectors are required to increase the ductility and energy dissipation capacity of CLT structures. All these properties call for dedicated models to predict the seismic response of CLT buildings.

* Corresponding author.
E-mail address: c.malaga@imperial.ac.uk (C. Málaga-Chuquitaype).

https://doi.org/10.1016/j.jobe.2023.106365
Received 11 December 2022; Received in revised form 8 March 2023; Accepted 22 March 2023
Available online 3 April 2023

To date, various research studies have investigated the hysteretic response of CLT walls and joints under seismic loads [5–7]. These experiments have shown that almost all the hysteretic energy dissipation in CLT walls concentrates at the steel connectors while the CLT panels remain largely elastic. The experimental evidence also points towards a complex coupling of rocking and sliding in CLT walls and their response is further complicated by pinching, friction, stiffness and strength degradation at the connection level. Therefore, as the number of mid-rise and tall CLT buildings increases worldwide and more cases of earthquake-hit timber buildings are expected, accurate models for predicting and assessing their seismic deformations are needed.

In this context, seismic drift demands are essential for the assessment of structures under earthquakes. In particular, the maximum inter-storey drift ratio (*MIDR*) and the global roof drift ratio (*RDR*) can be directly related to building damage and subsequently to cost. Although nowadays drift demands can be accurately estimated on a case-specific basis using nonlinear response history analyses (NRHAs) under a correctly selected suite of ground motions, NRHA still involves high computational costs, long execution times, as well as a good level of modelling expertise. This is especially true in the case of CLT structures, whose structural modelling demands a deep understanding of their intricate seismic response, including the pinching and degradation phenomena mentioned above. These aspects stress the significant need for alternatives to NRHA in the way of simplified estimation methods of seismic drift demands like those proposed in this study. This need is particularly dire in the case of simplified models for preliminary design iterations or building population-level assessments.

Although many simplified seismic drift predictive models have been proposed for steel [8,9] or concrete structures [10,11], to our knowledge, simplified models for seismic demand estimations in timber buildings are rare. For example, Heresi and Miranda [12] developed an expression for the spatial correlation of 5%-damped average spectral pseudo accelerations for *MIDR* estimations on low-rise single-family wood-frame constructions for regional risk assessments. In the case of CLT structures, in particular, Demirci et al. [13] constructed nonlinear regression models for the seismic drift estimation in multi-storey CLT buildings and highlighted the influence of the ground-motion frequency content on the drift demands of CLT walls. However, although regression models can reduce significantly the computational and time demands compared to NRHA, their predictions might be associated with considerable errors as simple functional forms are usually employed in their development. Moreover, traditional regression usually requires a good level of background knowledge on the phenomena at hand to be able to define *a priori* suitable functional forms and relationships among parameters. This makes regression burdensome for dealing with complex problems involving high-dimensional data and intricate parameter relationships which, on the other hand, can be more easily mapped using Machine Learning (ML) techniques [14]. Besides overcoming the constraints associated with the selection of simple functional forms used in traditional methods, ML models and can also be employed to inform the choice of features to be used as predictors of the target response parameter.

The application of ML techniques in seismic assessment and structural design has increased over recent years [17]. For example, Hwang et al. [18] used various ML techniques to predict the seismic response and structural collapse of reinforced concrete frames. Paral et al. [19] presented a deep learning-based method for the evaluation of semi-rigid connection's conditions in steel frames. Neuyen et al. [20] proposed ML-based prediction models for the estimation of maximum displacements in seismically isolated structures. Although to our knowledge ML techniques have not yet been applied to the estimation of seismic demands in timber buildings, ML has been used to examine the response of timber structures to other actions. For instance, Li et al. [21] used an Artificial Neural Network (ANN) to estimate the long-term damage accumulation and failure probability of timber structural members, while Xin et al. [22] proposed a non-destructive testing method using ML algorithms to predict the density and mechanical properties of ancient timber elements.

In this study, we develop ML-based predictive models for seismic drift demands in CLT-walled structures. Five different ML models are constructed for *MIDR* and *RDR*, which are considered as the target engineering demand parameters (*EDPs*). A database of seismic drift demands is first constructed by performing NRHA on 69 numerical models of mid-rise and tall CLT structures subjected to a large set of ground-motion records. The structures cover a broad spectrum of structural configurations and performance factors including practical ranges of panel fragmentation levels, storey heights, and force reduction factors. The total set of input variables comprises 30 features including intensity measures (*IMs*) and structural parameters (*SPs*). We use a range of feature selection techniques to screen important features which are expected to have the greatest impact on the drift demand prediction and arrive to a final set of six most efficient input features. This leads to the identification of the spectral acceleration at the building's fundamental period ($S_a(T_1)$), the Peak Ground Velocity (*PGV*), the tuning ratio (T_1/T_m), the behaviour factor (q), the wall height (H_w), and the wall subdivision ratio (W_r) as preferred predictors. After demonstrating the good accuracy of the model estimations, we use the SHapley Additive exPlanation method (SHAP) to explain the influence of key input features on the ML-model outputs and gain insight into the mechanics of CLT buildings' seismic response. Finally, the drift estimates of our newly proposed models are compared against outputs from previous research and design codes, and the potential causes of disagreement are discussed.

2. Drifts database development

This section deals with the development of the dataset of CLT buildings to be used in subsequent analyses presented later in this paper. First, the 69 CLT structures considered in this study are introduced. These structures cover a broad range of structural configurations and force reduction factors. The design basis and typical loads employed for their structural design and modelling are provided. Second, the numerical modelling approach adopted is described in detail. Finally, typical examples of the extensive model validation carried out against available experimental results are presented.

2.1. CLT buildings

69 CLT structures representing residential multi-storey CLT buildings of different heights (number of storeys) and panel

fragmentation levels are considered. Fig. 1a shows a 3D view of a typical 8-storey CLT building as an example. All buildings share the same plan layout (Fig. 1b) and have a typical storey height of 3.5 m, except for the ground level which height is 4.3 m, depicted in Fig. 1c. The CLT buildings were designed based on widely accepted capacity design principles and current codified guidance. Ductile design and failure mode control principles were followed in consistency with Eurocode 8, Part 1 [4,23]. Dead loads were calculated considering all finishing and insulation whereas a superimposed load of 2.0 kN/m^2 was assumed for both roof and floors. A combination of 100% dead and 30% superimposed load was used for seismic mass calculations. The response spectrum of Eurocode 8 Type 1 anchored at a Peak Ground Acceleration (*PGA*) of 3.0 m/s^2 was adopted with an importance factor of 1.0. More details of the design processes of the CLT buildings can be found in Ref. [13]. A number of parameters were varied during the design to cover a wide range of CLT buildings and to evaluate the effect of these variations on the corresponding seismic demands. These parameters include the number of storeys, the panel fragmentation level, as well as the behaviour factors. Buildings of 6, 8, 12, 16, and 20 storeys were considered.

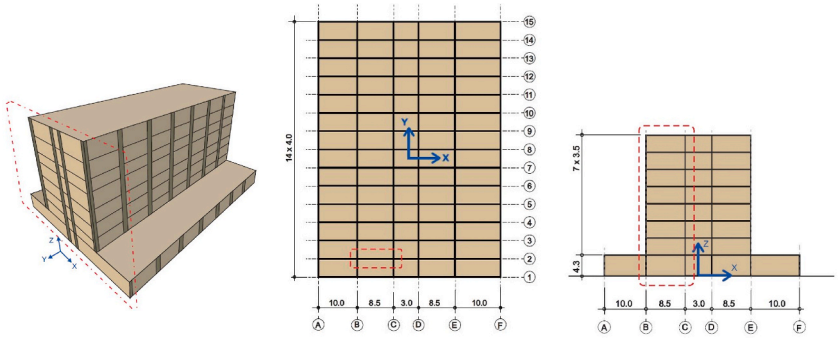
The variation in building height leads to different wall slendernesses. Likewise, the panel fragmentation level affects the joint density since it dictates the number of vertical joints per wall. Besides, the number of sub-panels (W_r) was varied from 1 to 4 per wall, where a $W_r = 1$ corresponds to a single monolithic CLT wall (i.e., no vertical joint is provided) while $W_r = 2$ to 4 indicate the walls are subdivided into 2 to 4 panels, respectively. It should be noted that a single CLT wall will be associated with a low seismic energy dissipation capacity and a correspondingly more brittle failure [24]. However, due to the widespread use of the single walls in low-rise CLT buildings in some parts of Europe, a value of $W_r = 1$ was included in this study but it is limited to structures of 6–12 storeys. For jointed CLT walls (i.e., $W_r \geq 2$), half-lap joints with self-tapping screws were provided for all the vertical step joints between panels. Moreover, to cater for the uncertainties still present around the appropriate q-factor for tall CLT buildings, the designs of the buildings under consideration are not done for a single value of q. Instead, a range of q-factor values (from 1.5 to 4.0) was used to design the CLT structures based on general guidance from EC8's specific rules for timber buildings. Note that a limit of q = 3 was enforced in the cases with single walls ($W_r = 1$) as a low dissipation capacity is expected in these cases. Table 1 summarises the structural configurations of the CLT walls adopted.

2.2. Numerical modelling

The buildings under consideration are regular in plan and elevation, and the CLT walls are symmetrically distributed throughout the buildings in plan and elevation. Assuming that a rigid diaphragm action is ensured, and considering only on the lateral deformations of the building, the use of one representative CLT wall is deemed to provide reasonable estimates of the building's overall lateral drift response. To this end, the portion of the building enclosing the 8.5-m long CLT Wall 2-BC (marked with dashed lines in Fig. 1c was selected as the focus of this study. We developed 2D numerical models in OpenSees [25] representing the selected wall for all buildings under study.

Fig. 2 shows a schematic representation of a typical model for a wall with $W_r = 2$. All CLT panels were discretised into a number of shell elements (*4-node Quad Element* [25]). Two-node link elements (*zero length element* [25]) were used to represent the three main types of steel connectors, namely hold-downs, shear brackets, and vertical screws. The link elements for hold-downs (presented in blue in Fig. 2) were located at the corners of the panels whereas shear brackets were distributed along the length of the panels in the X direction (presented in red). Self-threaded screws at the vertical joints were lumped into an equivalent single link element and placed at the mid-height of the panels (shown in green).

Since most of the nonlinear deformations in CLT walls under lateral actions occur at their joints and since the CLT panels themselves are expected to remain largely elastic [26,27], an equivalent isotropic material model with linear elastic properties was assumed for the CLT panels. To this end, we used a panel stiffness of 10^8 N/mm following the sensitivity study of [28]. In the case of steel connectors (angle brackets, hold-downs, and screws), a *hysteretic uniaxial material model* [25] that captures the global nonlinear responses of the connectors was used. The parameters of this model were calibrated to represent the strength and deformation values as well as pinching strength and stiffness degradation factors of each connector type employed. To this end, pinching factors of 0.7 and 0.3 for strain and stress, respectively, were assigned for angle brackets, while pinching factors of 0.9 (strain) and 0.2 (stress) were used for hold-downs. Importantly, both the shear and axial direction responses were modelled in all brackets and hold-downs considering



(a) 3D view (b) Plan (c) Elevation

Fig. 1. Typical (8-storey) building under consideration.

Table 1
Variation of CLT wall configurations.

