BIG DATA GENERATION AND COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS IN PREDICTING THE FUNDAMENTAL PERIOD OF STEEL STRUCTURES CONSIDERING SOIL-STRUCTURE INTERACTION

Ashley Megan van der Westhuizen[1], Nikolaos Bakas[2] and George Markou[3, 1]

[1] Department of Civil Engineering, University of Pretoria, South Africa
e-mail: u17179221@tuks.co.za; george.markou@up.ac.za

[2] National Infrastructures for Research and Technology – GRNET, 7 Kifisias Avenue, 11523, Athens, Greece
e-mail: nibas@grnet.gr

[3] Cyprus University of Technology, Department of Civil Engineering and Geomatics
e-mail: george.markou@cut.ac.cy

Abstract

The computing of the fundamental period of structures during seismic design is well documented in design codes but is mainly dependent on the height of the structure, which is considered to be the most influential parameter. It is, however, important to consider a phenomenon called the soil-structure interaction (SSI), as this has been found to have a detrimental effect, especially for buildings founded on soft soils. A pilot research project foresaw the use of machine learning (ML) algorithms trained on relatively limited data sets for the development of a more accurate and objective fundamental period formula. Therefore, a data set that consists of 98,308 fundamental period data points was created through the use of a High-Performance Computer (HPC), which is the largest data set of its kind. The HPC results were then used to train, test, and validate different ML algorithms. It was found that XGBoost-HYT-CV with hyperparameter tuning performed the best with a correlation of 99.99% and a Mean Average Percentage Error (MAPE) of 0.5%. Furthermore, the XGBoost-HYT-CV model outperformed all under-study ML models when using an additional data set that consisted of out-of-sample building geometries and soil properties, with a resulting MAPE of 9%. Finally, irregular buildings were also used to test the performance of the proposed predictive models.

1. Introduction

The fundamental period [1] is an important parameter used by engineers to understand the dynamic behaviour of structures during earthquake excitation. The fundamental period is used to determine the seismic loads during design. In the current design code formulae and literature, the height of the structure is the only parameter considered, but it has been found that other factors affect the natural period, and the inclusion of these parameters is crucial in designing safe and economic structures. The importance of the SSI phenomenon in calculating the fundamental period of steel structures forms the main objective of this research work.

When dealing with design code formulae, a fixed base condition is assumed, but in reality, structures are generally found within the soil. This results in a more complicated structural response that is controlled by the mechanical characteristics of the soil and the type of foundation [2-3]. It has been found that the SSI phenomenon has a detrimental influence on the fundamental period of structures, especially when they are founded on soft soils.

As it was discussed in [1], current formulae found in design codes for computing the fundamental period of steel structures are not accurate enough to predict the fundamental period of real structures. In this work, it was

suggested to use ML algorithms to obtain a new formula that accounts for different geometrical features of the superstructure, where the SSI effect is also considered. According to this pilot research project [1] which also served as preparatory work for the needs of this research work, experimentally validated finite element (FE) models are used to develop a data set that is then used to train ML algorithms for the development of predictive models.

In this research work the same approach proposed in [1] is used to develop a large data set through the use of HPC that foresees the solution of 49,154 3D finite element models of steel structures. The models take into account the SSI phenomenon, while the numerically obtained data set is then used to develop predictive models through the use of ML algorithms [19]. Markou et al. [19], proposed four new ML algorithms that exhibit advanced accuracy and extendibility characteristics when used for any structural mechanics-related problem. Two of the proposed ML algorithms proposed in [19] are adopted herein for the needs of the development of the proposed predictive models. The numerically obtained results and the response of the proposed predictive models will be presented after discussing the current knowhow on computing the fundamental period of steel framed structures.

2. Available fundamental period equations and predictive models

Computing the fundamental period of steel structures was a subject of numerous studies where international design codes offer different recommendations that derive from semiempirical knowledge. The formulae currently suggested by the design codes are presented below starting with that of Eurocode 8, SANS10160-4 and ASCE 7.

Eurocode 8 and SANS10160-4 **[4-5]**:

$$T_1 = C_t(H)^{0.75}. \quad (1)$$

where:

$C_t = 0.085$ for moment-resistant space steel frames.

$C_t = 0.075$ for eccentrically braced steel frames.

ASCE 7-05 **[6]:**

$T_1 = 0.0724(H)^{0.8}$ for steel moment-resisting frames. (2)

$T_1 = 0.0731(H)^{0.75}$ for eccentrically braced steel frames. (3)

ASCE 7-10 **[6]:**

$$T_1 = C_{t1}(H)^{x}. \quad (4)$$

$$T_1 = 0.01\,N. \quad (5)$$

The parameters in the formula presented above are altered so that they are similar to those used in this research work, for comparison purposes. H represents the height in meters of the building and N the number of storeys. The values for C_{t1} and x can be found in Table 1.