Structural Parameters	Range of values
Number of stories (m)	6, 8, 12, 16, 20
Number of subpanels per wall (W_r)	1, 2, 3, 4
Behaviour factor (q)	1.5, 2, 2.5, 3, 4

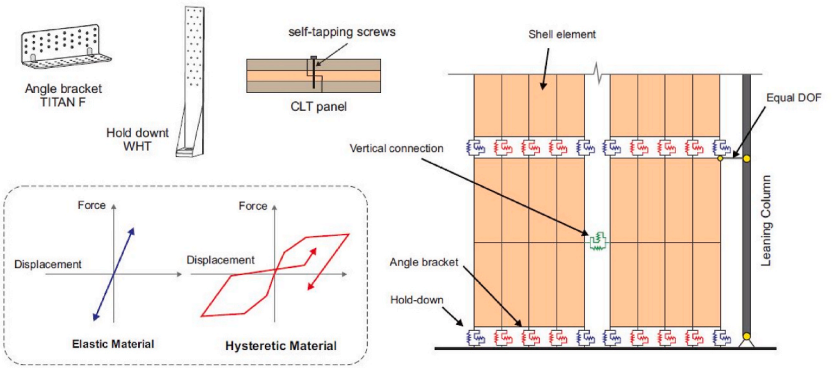


Fig. 2. Numerical model of CLT walls.

un-coupled response between the two directions. This is in contrast with a majority of previous studies that have usually ignored the shear capacity of hold-downs [5] The respective force-deformation relationships were taken from previous experimental studies [5–7] as shown in Table 2. The interested reader can find further particular and specific strength and stiffness values in these references or [28]. Moreover, to account for P-Delta effects caused by gravity loads on largely deformed structures (i.e., geometric non-linearities), a leaning column was added into the models using beam-column elements connected to the wall by means of *equal degree-of-freedom constraints* [25] to simulate a rigid diaphragm action. A Rayleigh damping model with a damping ratio of 0.05 in the first mode was assumed in all cases.

2.3. Finite Element model validation

The Finite Element models described in the previous section were validated against experimental results in a two-stage process. The first stage involved the validation of steel connectors. An example of this is presented in Fig. 3 which compares the experimental hysteretic responses [5] and their numerical counterparts for a BMF 90x116x48x3 bracket and a WHT 540 hold-down. As shown in this figure, the numerical models capture well the key characteristics of the connector's response in terms of strength, stiffness, degradation, and pinching in both directions.

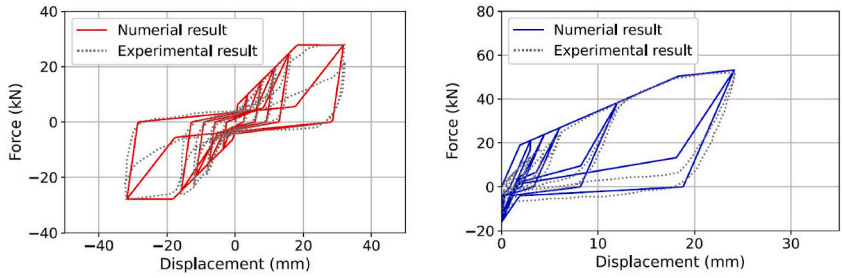
The second validation stage comprised the development of numerical models of CLT walls and wall assemblages. As an example, the results for the two CLT walls (single and coupled) tested by Ref. [26] are presented in Figs. 4 and 5. Fig. 4 compares the experimental and simulated hysteresis of the single wall case along with the cumulative dissipated energy whereas the corresponding comparisons for the coupled walls are shown in Fig. 5. The Solid lines and dotted lines in these figures correspond to the numerical and experimental results, respectively. It can be observed from these figures that our numerical models can reproduce the experimental hysteretic response of CLT assemblages with excellent accuracy.

3. Ground motions and acceleration demand database

In order to generate a large dataset of results using real ground motions without resorting to significant spectral scaling that may introduce undesirable bias [13,29,30], we used the set of 1656 records selected by Hancock et al. [29] from the NGA-West database [31]. These records come from shallow crust seismic events with moment magnitudes between 5.1 and 7.9 and an average *PGA* of 1 g.

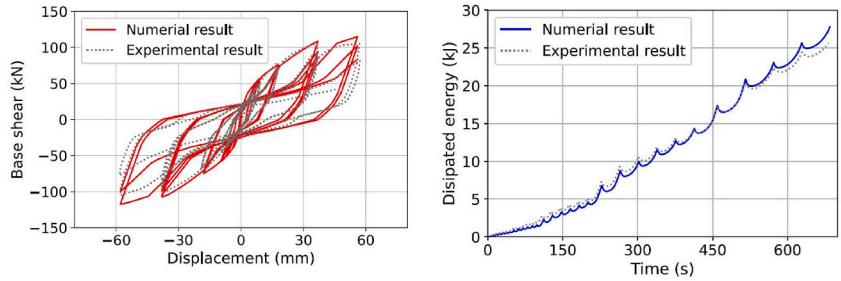
Table 2
Mechanical properties of CLT connections [5–7].

Connection	Loading direction	F_y (kN)	Δ_y (mm)	F_{max} (kN)	Δ_{max} (mm)	F_u (kN)	Δ_u (mm)
WHT 540 hold-down	Tension	40.46	8.810	48.33	20.30	38.79	23.75
WHT 540 hold-down	Shear	3.610	1.130	9.980	30.00	9.980	30.00
BMF 90x116x48x3 bracket	Tension	19.22	7.260	23.47	17.69	18.74	23.19
BMF 90x116x48x3 bracket	Shear	22.98	11.74	26.85	28.51	21.48	31.86
Half-lap screwed connection	Parallel	3.230	2.550	5.250	23.50	4.200	31.55
Half-lap screwed connection	perpendicular	2.910	2.340	4.910	17.51	3.930	23.76



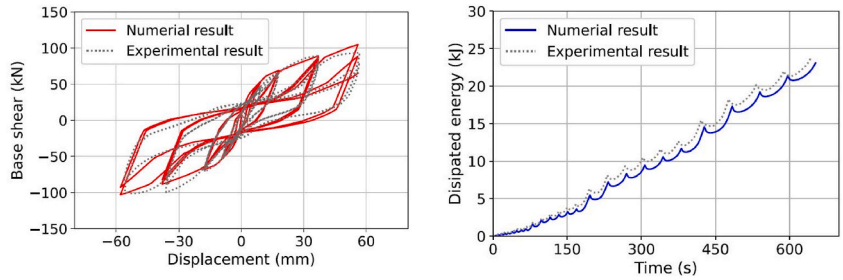
(a) Steel brackets **(b)** Hold-downs

Fig. 3. Validation example of the joint numerical models.



(a) Hysteretic response **(b)** Cumulative dissipated energy

Fig. 4. Validation of the numerical model for a single wall.



(a) Hysteretic response **(b)** Cumulative dissipated energy

Fig. 5. Validation of the numerical model for coupled walls.

More information regarding the ground motion set can be found in Ref. [29]. However, to increase the number of nonlinear responses, a factor of 2 was used to complement the original un-scaled group with records of increased intensity as discussed in Ref. [12]. Therefore, the total number of ground motion records used was 3312 leading to a total of 228,528 nonlinear response history analyses (NRHA) performed on the high-performance computing facility (HPC) at Imperial College London [32]. This extensive database of results forms the basis for the formulation of the ML models described below.

4. Data curation and feature selection

4.1. Data curation

As noted before, the *EDPs* of interest in this study are the maximum storey drift ratio (*MIDR*) and the maximum roof drift ratio (*RDR*) both of which were directly obtained from the NRHAs described above. Two main parameter types are considered for the development of ML models, namely ground-motion intensity measures (*IMs*) that describe the earthquake action, and structural parameters (*SPs*) that describe the buildings. We chose nine commonly used *IMs* with practical levels of hazard computability for our initial explorations. These *IMs* are: Peak Ground Acceleration (*PGA*), Peak Ground Velocity (*PGV*), Peak Ground Displacement (*PGD*), total Arias intensity at the end of the time series (I_a), significant duration (D_{5-75} and D_{5-95}), and spectral acceleration values at the building's fundamental period ($S_a(T_1)$) and at n times T_1 ($S_a(nT_1)$) where n was varied from 1.1 to 3.0. Moreover, due to the strong

influence of the frequency content on structural systems, including CLT buildings, identified by previous researchers [13,33], the mean period of the ground motion (T_m) and the ratio between the fundamental structural period and the ground-motion mean period (or the tuning ratio, T_1/T_m) were also considered.

The mean period (T_m) is an indicator of the frequency content of a ground motion which is determined as the weighted mean of the periods of the Fourier Amplitude Spectrum (FAS) over a pre-defined frequency range. The weights are assigned based on the Fourier amplitudes. The mean period (T_m) is calculated using the following relationship:

$$T_m = \frac{\sum_{i=1}^{n} C_i^2 \times \frac{1}{f_i}}{\sum_{i=1}^{n} C_i^2} \tag{1}$$

where C_i is the Fourier amplitude coefficient corresponding to a frequency (f_i) obtained from a discrete Fast Fourier Transform (FFT) for frequencies between 0.25 and 20 Hz, and Δf_i is the spacing of the frequencies for which the FFT is performed.

Table 3 summarises the intensity measures examined as candidate input features for our ML models. Besides, the Pearson correlation among *IM*s is provided in Fig. 6. This figure depicts a Pearson correlation heatmap with a colour gradation that goes from blue which represents negative colinearity (−1.0), to red which represents positive correlation (1.0). It is observed from Fig. 6 that most of the features are positively correlated with each other such as *PGA*, *PGV*, *PGD*, and spectral acceleration values, as expected. Negative correlations are observed in the case of significant durations and *PGA*, or T_1/T_m and spectral acceleration values.

In addition to the *IM*s, 6 structural parameters (*SP*s) are also considered as candidate input features. These include the behaviour factor (q), building height (H_w), slenderness (λ_w), length of fastening lines (F_l), joint density (β_w), and wall subdivision ratio (W_r) as shown in Table 4. These *SP*s have been reported in the literature to significantly impact the seismic response of CLT wall structures [28,34,35].

4.2. Data partitioning

A key concern around the development of ML models is the need to balance generalisation and overfitting. The latter term refers to the process of training a model so that it "memorises" the training data but is unable to make reasonably accurate estimates outside of it [36]. In order to prevent overfitting during the development of ML predictive models, we randomly divided the whole database into training and test sets following standard practice [37]. The training set contains 70% of the database and was used to train the model. The remaining 30% of the data was kept as unseen data and was used for performance testing purposes. Moreover, the common assumption of log-normality was also adopted for *IM*s [38,39] and hence we have transformed the *IM*s to the log-normal space together with the seismic drift demands of interest (*MIDR* and *RDR*).

4.3. Feature selection

It is usually not necessary nor practical to use all of the parameters identified in Tables 3 and 4 as predictors to estimate the drift demands even if a high accuracy was warranted. Instead, a sub-set of input features, which are most relevant to the drift demand prediction, is selected. Note that due to limited information on the seismic response of multi-storey CLT buildings being available to date, the key influencing features cannot be directly identified *a priori*. Thus, various feature selection techniques based on ML are applied to this end including Sequential Forward Floating Selection, LASSO, and Random Forest. Therefore, the efficiency of all the 30 input features (both *IM*s and *SP*s) for predicting drift demands in CLT structures is evaluated using the training dataset alone. This is done prior to the formulation of the ML predictive models and only to assess the predictive potential of each of the parameters being considered and to determine an optimal subset to be carried out for the formulation of our ML models. The mean square error (*MSE*) was used to evaluate the fit and a 10-fold cross validation was applied to generalise the results. Cross validation (CV) is a re-sampling method that not only improves the model's stability, but also mitigates overfitting [40]. A 10-fold is adopted following a preliminary analysis as the optimal number that provides good results within reasonable computational costs [41]. Therefore, the data is divided into 10 subsets of equal size. Then, for each training time, nine of the ten folds are used to train the model while the remaining subset is kept for measuring its performance. This process is repeated ten times. The performance of the model is obtained by averaging over the 10 folds [42,43].

4.3.1. Sequential Forward Floating Selection (SFFS)

Sequential Forward (or Backward) Selection accounts for the interactions between features during the process of feature selection by adding (or removing) one feature at a time until the best model is obtained based on a specified criterion. In SFFS, a number of backward steps are added at each forward step until the obtained result is better than that of the previous version of the model [44]. As a result, SFFS is able to leverage the trade-offs between forward and backward steps to find the optimal solution [45]. In this study, we carried out SFFS via Python's MLxtend Library [46].

Fig. 7a shows the results of SFFS in terms of the evolution of the 10-fold CV score *MSE* as a function of the number of features in the case of *MIDR* estimations, as an example. It can be seen from this figure that the improvement of prediction power is relatively insignificant after the 7th feature is added. The selected features are: ln(*PGV*), ln($S_a(T_1)$), ln($S_a(1.1T_1)$), ln($S_a(1.4T_1)$), ln(T_m), ln(T_1/T_m), and W_r, in that order. A similar process was applied for *RDR*, however, in this case, a set of 8 features is retrieved including: ln(*PGV*), ln($S_a(T_1)$), ln($S_a(1.4T_1)$), ln(T_m), ln(T_1/T_m), q, H_w, and W_r. The results from SFFS show remarkable consistency in the features selected for *MIDR* and *RDR* with *PGV*, T_m, T_1/T_m, W_r, and spectral acceleration values appearing in both fronts. *PGV* is widely recognised as a good *IM* for the structural seismic response [12,47] whereas spectral acceleration values, especially $S_a(T_1)$, is well

Table 3
IMs considered.

Intensity measure	Abbreviation	Min.	Max.	mean (μ)	SD (σ)
Peak ground acceleration (g)	PGA	0.003	3.559	0.145	0.177
Peak ground velocity (cm/s)	PGV	0.271	332.1	14.07	15.75
Peak ground displacement (cm)	PGD	0.073	525.5	6.373	8.748
Total arias intensity at end of time series	I_a	0.000	90.99	25.30	22.02
Significant duration between when 5% and 75% of the arias intensity occurred (s)	D_{5-75}	0.600	80.95	13.41	11.61
Significant duration between when 5% and 95% of the arias intensity occurred (s)	D_{5-95}	3.015	143.4	25.30	22.02
Mean period of the ground motion (s)	T_m	0.118	2.563	0.668	0.350
Ratio between the fundamental period of a building and the mean period or tuning ratio	T_1/T_m	0.126	23.53	2.493	2.075
Spectral acceleration at the fundamental period of the building (g)	$S_a(T_1)$	0.000	6.470	0.151	0.253
Spectral acceleration at n times of the building's fundamental period (g) ($n = 1.1, 1.2, 1.3. 1.4, 1.5, 1.6, 1.7,$ $1.8, 1.9, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0$)	$S_a(nT_1)$	0.000	6.309	0.089	0.168

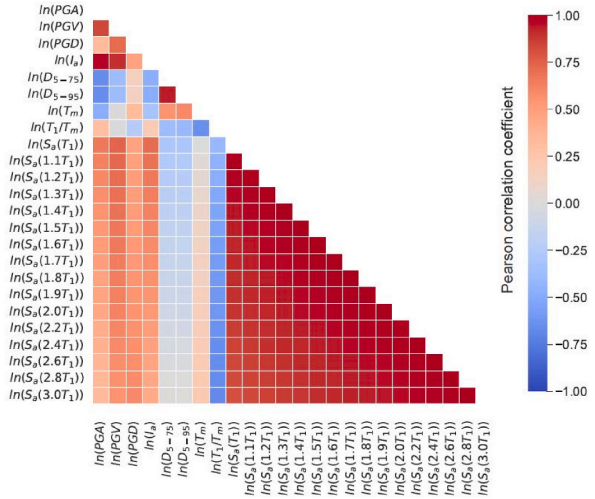


Fig. 6. Correlation heatmap among the considered *IMs*.

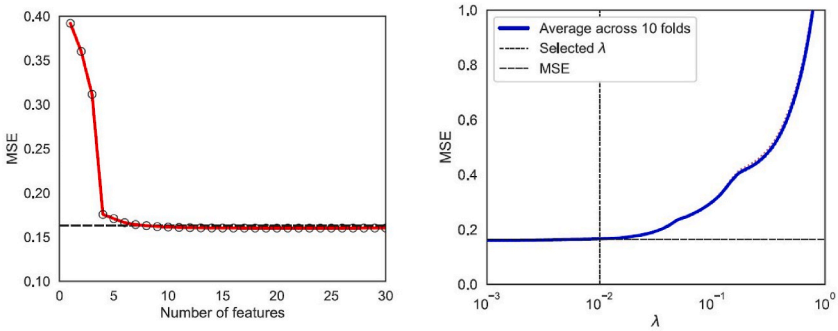
Table 4
Structural parameters considered.

Structural parameter	Abbreviation
Seismic behaviour factor	q
Height of buildings	H_w
The slenderness of the CLT wall	λ_w
Length of the fastening line	F_l
Ratio of fastening lines and perimeter	β_w
The number of sub-panels per wall	W_r

acknowledged to reduce the variability of the estimate in seismic responses [38]. The significant influence of T_m and T_1/T_m on the structural response has also been highlighted in the scientific literature [13,28]. In the case of structural parameters, the wall ratio (W_r) which represents the level of panel fragmentation is found to be the main factor to influence the seismic response in CLT buildings [13, 26,28].

4.3.2. LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) is a modified linear regression. The main distinct feature of LASSO, is that all input features are shrunk by adding a penalty to reduce their impact on the regression model. As a result, the regression coefficients of the features which have less importance on the target are penalised towards zero by increasing their shrinkage factors leading to a better generalisation of the regression model [42]. To this end, LASSO regression minimises the residual sum of squares as defined in Equation (2):



(a) SFFS **(b)** LASSO

Fig. 7. Evaluation of *MSE* for *MIDR*.

$$\widehat{\beta}_{lasso} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum\nolimits_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2}$$

where $\widehat{\beta}_{lasso}$ is the parameter estimated, β_0 the intercept, β_j a feature's coefficient for variable j, p the number of input features, x_{ij} a set features from the given data (n), λ the penalty factor, and the term $\sum_{j=1}^{p}|\beta_j|$ indicates the L1 norm of β_j. It should be noted that all predictor features are standardised in order to avoid dimensionality issues.

Fig. 7b shows the reduction in average *MSE* obtained from the 10-fold CV for different values of λ for *MIDR*, as an example. The minimum averaged *MSE* is associated with a λ close to zero, meaning that all features remain in the model. As the goal of our feature selection process is to shrink insignificant features, we selected a value of $\lambda = 0.01$. This marks the point of inflection in Fig. 7b and is consistent with previous applications of ML in the area of seismic engineering [27,48]. The features selected by LASSO, on the basis of $\lambda = 0.01$ are provided in Table 5. In general, LASSO's selection is broadly consistent with the findings of the previous section (SFFS), however, the number of features selected by LASSO is larger than those indicated by SFFS and most of the features are *IM*s. Moreover, the *MSE* of LASSO (= 0.18) is higher than that of SFFS (= 0.16). This can be attributed to the multicollinearity of some *IM*s contained in the LASSO model [42]. In particular, features such as spectral accelerations at different periods are highly colinear between each other and can be removed.

4.3.3. Random forest

In Random Forest, the entire dataset is partitioned into multiple subsets to develop individual Regression Trees. The results from these trees are averaged leading to a better precision of the estimations (i.e., variance reduction). Random Forest evaluates the importance of a feature based on an impurity reduction. For this, the algorithm ranks the input features based on how much they reduce the impurity during the tree training. A major advantage of using Random Forest feature importance measures is that the influence of each input feature is considered not only individually but also as part of multivariate interactions with other features [49]. Moreover, the algorithm is capable of informing relevant features in high-dimensional data with complex relationships [49,50]. In this study, the model's hyper-parameters were tuned using Grid Search technique with cross-validation to obtain a suitable set of hyper-parameters that minimises the average *MSE*. Once the combination of hyper-parameters is obtained, the Random Forest model is developed and the feature importance is evaluated. More details regarding Random Forest and hyper-parameter tuning are presented later as part of our discussion on model training.

The top 7 features with the highest importance as selected by Random Forest are summarised in Table 5. The selected features are: $\ln(PGV)$, $\ln(S_a(T_1))$, $\ln(T_m)$, $\ln(T_1/T_m)$, q, H_w and W_r, and $\ln(PGV)$, $\ln(S_a(T_1))$, $\ln(S_a(1.1T_1))$, $\ln(T_1/T_m)$, q, H_w and W_r, for *MIDR* and *RDR*, respectively. Among the features, *PGV* is the most important sharing around 60% of weight in the model prediction for both drift demand parameters considered herein.

4.3.4. Discussion on feature selection results

Table 5 compares the results of the different feature selection methods applied in this study for the estimation of *MIDR* and *RDR*. It should be noted that only the sub-set of selected features (among the original 30 features explored) are listed in this table. The table shows that different techniques lead to different sets and numbers of features. Instead of selecting the result coming from a single method, we have considered all feature sets in conjunction with the results from the correlation analysis (Fig. 6), as well as engineering judgement to arrive to the final set. We reasoned that in order to keep the model simple while ensuring high predictive performance, the number of the features included should be small i.e., < 7 as per the results of previous sections. The use of a small number of features also has a positive impact on the computational complexity of the models. Besides, in order to enable an easy adoption, and since most of the selected features for *MIDR* and *RDR* coincide, the same set of features was adopted for both drift demands.

First, $\ln(PGV)$ was selected since its role as a good predictor has been recognised by all feature-selection techniques applied for both *MIDR* and *RDR* in this study. Besides, *PGV* has been previously identified as an efficient *IM* for the estimation of drift demands in wooden structures [12]. Similarly, spectral acceleration-related features like $\ln(S_a(T_1))$, $\ln(S_a(1.1T_1))$, $\ln(S_a(1.2T_1))$, and $\ln(S_a(1.4T_1))$

Table 5
Feature selection using various ML techniques.

Feature	*MIDR*			*RDR*			Choose
	SFFS	LASSO	Random Forest	SFFS	LASSO	Random Forest	
$\ln(PGA)$		●			●		
$\ln(PGV)$	●	●	●	●	●	●	●
$\ln(T_m)$	●	●	●	●	●		
$\ln(S_a(T_1))$	●	●	●		●	●	●
$\ln(S_a(1.1T_1))$	●	●		●	●	●	
$\ln(S_a(1.2T_1))$		●			●		
$\ln(S_a(1.3T_1))$		●			●		
$\ln(S_a(1.4T_1))$	●	●		●	●		
$\ln(S_a(1.5T_1))$		●			●		
$\ln(I_a)$		●					
$\ln(T_1/T_m)$	●	●	●	●	●	●	●
q		●	●	●	●	●	●
H_w			●	●	●	●	●
W_r	●		●	●		●	●

have also been consistently identified as good predictors in Table 5. However, they are all multicollinear as can be seen from the heatmap of Fig. 6. This can threaten the predictability of regression models [51,52] and hence, only $\ln(S_a(T_1))$ was finally selected. $S_a(T_1)$ is the most commonly used *IM* [50] and has also been shown in past research to improve the estimates of seismic responses in a variety of structures [38]. With regard to frequency content-related features, both $\ln(T_1/T_m)$ or $\ln(T_m)$ appear in Table 5, however, $\ln(T_1/T_m)$ is structure-specific and has been highlighted in previous studies to improve the seismic response estimation of structures [13,53]. Therefore, $\ln(T_1/T_m)$ was considered as the third-IM feature of the final set.

In terms of structural parameters, q was the most frequently selected parameter by all algorithms. The other two consistently selected features are H_w and W_r, and can be associated with different aspects of the configuration of CLT structures. H_w is a proxy for the slenderness and flexibility of the building, while W_r represents the level of panel fragmentation which is related to the number of joint lines and correlates with the energy dissipation capacity of the buildings. In conclusion, the final set of features to be carried over for the development of ML models is comprised by: $\ln(PGV)$, $\ln(S_a(T_1))$, $\ln(T_1/T_m)$, q, H_w and W_r.