Table 1: Values for C_{t1} and x

Structure type	C_{t1}	x
Steel moment-resisting frame	0.028	0.8
Eccentrically braced steel frame	0.03	0.75
Concentrically braced steel frame	0.02	0.75

Two additional formulae were proposed in [7] that consider the plan view geometry of the building $(L \times B)$ as shown in "Eq. (6)", which is applicable for low-rise buildings (up to 10 m), where "Eq. (7)" is recommended to be used for medium-rise buildings (up to 30 m).

$$T_1 = 0.056 \times (L \cdot B)^{0.3289}. \tag{6}$$

$$T_1 = C_0 (L \cdot B)^{0.3289 \cdot \alpha}. \tag{7}$$

with:

$$C_0 = 0.0247 e^{0.1305 \cdot H}. \tag{8}$$

$$\alpha = 0.4473 e^{-0.0441 \cdot H}. \tag{9}$$

As it was presented in [8], the fundamental period of irregular moment-resisting steel frame structures was investigated, and it was found that structures without irregularities tended to have longer periods than irregular structures. The formula presented in "Eq. (10)" was found to statistically match Rayleigh's method, however this method requires to be solved through the use of a computer program.

$$T_1 = 0.071 (H)^{0.75} \left(\frac{H_{\text{av}}}{H}\right)^{0.35} \left(\frac{D_{\text{av}}}{D}\right)^{0.2}. \tag{10}$$

where:

H_{av} = average height in meters

D, D_{av} = depth and average depth in the direction of the earthquake forces in meters.

It is important to note here that Rayleigh's equation is computed through the use of "Eq. (11)" (units in meters and seconds).

$$T_1 = 2\pi \sqrt{\sum_{i=1}^{N} \frac{w_i \delta^2{}_i}{g} \sum_{i=1}^{N} f_i \delta_i}. \tag{11}$$

where:

w_i = total weight of structure assigned to level i

f_i = lateral force at level i

δ_i = deflection at level i relative to the base, due to lateral forces

g = acceleration due to gravity

As seen in the formulae described above, there is no parameter that considers the SSI effect, where the main parameter used to compute the fundamental period practically is the height of the buildings. It has been found that the SSI can increase the fundamental period, thus, it is an important consideration when determining natural frequencies of building-like structures [1, 3, 9-11]

One of the research works that delt with accounting for the SSI phenomenon during the fundamental period computation of steel structures, was presented in [9]. According to [9], the influence that SSI has on the fundamental period of buildings was investigated and it was found that the influence depends mainly on the soil-structure relative rigidity (K_{ss}). The relative rigidity is expressed in terms of soil shear wave velocity (V_s), foundation area (A), the flexural rigidity of building columns (I_c, E_c), storey height (H), number of storeys (N_s) and spans (N_{bt}, N_{bl}), whereas K_{ss} can be calculated based on the relationship provided in "Eq. (12)".

$$K_{ss} = \frac{N_{bt} \times N_{bl} \times \rho \times V_s^2 \times H^3 \times \sqrt{\frac{A}{A_0}}}{N_s \times E_c \times (I_c)^{3/4}}. \tag{12}$$

The influence of K_{ss} on the fundamental period of steel structures is shown in Figure 1. It can be seen that for flexible buildings (log (K_{ss})>1.5) the SSI could be neglected, but, in other cases, if the SSI is ignored, the fundamental period will be significantly misestimated consequently leading to a poor prediction [9].

A number of recent investigations have been conducted where the influence of SSI on the fundamental period of buildings was assessed. All studies involved the use of ML algorithms in the determination of an improved fundamental period formula and showcased lower errors than the current design code formulae [1, 3]. The first published investigations that paved the way for developing more accurate formulae will be presented in this research work, which mainly involved determining the influence of SSI on the fundamental period of reinforced concrete (RC) structures [3].