5. Development of ML models

5.1. Model description and training

Five ML algorithms are constructed to predict *MIDR* and *RDR* ratios in CLT structures in this paper. The algorithms used are: Multiple Linear Regression, Regression Tree, Random Forest, K-nearest Neighbors Regression, and Support Vector Machine Regression. All these ML algorithms are widely adopted in structural and earthquake engineering [15,16] and are therefore readily available in commercial and open-source platforms which is a desirable feature that should facilitate the model adoption in practice. In this study, we have implemented the ML algorithms using Scikit-learn package in the Python-based computational platform [54] and all models and hyper-parameters are available from https://github.com/CMalagaC/ML-CLT-Drifts.[1]

This section begins with an overview of the ML algorithms considered. Then, the process followed to optimise their hyper-parameters is discussed together with the performance measures adopted. A more indepth explanation of the ML algorithms and examples of their practical use can be found in Refs. [37,42,55].

5.1.1. Multiple Linear Regression

Multiple Linear Regression [37] is a regression method that assumes a linear relationship between the regression function and the p input features (x_1, x_2, …, x_p). Linear regression models have been effectively used in engineering practices to date due to their simplicity and easy interpretability. They can sometimes outperform more complex nonlinear ML models, especially in scenarios with sparse or limited training data [37]. Multiple linear regression can be described by:

$$\widehat{y}_i = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \tag{3}$$

where $\widehat{y}_i$ is the estimated target output, β_0 is the intercept, β_j is the feature's coefficient for x_j, p is the number of input features. The coefficients β_j are generally estimated using a training dataset (x_1,y_1), (x_2,y_1), …, (x_n,y_n) by means of a least squares approach. Given a number of features (p), each x_i can be written as a vector $(x_{i1},x_{i2}, \ldots,x_{ip})^T$. A vector of coefficients $(\beta_0,\beta_1, \ldots,\beta_p)^T$ can be found to minimise the sum of squared residuals such that:

[1] The models will be freely accessible from this repository address (https://github.com/CMalagaC/ML-CLT-Drifts) upon acceptance of the paper.

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{4}$$

5.1.2. Regression Tree

Regression Tree [56] is a decision tree algorithm applied to regression problems [42]. This method fits the data by constructing a set of splitting rules from top to bottom [56] as shown in Fig. 8a, creating what is known as a decision tree. This process starts with dividing the whole dataset into m non-overlapping regions (R_1, R_2, …, R_m) called "terminal nodes" or "leaves". The points at which the predictor space is subdivided are "internal nodes" while the tree segments that link the nodes are known as "branches" [37,42]. To construct the regions, the algorithm automatically selects the internal nodes to split input features and define the shape of the tree. The same prediction is made based on the output target y_i in each region R_m, represented by the mean of y_i. Subsequently, the regions (R_1, R_2, …, R_m) that minimise the RSS (function $f(x)$) are found:

$$f(x) = \sum_{m=1}^{M} \sum_{i \in R_m} \left(y_i - \hat{y}_{R_m} \right)^2 \tag{5}$$

where y_{R_m} is the mean of the target output observed in the training data in m. A greedy approach is adopted to split the input features j based on the lowest RSS. A splitting point j and s can be defined as the pair of half-planes:

$$R_1(j,s) = \left(X | X_j \le s \right), R_2(j,s) = \left(X | X_j > s \right) \tag{6}$$

and the value of j and s are determined to solve

$$\sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1} \right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2} \right)^2 \tag{7}$$

where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ is the mean target of the training data in $R_1(j, s)$ and $R_2(j, s)$, respectively. The process is repeated for all regions until a specific criterion is attained. However, Regression Tree models can be overfitted if the tree is very large while a small tree might result in underfitting [37,42].

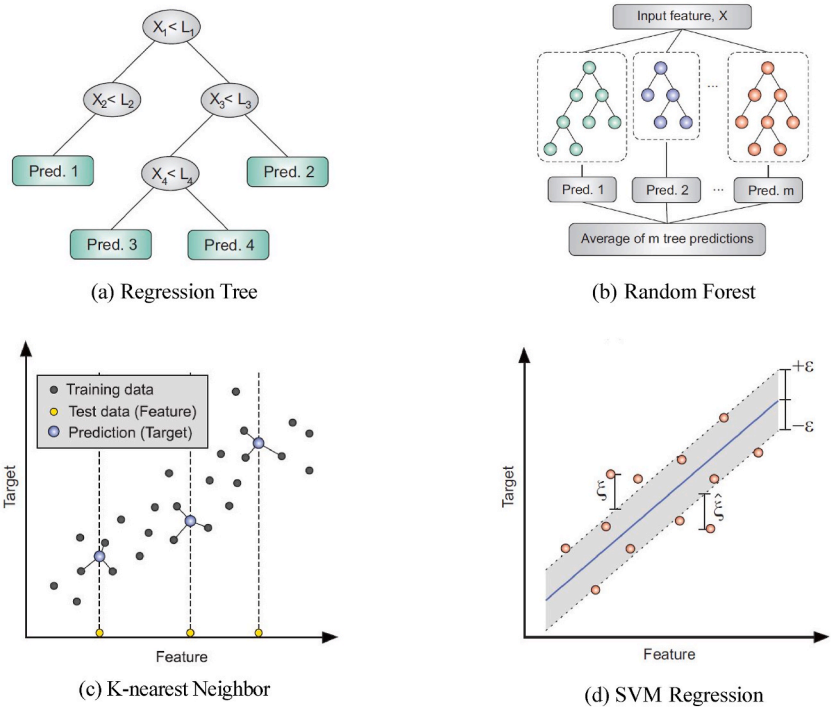


Fig. 8. Machine learning schemes.

$$\hat{f}_{RF} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{8}$$

5.1.3. Random forest

Random Forest [57] is an ensemble of Decision Trees. This method modifies the bagging approach by reducing the correlation between the trees [37]. The choice of selecting feature subsets is the key distinction between random forests and bagging in the tree-growing process. The process of considering each split, called "tree decorrelation", leads to a substantial reduction in variance over the trees, and to a more reliable model [42]. In other words, the decision tree algorithm is trained several times on a random variety of subsets in the training data. The final output value of the approach is determined by averaging the predictions obtained from a number of different decision trees (Fig. 8b), expressed by $\widehat{f}_{RF}\left(x\right) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$, where $\widehat{f}_{RF}(x\Big)$ is the final prediction averaged over the output of the trees T_b, and B is the total number of Decision Trees. Since there are numerous independent trees considering in the prediction, Random Forest is superior to Decision Tree in terms of model generalisation. However, it demands a high computational cost to construct a large number of trees and integrate their results.

5.1.4. K-nearest Neighbors Regression

K-nearest Neighbors Regression or KNN Regression [42] predicts the target by interpolating the target values selected from the nearest k points from the training set [42], as illustrated in Fig. 8c. The KNN algorithm is significantly faster than other ML algorithms especially when the dataset size and the number of input features are small [58]. It is also easy to add new data without affecting the accuracy of the algorithm. However, KNN is sensitive to noisy data and outliers. Three main KNN hyperparameters are required to be tuned, namely the number of neighbors (k), weight function, and the distance metric [58]. The number of neighbors k has to be defined before training. The distance between neighbors is specified by the distance function as

$$D_{mink} = \left(\sum_{i=1}^{k} |x_i - y_i|^p\right)^{\frac{1}{p}} \tag{9}$$

A value $p = 1$ leads to the Minkowski distances whereas a p of 2 refers to the Euclidean and Manhattan distances. Lastly, the weight function is neither an average nor an inverse of the distance-weighted average over the k neighbors.

5.1.5. Support Vector Regression

Support Vector Machine (SVM) [59] is commonly used in classification problems. However, SVM can also be applied to solve regression problems, called Support Vector Regression (SVR) which has a typical mapping function:

$$f(x) = \sum_{m=1}^{M} w_m \varphi_m(x) + b \tag{10}$$

where M is the number of mapping functions, w_m are weights of the mapping functions from the original to high-dimensional feature spaces, φ_m are the vectors of the input features (x), and b is the bias parameter.

The goal of the SVR algorithm is to find a hyperplane such that the majority of the training data points are within the margin or the tolerance level (ε). ε is a hyper-parameter that controls the maximum error margin between the target and the fitted hyperplane. Therefore, the algorithm tries to minimise their coefficients while constraining the error term depicted in Fig. 8d. The error function that is optimised by SVR can be expressed as:

$$C\sum_{i=1}^{n}(\xi_i + \widehat{\xi}_i) + \frac{1}{2}\|w\|^2 \; subject\; to \begin{cases} \left(\sum_{m=1}^{M} w_m \varphi_m\left(x_{ij}\right) + b\right) - y_i \le \varepsilon + \widehat{\xi}_i \\ y_i - \left(\sum_{m=1}^{M} w_m \varphi_m\left(x_{ij}\right) + b\right) \le \varepsilon + \widehat{\xi}_i \end{cases} \tag{11}$$

where ξ_i and $\widehat{\xi}_i$ are the slack variables for the data points i, n is the total number of the training data, w is a dimensional vector, y_i is a target output value, x_i is a vector of features for y_i, and C is a hyper-parameter representing regularization levels that shows the allowance for training data points outside of the error margin ε.

Several mapping functions can be applied to develop a highly accurate prediction model. A kernel function, which can map nonlinear relationships [59,60], is often used to improve the accuracy. In this study, the radial basis function kernel is employed. Moreover, SVR hyper-parameters such as C and ε are needed to control the bias-variance trade-off.

5.1.6. Hyper-parameter optimisation

Each ML algorithm was tuned by varying its hyper-parameters to minimise a cost function using a tuning process called GridSearchCV available from the Scikit-learn package [54]. GridSearchCV stands for Grid Search Cross-validation approach that involves generating grids with varying hyper-parameter values, evaluating the performance at each grid point, and iterating until the optimal values are found. Thus, the use of GridSearchCV to optimise hyper-parameters during the training process does not only improve the model robustness but also minimises the risk of overfitting via the CV. A 10-fold CV was employed. It should be noted that GridSearchCV is computationally expensive due to the large number of components to be evaluated. We performed the hyper-parameter tuning on the high-performance computing facility (HPC) [32]. Optimised values of hyper-parameters for each model are summarised in Table 6. Finally, overfitting was checked by comparing the training and test error of the ML models through the 10-fold CV score.

Table 6
Optimised values of hyper-parameters for regression models applied.

Model	Optimised hyper-parameters
Multiple Linear Regression	None
Regression Tree	maximum depth = None (both *MIDR* and *RDR*) minimum samples split = 6 (*MIDR*), 9 (*RDR*) minimum samples leaf = 12 (both *MIDR* and *RDR*) n estimators = 380 (*MIDR*), 480 (*RDR*)
Random Forest	maximum depth = None (both *MIDR* and *RDR*) minimum samples split = 2 (both *MIDR* and *RDR*) minimum samples leaf = 1 (both *MIDR* and *RDR*) n neighbors = 12 (*MIDR*), 13 (*RDR*)
K-nearest Neighbor	weights = uniform (both *MIDR* and *RDR*) p = 1 (*MIDR*), 2 (*RDR*) leaf size = 5 (both *MIDR* and *RDR*)
Support Vector Regression	ε = 0.1 (both *MIDR* and *RDR*) C = 1 (both *MIDR* and *RDR*)

The model which performed well on the test was considered effective and generalisable.

5.2. Performance measures

To evaluate the prediction performance of the newly developed ML models, we have used two performance metrics: i) the coefficient of determination (R^2), and ii) the root mean square error (*RMSE*). These two indicators, which are widely adopted in ML performance studies, are calculated on the basis of the 'unseen' test dataset mentioned previously. The expressions for R^2 and *RMSE* are given in Equations (12) and (13):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \tag{12}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{13}$$

where $\hat{y}_i$ is the estimated target output, y_i is the real target value, and n is the number of data points.

Table 7 reports the values of the above-mentioned metrics for the training and test datasets and for all ML algorithms used. The corresponding predictions versus actual drift ratios are depicted in Fig. 9. Overall, all models show a comparable prediction performance between the training and the test set, indicating that they do not suffer from overfitting and should perform well on unseen data. Except for multiple regression, all other ML algorithms show high predictive power evidenced by the high values of R^2 and low values of *RMSE* shown in Table 7. Even the worst performing model among the set, multiple regression, presents a value of $R^2 = 0.777$ on the test set for *MIDR* which is of itself not particularly low in comparison with other similar studies in earthquake engineering, e.g, Refs. [9, 61]. A notable mismatch between predictions and numerical results is clearly observed in Fig. 9a for Multiple Regression. This indicates that the traditional linear regression, which has been and continues to be broadly applied for estimating seismic demands in a variety of structures, may not capture the full range of drift responses observed in CLT structures. This can be partially attributed to its inability to describe highly nonlinear processes and relationships between the input features and the drift demands. A comparatively higher prediction performance was observed from all other ML models where values of R^2 in excess of 0.90 on both training and test sets are found in all algorithms and for both *MIDR* and *RDR* demand parameters.

This indicates that the ML algorithms applied herein can better capture the non-linearity in the seismic response of CLT walls. The highest R^2 and lowest *RMSE* were obtained for estimates coming from Random Forest for both inter-storey (*MIDR*) and roof drift (*RDR*) demands. Although Random Forest is the best-performing ML model overall, and hence our recommended choice, different users may find the other ML models more appealing depending on their background and familiarity with the underlying algorithms.[2]

It is worth noting that *MIDR* estimations had higher error rates than those of *RDRs* for any single predictive model. These relative differences obey, at least partially, to the influence of higher non-linear patterns caused by higher mode effects and rocking that will affect *MIDRs* more than *RDRs* making them relatively more difficult to predict. Structurally, higher mode effects tend to increase with increasing slenderness of structures whereas rocking is exacerbated in CLT structures with higher fragmentation levels.

6. ML model interpretation using SHAP

Even though the ML models proposed in the previous section have superior accuracy than previously proposed traditional prediction methods [13], as judged by their performance metrics. ML models are sometimes perceived as black boxes due to the complexity of their structure [58]. Therefore, it is essential to understand why and how an ML model makes predictions, not only to

[2] The models will be freely accessible from this repository address (https://github.com/CMalagaC/ML-CLT-Drifts) upon acceptance of the paper.

Table 7
Comparison of the predictive performance of ML algorithms.

ML model	*MIDR*				*RDR*			
	Training set		Test set		Training set		Test set	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Multiple Linear Regression	0.778	0.368	0.777	0.369	0.841	0.316	0.842	0.316
Regression Tree	0.926	0.182	0.919	0.242	0.929	0.173	0.924	0.228
Random Forest	0.950	0.102	0.946	0.198	0.953	0.094	0.949	0.185
K-nearest Neighbor	0.940	0.215	0.937	0.222	0.941	0.213	0.939	0.214
Support Vector Regression	0.927	0.216	0.923	0.217	0.946	0.186	0.947	0.185

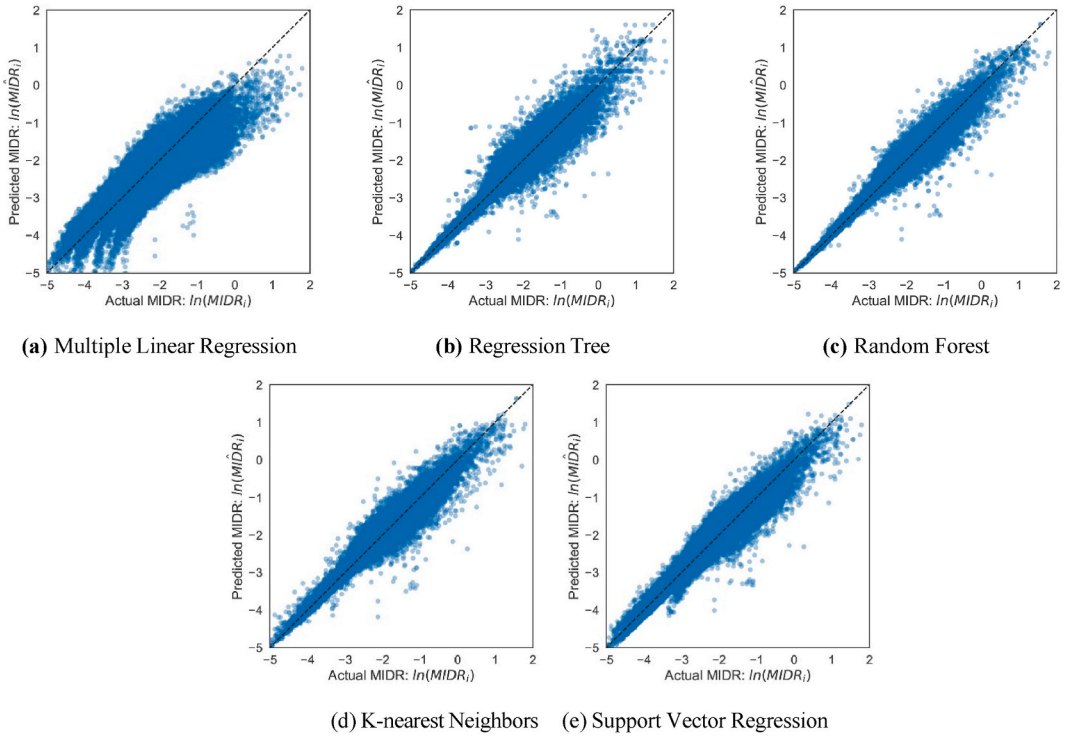


Fig. 9. Performance of ML models for predicting *MIDR*.

evaluate its accuracy [62]. In this section, we use the SHAP methodology [62] to provide an interpretation of our ML models. To this end, the Random Forest model is taken as the case study given its high accuracy in comparison with the alternatives put forward in the preceding section.

SHapley Additive exPlanations (SHAP) is a collection of explainers based on a game theory approach [63,64] that estimates Shapley values from an absolute average of the feature contributions over several simulations [62,65]. Since the original ML model is complex, this approach uses additive feature importance measures based on a linear explanation model to approximately interpret the original model. The function $g(z')$ is formulated using an explanation model that is a linear combination of binary variables [62] expressed as

$$f(x) \approx g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i' \tag{14}$$

where $f(x)$ is the original model and x is the original input feature. The explanation model uses x' as a simplified input feature and links it with a mapping function $x = h_x(x')$, while local methods attempt to guarantee that $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$. The value φ_i is the Shapley value which is expressed as:

$$\varphi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} (f_x(z') - f_x(z' \setminus i)) \tag{15}$$

when three desirable properties (*local accuracy, missingness, and consistency*) are satisfied [62]. $|z'|$ is the number of non-zero entries in

z', and all z' vectors are a subset of x'. More information regarding SHAP can be found in Refs. [62,63,65].

The SHAP summary plot for *MIDR* is presented in Fig. 10. Each point in this graph shows the relation of the input features with a SHAP value obtained from the same amount of data for the six input features under consideration (i.e., $\ln(S_a(T_1))$). These input features are arranged from top to bottom according to their importance with the most important feature according to its SHAP value being $\ln(S_a(T_1))$ and appearing at the top followed by H_w, $\ln(PGV)$, W_r, q, and $\ln(T_1/T_m)$, in this order of importance. Also, the variation of the input feature SHAP values is depicted by the variation in colour tones which changes from low (in blue) to high values (in red). Besides, data points which lead to the same SHAP value are scattered resulting in different distribution shapes for each feature.

Fig. 10 indicates that *MIDR* is directly proportional to almost all features except $\ln(T_1/T_m)$. That is, an increase in $\ln(S_a(T_1))$, H_w, $\ln(PGV)$, W_r, or q, leads to an increase in SHAP value which means that these features have a positive impact on *MIDR*. In the case of $\ln(T_1/T_m)$, it can be observed that lower $\ln(T_1/T_m)$ values have a stronger influence on *MIDR* than higher values. Moreover, small values of SHAP (SHAP values < 0 in our case) are sensitive to the variation all features. This is reflected in the long blue tails observed to the left in Fig. 10. This indicates that all features have a significant impact on the low range of *MIDR*. It should be recalled that in this range of responses *MIDR* is highly dependent on the elastic stiffness of the CLT structures, their yield strength and their structural configuration. On the other hand, $\ln(PGV)$ is the only significant feature to play a vital role in the prediction of large *MIDR* as understood from the large SHAP values associated with it shown in Fig. 10. Other features become progressively less influential on the drift prediction after the significant reduction in the structural stiffness after yielding.

A further examination of our ML model interpretability can be done with reference to Fig. 11. This figure shows scatter diagrams displaying local feature importance over a sub-set of input features where the interaction between the two features considered in each plot can be appreciated. In addition, the transition from one feature to another is also shown on the right of Fig. 11 represented by the transition from blue (low values) to red (high values). Overall, a markedly nonlinear influence of all the *IM*s examined over the maximum inter-storey drift (*MIDR*) is observed in Fig. 11 $\ln(S_a(T_1))$ and $\ln(PGV)$ have a positive relation with *MIDR*, although they follow different trends. This can be explained by the fact that higher values of $\ln(S_a(T_1))$ and $\ln(PGV)$ lead to higher SHAP values, which are associated with larger structural deformations. The impact of $\ln(S_a(T_1))$ on the estimation of *MIDR* decreases moderately while $\ln(PGV)$ tends to increase exponentially. In the case of $\ln(T_1/T_m)$, a positive impact on the drift estimation is observed until $\ln(T_1/T_m)$ reaches a value of 1. After that, $\ln(T_1/T_m)$ has no impact on the prediction denoted by the almost unchanged SHAP values. This marks a clear limitation to the predictive power of T_1/T_m that peaks at "resonance" and decays afterwards. This reduction in the predictive ability of T_1/T_m can be partially due to the raise in the importance of structural yielding. Fig. 11b, d, and 11f depict the SHAP dependency plots for the three structural parameters used for the formulation of our model. They show that H_w and q have a positive impact on the *MIDR* prediction. However, for q values of 3 or larger, *MIDR* becomes less sensitive to q. On the other hand, the trend of W_r changes from positive to slightly negative after a value of 3. This suggests that subdividing a CLT wall into more than 3 sub-panels is unnecessary.

7. Comparative studies

7.1. Comparison with previous studies

As highlighted in the introduction, there is a dearth of predictive models for seismic drifts in CLT structures. To our knowledge, only the work by Demirci et al. [13] covers similar grounds. However, Demirci et al.'s model does not incorporate *PGV* or $S_a(T_1)$ into its functional form and relies solely on the tuning parameter T_1/T_m for its predictions. This means that a strict one-to-one comparison between our model and previous proposals is limited. Nevertheless, some comparisons with the predictive relationships suggested by Demirci et al. for drifts can be done and are provided in this section.

Demirci et al.'s nonlinear regression models [13] were developed on the basis of numerical models for CLT walls of 6–20 stories with varying joint density, wall slenderness, and behaviour factor (q). They proposed predictive expressions for maximum inter-storey and global drift demands in the form of maximum (θ_{mod}) and global (δ_{mod}) drift modification factors, defined as:

$$\theta_{mod} = \frac{\theta_{max}}{q \times \theta_{1,max}} \tag{16}$$

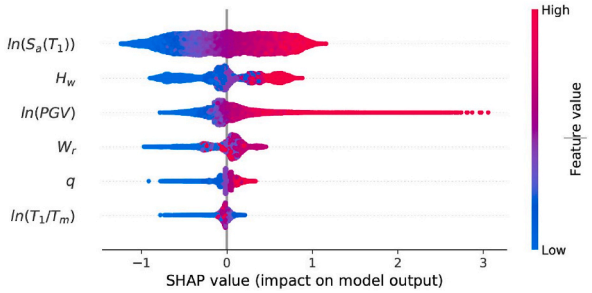


Fig. 10. Impact of the predictor features on *MIDR*.

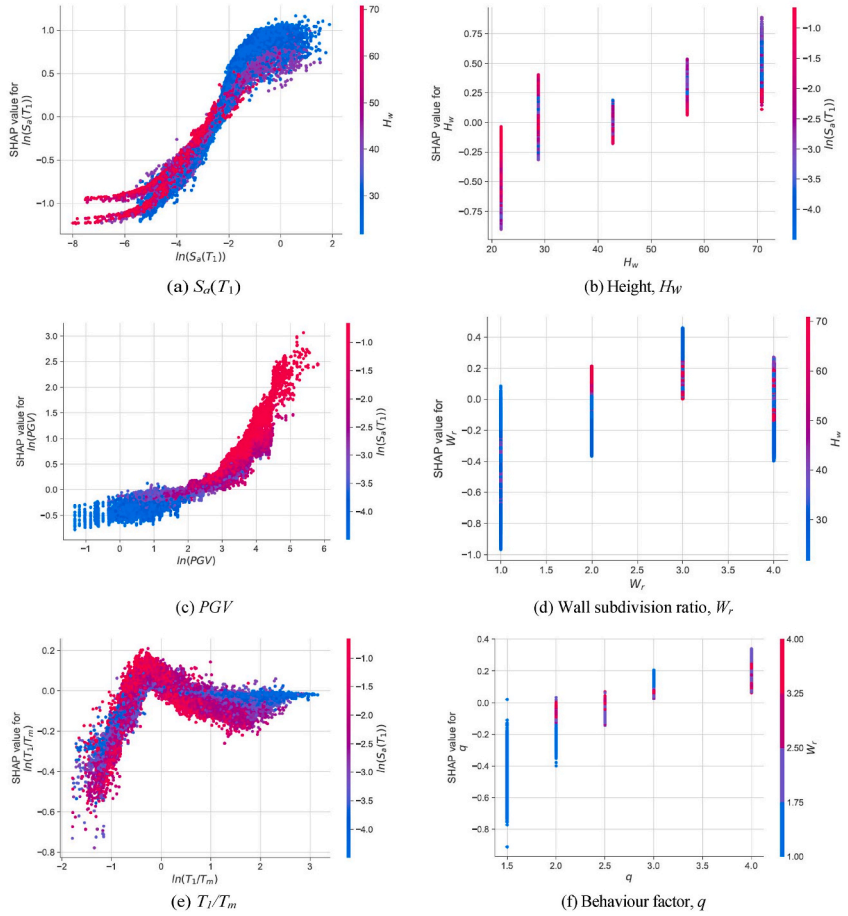


Fig. 11. SHAP dependency plots for *MIDR*.

$$\delta_{mod} = \frac{\Delta_{max}}{q \times \Delta_{1,roof}} \tag{17}$$

where θ_{max} is the maximum inter-storey drift ratio, $\theta_{1,max}$ is the maximum inter-storey drift ratio at first yielding obtained from the building's pushover curve, Δ_{max} is the maximum roof displacement, $\Delta_{1,roof}$ is the roof yield displacement at the first yielding also from a pushover analysis, and q is the behaviour factor. Worthy of note, the requirement of performing a nonlinear static (Pushover) analysis is avoided in our ML predictive models.

However, given the difference in *IM* s considered, the comparisons presented below are organised into different levels of $S_a(T_1)$, *PGV*, and presented as a function of T_1/T_m as illustrated in Fig. 12 to facilitate comparisons. To this end, global roof drift demands (*RDRs*) were estimated for different levels of earthquake intensities following the limiting values suggested in the Modified Mercalli Intensity scale (MMI). We acknowledge the empirical character of a MMI categorisation, however, for the specific case of the comparisons that follow, the use of this scale allows us to divide the seismic demands into *PGV* bins that are broadly indicative of their structural damage potential. Consequently, bins are created for 6 levels of *PGV* representing the Median *PGV* and the median plus and minus one standard deviation (*SD*) (Median *PGV* ± *SD*) for violent and extreme MMI events. Likewise, $S_a(T_1)$ bins are constructed around the median (Median $S_a(T_1)$), shown in orange in Fig. 12) and the median plus and minus one *SD* (Median $S_a(T_1)$ ± *SD*) presented in red and green, respectively. For example, in the 6-storey building, three bins of *PGV* values in the lower intensity are provided with means of 66, 78, and 92 cm/s. For each *PGV*, the global roof drifts for three $S_a(T_1)$ bins are shown consisting of the Median $S_a(T_1)$ = 1.2 g, the Median $S_a(T_1)$+*SD* = 2.1 g, and the Median $S_a(T_1)$-*SD* = 0.7 g.

The estimated *RDRs* obtained by means of our newly proposed ML model and the corresponding predictions from Demirci et al.'s relationships are shown in Fig. 12 for three different building heights (6, 12, and 20 storeys), representative of low-, mid- and high-rise CLT walled structures. The 6-storey wall with $W_r = 1$ and $q = 2$ is used as an example of a wall with no joints whereas the 12-storey case presented has $W_r = 3$ and $q = 3$ and the 20-storey has $W_r = 2$ and $q = 2$. These two latter structures use half-lap joints. Fig. 12 depicts the values of *RDR* as a function of the tuning parameter (T_1/T_m) for different levels of $S_a(T_1)$ and *PGV*, as discussed above.

In general, it can be observed from Fig. 12, that the influence of T_1/T_m on *RDR* identified by Demirci et al. [13] has been learnt by our ML model leading to larger *RDR* demands in the region of shorter T_1/T_m period ratios. This trend is also consistent with previous

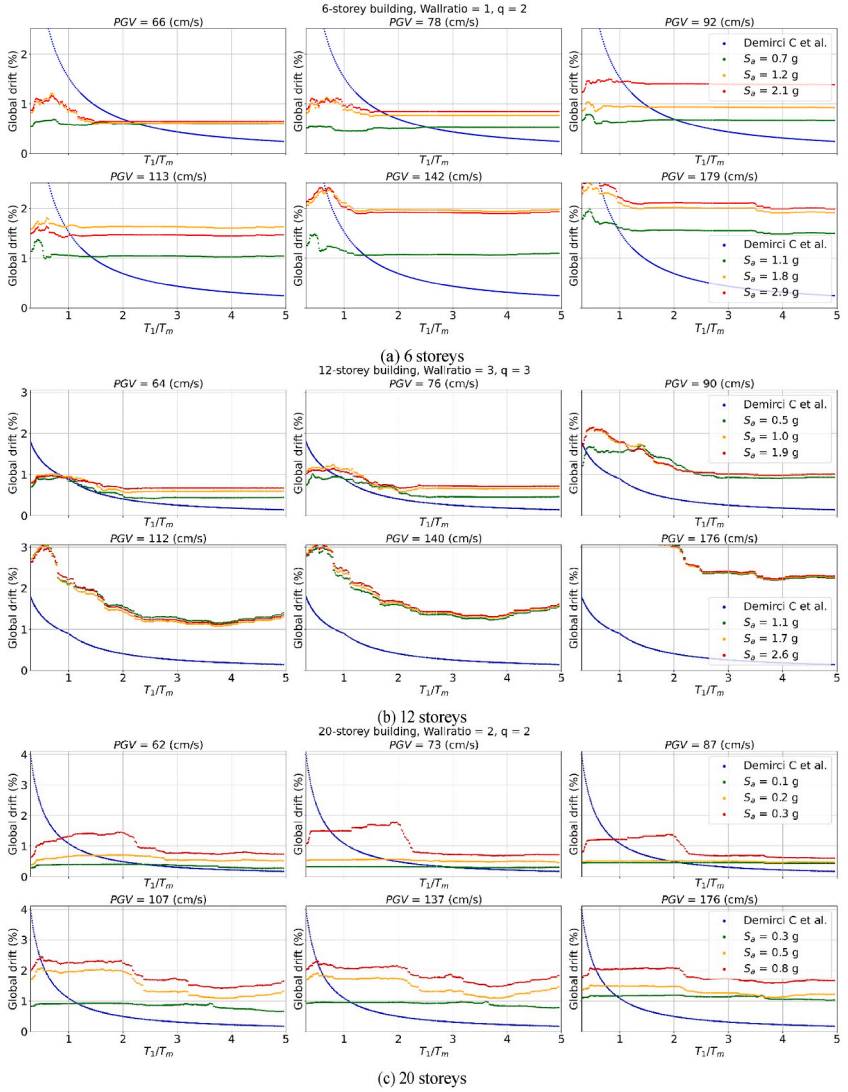


Fig. 12. Comparison of *RDR* obtained by the proposed model and previous studies.

studies for other structural types [8,66]. However, the sensitivity of *RDR* to T_1/T_m appears to be stronger in the model by Ref. [13] than in our newly proposed ML model, and this is specially evident at shorter periods. Interestingly, our ML model captures a more rapid change in the *RDR* at around $T_1/T_m = 1$ compared to Demirci et al.'s model which cannot detect this behavioural feature. This is probably due to the limitations of the functional form adopted by Demiric and their lack of consideration of the effects of $S_a(T_1)$ and *PGV* in the deformation demands of CLT buildings. This leads to an overestimation of *RDR* for short periods and an underestimation at large T_1/T_m periods if their expression is used.

In the case of the 12-storey wall (Fig. 12b), *RDR* predictions from our ML model and the regression model of Demirci et al. [13] are largely consistent for *PGV*s of 64–76 cm/s, approximately. The underestimations at longer periods from Demirci et al.'s model are still evident in the 12-storey structure. This is attributed to the lack of consideration of other ground-motion *IM*s discussed earlier. It is also interesting to note that Demirci et al.'s model predictions for the higher range of *PGV* ($>$ 112 cm/s) lead to noticeable low estimates of *RDR*. The fact that Demirci et al. relied solely on un-scaled records whereas scaling factors of 1 and 2 were used in the development of our model may account for these discrepancies at higher *IM*s.

In the case of the 20-story structure shown in Fig. 12c, the ML model proposed generally produces low *RDR* values that are more in accordance with Demirci et al.'s model predictions at longer periods for lower levels of *PGV* ($\approx$ 73 cm/s). Also interestingly, the noticeable drop in *RDR* happens at a period ratio of around $T_1/T_m = 2.0$, while in the case of 6 and 12 storeys such drop takes place at

approximately $T_1 = 1.0$. This can be attributed to the shift of the fundamental period caused by the activation of rocking modes in taller CLT walls subjected to stronger earthquakes.

7.2. Comparison with design provisions

An alternative comparison between different models can be offered in terms of drift modification factors. If the structural overstrength is ignored ($\Omega = 1$), the global drift modification factor can be expressed in terms of EC8, Part 1 or US provisions as:

$$\delta_{mod} = 1(\text{for EC8}) = \frac{C_d}{R}(\text{for US provision}) \tag{18}$$

where C_d is the deflection amplification factor and R is the response modification factor. In the case of US provisions [67], the C_d values proposed by Van de Lindt et al. for CLT structures [34] can be assumed. Therefore, δ_{mod} becomes 0.75 (for $C_d = 3$ with $R = 4$) and 1 (for $C_d = 3$ with $R = 3$, and $C_d = 4$ with $R = 4$). These δ_{mod} assumptions are compared with the ML predictions in Fig. 13. In these figures, two dashed horizontal lines are shown for δ_{mod} based on EC8 and US guidelines. As before, three CLT buildings, of 6-, 12-, and 20-storeys are selected for comparison. Likewise, the estimations are reported as a function of the period ratio (T_1/T_m) for different levels of $S_a(T_1)$ and *PGV* which are grouped using the bin limits described in the previous section. This leads to three different *PGV*s (Median *PGV* and Median *PGV* ± *SD*) per *PGV*-intensity group and three values of $S_a(T_1)$ overall, where the ML results corresponding to the Median $S_a(T_1)$ are presented in orange in Fig. 13 and the Median $S_a(T_1) \pm SD$ are shown in red and green, respectively.

In the 6-storey wall (Fig. 13a), our ML predictions are consistent with EC8 demand estimations in the short T_1/T_m range but become more in agreement with the US provisions at longer T_1/T_m ratios for lower *PGV* levels (i.e. *PGV*s between 66 and 78 cm/s). For earthquakes of higher intensity (i.e. *PGV* ≈ 92 cm/s) the ML predictions are closer to the EC8 demand levels. This indicates that the expected displacement of low or mid-rise CLT structures designed using EC8 or US provisions are broadly consistent with the assumed seismic demands. However, for taller structures like the 12-storey building, the δ_{mod} obtained from the EC8 and US codes are higher than their ML estimates at smaller *PGV*s (≈ 76 cm/s) for almost the whole range of T_1/T_m (Fig. 13b). Nonetheless, at higher *PGV*s (≈ 90 cm/s) our ML estimates are closer to the EC8 demand assumptions in the short T_1/T_m range and to the US assumptions at longer T_1/T_m ratios. Relatively good matches are also appreciable in the range 112 cm/s < *PGV* < 140 cm/s between the code and our ML model at $T_1/T_m > 2.0$, approximately. However, at very large *PGV* demands (≈ 176 cm/s), the code provisions seem to significantly underestimate δ_{mod} at all periods.

Finally, in the of 20-storey CLT wall case, both codes significantly overestimate the displacement of CLT walls at lower *PGV* levels (PGV ≈ 73 cm/s), except when the $S_a(T_1)$ is large (i.e., Median $S_a(T_1) + SD$) as seen in Fig. 13c. These discrepancies can be explained by the large difference between the simplified period estimations of the code and their numerically obtained counterparts. This is illustrated in Fig. 14 which compares the T_1 estimates of our numerical FE models, presented by the dark blue line, and the estimations made with the EC8 empirical formula presented in the form of bars. It should be noted that the numerical values of T_1 obtained from our Eigenvalue analyses are consistent with those reported in the scientific literature [68–71]. The variability of the numerical estimations for T_1 coming from buildings with the same height are due to the differences in wall partitioning and q-factors. As a result of the discrepancies shown in Fig. 14a, lower $S_a(T_1)$ values (around 30% lower than the EC8 calculations for the 12-storey buildings under consideration) are obtained. When higher *PGV* are expected, for example for *PGV*s over 107 cm/s in Fig. 13c, a better consistency between the ML model predictions and those of the codes are observed, especially at $T_1/T_m > 2$. Overall, the code provisions seem to lead to consistent deformation demand estimates for low earthquake intensities in short or mid-rise structures and large earthquake intensities in taller structures. In other words, the code-based estimations for tall CLT buildings are more conservative than those of short to medium CLT buildings. The above comparisons also highlight the influence of *PGV* and T_1/T_m on the drift demand in CLT structures and stress the need for their consideration in the design process. It should be noted that the comparison herein is based on the results from 2D-walled models and are therefore suitable for symmetric and regular structures. Additional validation results from the 3D models can be conducted in future work.

8. Conclusion

In this paper, we have examined the use of machine learning (ML) to predict peak seismic deformations in multi-storey CLT buildings. To this end, we have developed numerical models of 69 CLT walled structures with different structural configurations and design characteristics that represent a wide range of CLT building configurations. After calibrating them against available experimental results, our Finite Element models are used, in conjunction with a database of unscaled and mildly scaled 3312 ground-motion records, to conduct extensive nonlinear response history analyses and obtain a large dataset of seismic drift demands to be used as output targets.

A total of 30 features, including ground-motion intensity measures and structural parameters, have been considered as input features. Various ML feature selection techniques have informed our identification of the 6 most important predictors of seismic drift demands. Three of these features – Spectral Acceleration at T_1 ($S_a(T_1)$), Peak Ground Velocity (*PGV*), and the period ratio (T_1/T_m) – are related to the intensity of the ground motion while the other three – behaviour factor (q), structure height (H_w), and wall ratios (W_r) – are structural parameters.

Five models, constructed using widely employed ML algorithms, are proposed for the estimation of maximum inter-storey drift ratio (*MIDR*) and maximum roof drift ratio (*RDR*). These models are based on Multiple Regression, Regression Tree, Random Forest, K-nearest Neighbor, and Support Vector Regression. Our results show that these ML algorithms have superior prediction power compared to traditional regression approaches. This is reflected in a good agreement between model estimates and numerical results

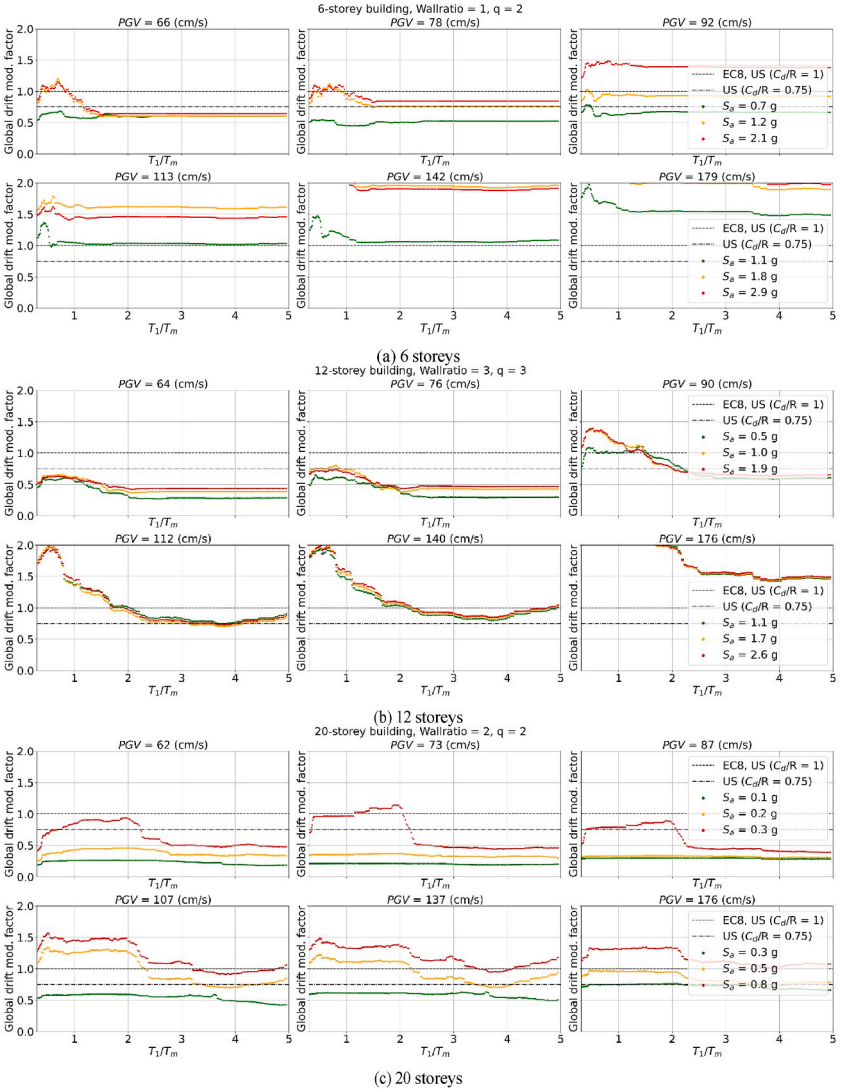
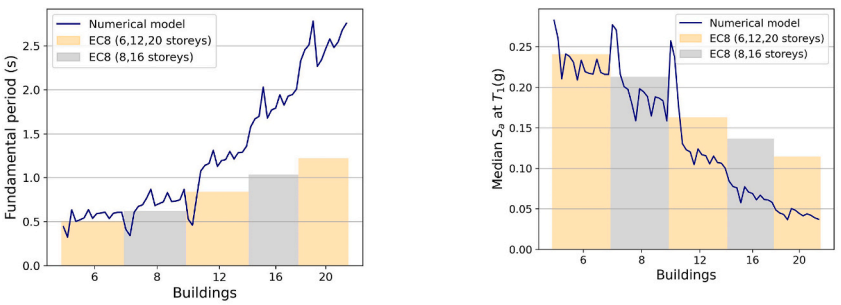


Fig. 13. Comparison of δ_{mod} obtained by the proposed model and codes.



(a) Comparison of T_1 **(b)** Comparison of $S_a(T_1)$

Fig. 14. Comparison of the fundamental period (T_1), and spectral acceleration at T_1 ($S_a(T_1)$) between modal analysis and EC8.

and expressed in high R^2 and low *RMSE* for both *MIDR* and *RDR*. The Random Forest model moderately outperforms the rest in terms of performance metrics (with R^2 of 0.946 and 0.949 for *MIDR* and *RDR*, respectively) while Multiple Regression shows the most limited predictive power (with an $R^2 \approx 0.7$). All models can be accessed from the following repository: https://github.com/CMalagaC/ML-CLT-Drifts.

Our examinations have also revealed that q, H_w, and W_r have a greater influence on *MIDR* than on *RDR* and particularly at lower levels of demand. This is partially explained by the relative dependency of *MIDR* on the building strength and stiffness distributions as well as its higher susceptibility to higher mode effects. This observation was further reinforced by the feature importance and the dependent plots developed using the SHAP technique. Likewise, $S_a(T_1)$ and T_1/T_m actively define the drift prediction at small drifts where the structure remains (or is close to) elastic. However, *PGV* dominates the model outputs for large displacements. Besides, SHAP also served to identify that a $q = 3$ and a $W_r = 3$ are optimal for multi-storey CLT buildings and an increase in their value beyond those levels (by, for example, increasing the level of fragmentation of a CLT wall beyond 3 sup-panels) will bring practically no benefit to the overall ductility of CLT buildings.

After comparisons with EC8 and US provisions, code-based predictions appear consistent with our ML estimations for low earthquake intensities in short or mid-rise structures and large earthquake intensities in taller structures. Besides, for taller CLT buildings and larger *PGV*s our ML estimates are closer to the EC8 demand assumptions in the short T_1/T_m range and to the US assumptions at longer periods. The discrepancies between the simplified calculations of natural periods offered by codes serve to explain some of these differences.

Our findings have stressed the limitations of current frameworks for drift estimation, and have highlighted the use of more advanced algorithms to improve their predictive accuracy. The comparative advantages of the ML models proposed herein lay in their ability to predict drift demands of CLT walls for a wide range of earthquake intensities and provide insight into the nonlinear behaviour of multi-storey CLT structures outside the constraints of a simplified functional form.

Author statement

Eknara Junda: Investigation, Methodology, Software, Analyses, Writing. Christian Málaga-Chuquitaype: Conceptualization, Methodology, Investigation, Supervision, Writing, Reviewing and Editing. Ketsarin Chawgien: Methodology, Software, Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be uploaded to https://github.com/CMalagaC/ML-CLT-Drifts upon acceptance.

References

[1] K. Sullivan, T.H. Miller, R. Gupta, Behavior of cross-laminated timber diaphragm connections with self-tapping screws, Eng. Struct. 168 (2018) 505–524.
[2] M. Schenk, C. Hübner, J.M. Cabrero, Cross-Laminated timber: a survey on design methods and concepts in practice, Civil Eng. 3 (2022) 610–629.
[3] M. Fragiacomo, M. Follesa, D. Casagrande, M. Piazza, *The Revision of the Timber Chapter of the Eurocode 8* in *Proc*, 5th Int. Conf. on Structural Health Assessement of Timber Structures, 2019.
[4] C. Málaga-Chuquitaype, *Chapter 8: Design of Timber Structures* in *Seismic Design of Buildings to Eurocode 8*, CRC Press, 2016, pp. 213–234.
[5] I. Gavric, M. Fragiacomo, A. Ceccotti, Cyclic behaviour of typical metal connectors for cross-laminated (CLT) structures, Mater. Struct. 48 (2015) 1841–1857.
[6] I. Gavric, M. Fragiacomo, A. Ceccotti, Cyclic behavior of typical screwed connections for cross-laminated (CLT) structures, Eur. J. Wood and Wood Prod. 73 (2015) 179–191.
[7] I. Gavric, M. Fragiacomo, A. Ceccotti, Cyclic behavior of CLT wall systems: experimental tests and analytical prediction models, J. Struct. Eng. 141 (2015), 04015034.
[8] C. Málaga-Chuquitaype, Estimation of peak displacements in steel structures through dimensional analysis and the efficiency of alternative ground-motion time and length scales, Eng. Struct. 101 (2015) 264–278.
[9] X. Guan, H. Burton, M. Shokrabadi, Z. Yi, Seismic drift demand estimation for steel moment frame buildings: from mechanics-based to data-driven models, J. Struct. Eng. 147 (2021), 04021058.
[10] L.A. Laughery, S. Pujol, Drift demands on reinforced concrete structures subjected to strong ground motions, ACI Struct. J. 116 (2019) 103–115.
[11] R. Napolitano, A. Bilotta, E. Cosenza, Seismic lateral deformations demand in conceptual design of reinforced concrete framed structures, J. Build. Eng. 45 (2022), 103565.
[12] P. Heresi, E. Miranda, Intensity measures for regional seismic risk assessment of low-rise wood-frame residential construction, J. Struct. Eng. 147 (2021), 04020287.
[13] C. Demirci, C. Málaga-Chuquitaype, L. Macorini, Seismic drift demands in multi-storey cross-laminated timber buildings, Earthq. Eng. Struct. Dynam. 47 (2018) 1014–1031.
[14] C. Málaga-Chuquitaype, Machine learning in structural design: an opinionated review, Front. Built Environ. 8 (2022) 2297–3362.
[15] H. Sun, H.V. Burton, H. Huang, Machine learning applications for building structural design and performance assessment: state-of-the-art review, J. Build. Eng. 33 (2021), 101816 issn: 2352- 7102.
[16] Y. Xie, M. Ebad Sichani, J.E. Padgett, R. DesRoches, The promise of implementing machine learning in earthquake engineering: a state-of-the-art review, Earthq. Spectra 36 (2020) 1769–1801.
[17] F. Zahra, J. Macedo, C. Málaga-Chuquitaype, Hybrid data-driven hazard-consistent drift models for SMRF, Earthq. Eng. Struct. Dynam. 52 (2023), 1112-1135.
[18] S.-H. Hwang, S. Mangalathu, J. Shin, J.-S. Jeon, Machine learning-based approaches for seismic demand and collapse of ductile reinforced concrete building frames, J. Build. Eng. 34 (2021), 101905 issn: 2352-7102.
[19] A. Paral, D.K.S. Roy, A.K. Samanta, A deep learning-based approach for condition assessment of semi-rigid joint of steel frame, J. Build. Eng. 34 (2021), 101946.
[20] H.D. Nguyen, N.D. Dao, M. Shin, Machine learning-based prediction for maximum displacement of seismic isolation systems, J. Build. Eng. 51 (2022), 104251.
[21] Z. Li, et al., Prediction of damage accumulation effect of wood structural members under long-term service: a machine learning approach, Materials 12 (2019) 1243.
[22] Z. Xin, D. Ke, H. Zhang, Y. Yu, F. Liu, Non-destructive evaluating the density and mechanical properties of ancient timber members based on machine learning approach, Construct. Build. Mater. 341 (2022), 127855.

[23] P. Code, Eurocode 8: Design of Structures for Earthquake Resistance-Part 1: General Rules, Seismic Actions and Rules for Buildings, European Committee for Standardization, Brussels, 2005.
[24] C. Málaga-Chuquitaype, J. Skinner, A. Dowdall, J. Kernohan, *Response of CLT Shear Walls under Cyclic Loads* in *World Conference on Timber Engineering*, 2016. Vienna, Austria.
[25] F. OpenSees McKenna, A framework for earthquake engineering simulation, Comput. Sci. Eng. 13 (2011) 58–66.
[26] I. Gavric, M. Fragiacomo, A. Ceccotti, Cyclic behavior of CLT wall systems: experimental tests and analytical prediction models, J. Struct. Eng. 141 (2015), 04015034.
[27] E. Junda, C. Málaga-Chuquitaype, J.M. Escudero, *Influence of Panel Fragmentation on the Seismic Response of Cross-Laminated Timber Buildings* in *World Conference on Timber Engineering*, Chile, Santiago, 2021.
[28] C. Demirci, *Seismic Response of Multi-Storey Cross-Laminated Timber Buildings* PhD Thesis (Civil and Environmental Engineering, Imperial College London, SW7 2AZ, UK, 2019.
[29] J. Hancock, J.J. Bommer, P.J. Stafford, Numbers of scaled and matched accelerograms required for inelastic dynamic analyses, Earthq. Eng. Struct. Dynam. 37 (2008) 1585–1607.
[30] J.W. Baker, Measuring bias in structural response caused by ground motion scaling in 8th Pacific Conference on Earthquake Engineering, Singapore 8 (2007).
[31] C.A. Goulet, et al., PEER NGA-east database, Earthq. Spectra 37 (2021) 1331–1353.
[32] HPC-Imperial, Imperial College Research Computing Service, 2019.
[33] C. Demirci, C. Málaga-Chuquitaype, L. Macorini, Seismic shear and acceleration demands in multi-storey cross-laminated timber buildings, Eng. Struct. 198 (2019), 109467.
[34] J.W. Van de Lindt, et al., Seismic performance factors for cross-laminated timber shear wall systems in the United States, J. Struct. Eng. 146 (2020), 04020172.
[35] A. Aloisio, R. Alaggio, M. Fragiacomo, Fragility functions and behavior factors estimation of multi-story cross-laminated timber structures characterized by an energy-dependent hysteretic model, Earthq. Spectra 37 (2021) 134–159.
[36] C. Sammut, G.I. Webb, Encyclopedia of Machine Learning, Springer Science & Business Media, 2011.
[37] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.
[38] N. Shome, C.A. Cornell, P. Bazzurro, J.E. Carballo, Earthquakes, records, and nonlinear responses, Earthq. Spectra 14 (1998) 469–500.
[39] N. Jayaram, J.W. Baker, Statistical tests of the joint distribution of spectral acceleration values, Bull. Seismol. Soc. Am. 98 (2008) 2231–2243.
[40] H. Sun, A Data-Driven Building Seismic Response Prediction Framework: from Simulation and Recordings to Statistical Learning, University of California, Los Angeles, 2019.
[41] R. Kohavi, et al., *A study of cross-validation and bootstrap for accuracy estimation and model selection* in, Ijcai 14 (1995) 1137–1145.
[42] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R* 15–57, Springer US, New York, NY, 2021, 978-1-0716-1418-1.
[43] K. Chawgien, S. Kiattisin, Machine learning techniques for classifying the sweetness of watermelon using acoustic signal and image processing, Comput. Electron. Agric. 181 (2021), 105938.
[44] P. Pudil, J. Novovicová, J. Kittler, Floating search methods in feature selection, Pattern Recogn. Lett. 15 (1994) 1119–1125.
[45] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance. eng, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 153–158, issn: 0162-8828.
[46] S. MLxtend Raschka, Providing machine learning and data science utilities and extensions to Python's scientific computing stack, J. Open Source Softw. 3 (2018) 638.
[47] K. Morfidis, K. Kostinakis, Seismic parameters' combinations for the optimum prediction of the damage state of R/C buildings using neural networks, Adv. Eng. Software 106 (2017) 1–16.
[48] S. Mangalathu, J.-S. Jeon, R. DesRoches, Critical uncertainty parameters influencing seismic performance of bridges using Lasso regression, Earthq. Eng. Struct. Dynam. 47 (2018) 784–801.
[49] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC Bioinf. 9 (2008) 1–11.
[50] C. Liu, J. Macedo, Machine learning-based models for estimating seismically-induced slope displacements in subduction earthquake zones, Soil Dynam. Earthq. Eng. 160 (2022), 107323.
[51] T. Næs, T. Isaksson, T. Fearn, T. Davies, A User-Friendly Guide to Multivariate Calibration and Classification, NIR Chichester, 2002.
[52] T.-T. Liu, D.-G. Lu, X.-H. Yu, Development of a compound intensity measure using partial least-squares regression and its statistical evaluation based on probabilistic seismic demand analysis, Soil Dynam. Earthq. Eng. 125 (2019), 105725.
[53] C. Málaga-Chuquitaype, M. Psaltakis, G. Kampas, Wu, J. Dimensionless fragility analysis of seismic acceleration demands through low-order building models, Bull. Earthq. Eng. 17 (2019) 3815–3845.
[54] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[55] G. Bonaccorso, Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning, Packt Publishing Ltd, 2018.
[56] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[57] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[58] V.V. Degtyarev, K.D. Tsavdaridis, Buckling and ultimate load prediction models for perforated steel beams using machine learning algorithms, J. Build. Eng. 51 (2022), 104316.
[59] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.
[60] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[61] L. Di Sarno, J.-R. Wu, Fragility assessment of existing low-rise steel moment-resisting frames with masonry infills under mainshock-aftershock earthquake sequences, Bull. Earthq. Eng. 19 (2021) 2483–2504.
[62] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[63] B. Peleg, P. Sudhölter, Introduction to the Theory of Cooperative Games, Springer Science & Business Media, 2007.
[64] S. Mangalathu, H. Shin, E. Choi, J.-S. Jeon, Explainable machine learning models for punching shear strength estimation of flat slabs without transverse reinforcement, J. Build. Eng. 39 (2021), 102300.
[65] S. Masís, Interpretable Machine Learning with Python: Learn to Build Interpretable High-Performance Models with Hands-On Real-World Examples, Packt Publishing Ltd, 2021.
[66] C. Málaga-Chuquitaype, A.Y. Elghazouli, Inelastic displacement demands in steel structures and their relationship with earthquake frequency content parameters, Earthq. Eng. Struct. Dynam. 41 (2012) 831–852.
[67] A.T. Council, Quantification of Building Seismic Performance Factors, US Department of Homeland Security, FEMA, 2009.
[68] D. Vassallo, M. Follesa, M. Fragiacomo, Seismic design of a six-storey CLT building in Italy, Eng. Struct. 175 (2018) 322–338.
[69] X. Sun, M. He, Z. Li, Z. Shu, Performance evaluation of multi-storey cross-laminated timber structures under different earthquake hazard levels, J. Wood Sci. 64 (2018) 23–39.
[70] M. Shahnewaz, Y. Pan, M. Shahria Alam, T. Tannert, Seismic fragility estimates for cross-laminated timber platform building, J. Struct. Eng. 146 (2020), 04020256.
[71] S. Yang, H. Hong, F. Bartlett, Responses and capacity curves of mid-and high-rise wood buildings subjected to seismic excitations, Can. J. Civ. Eng. 47 (2020) 63–76.