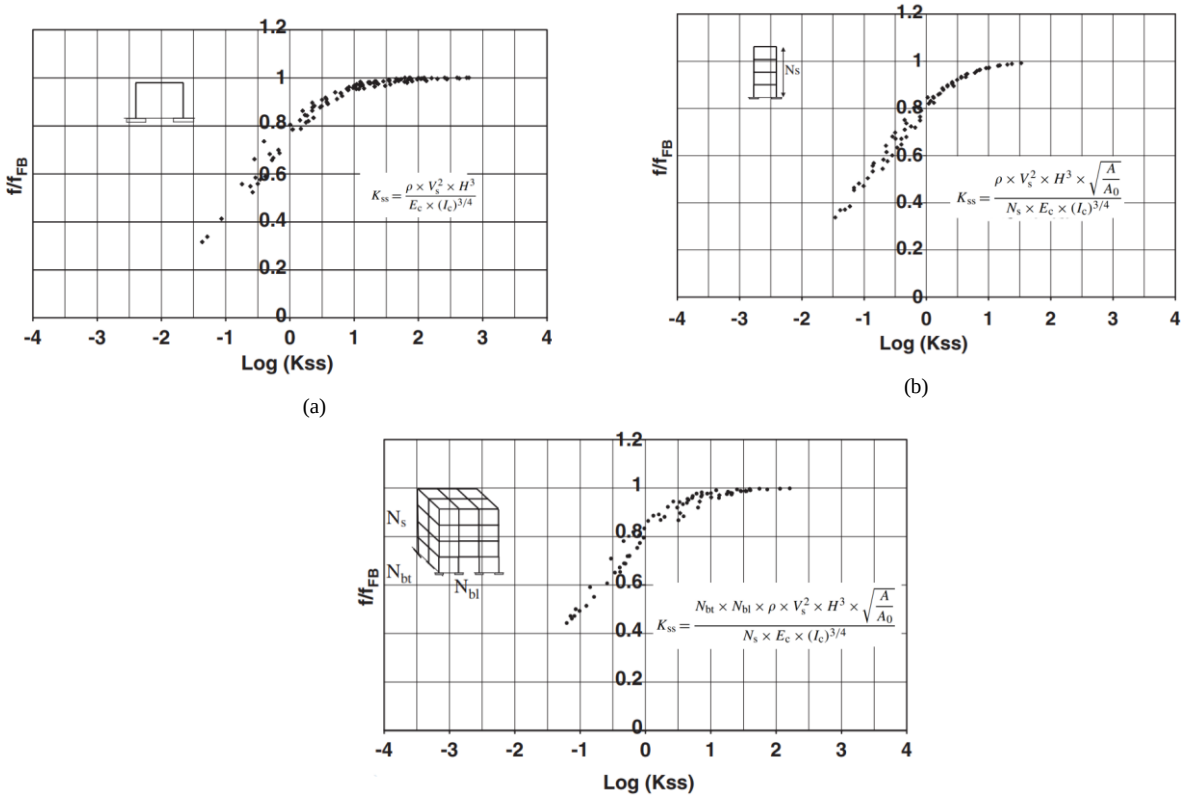


Figure 1: Influence of K_{ss} on the fundamental frequency of buildings: (a) one-storey buildings (b) multi-storey buildings with one span in the lateral and transverse directions and (c) multi-storey buildings with multiple spans [9].

It is important to state here that the preparatory study in [1], foresaw the investigation of the influence of SSI on the fundamental period of steel structures, where a higher-order nonlinear regression (NLR) model was used to obtain a fundamental period formula. The data set used to train and test the ML algorithms consisted of 1,152 numerical data points generated using finite element analysis (FEA) software [12]. As it was reported by [1] the proposed 40-feature formula, which was found to be the model that best fitted the data set, could predict the fundamental period with a correlation of 99.95%. The formula can be seen in "Eq. (13)" [1] and it will be used to evaluate the newly proposed predictive models and formulae that will be presented in this research work.

Furthermore, according to the parametric investigation presented in [1] it was also found that a correlation of 99.71% and a MAPE of 2.8% was achieved when the formula was tested on the out-of-sample steel structures' data set. Although the results obtained are accurate, it was deemed necessary that a larger data set should be developed to yield more accurate results that can predict the fundamental period for a wider range of structures, given that the previously developed data set foresaw the use of a limited number of structural geometries.

According to the current state-of-the-art, this research work's objective foresaw the creation of a large data set that consists of nearly 100,000 results that were obtained through developing numerical models that were analysed by utilizing an HPC. Thereafter, this data set (which is currently considered to be the largest of its kind in the literature) is used to parametrically investigate the numerical response of different ML algorithms, including that of Artificial Neural Networks (ANN), in an attempt to develop an objective, accurate and robust fundamental period predicting model for steel structures that account for the SSI effect. The accuracy of the proposed predictive models is compared to that of design code formulae and proposed formulae found in the literature.

It is important to note here that the ML algorithms that were used for the needs of this research work are integrated within the research software *nbml*[1] which is an open-source code, where the developed dataset presented in this research work can also be found within the folder named "datasets" under the name "Dataset 98308 Steel Frames with SSI.xlsx". Furthermore, the performance of the newly proposed POLYREG-HYT and XGBoost-HYT-CV ML algorithms that are used herein for the needs of the numerical analysis performed in this research work, is found in [19], where their numerical response is tested for different data sets.

$$
\begin{aligned}
T = {} & (0.194630 \cdot lH^2) + (0.0580556 \cdot CO^2 \cdot B) - (9.39027 \cdot InvCO \cdot InvB \cdot lB) - (8.49213 \cdot InvL \cdot CO \cdot H) \\
& - (41.8498 \cdot InvCO \cdot lL \cdot H) - (8.14564 \cdot InvE \cdot E \cdot H) - (0.800465 \cdot CO \cdot B \cdot H) \\
& + (114.808 \cdot InvCO \cdot InvB \cdot H) + (46.6778 \cdot InvCO \cdot InvB^2) + (0.0631499 \cdot B^2 \cdot H) \\
& + (4.20803 \cdot lB \cdot CO \cdot H) - (0.144945 \cdot lL \cdot H \cdot L) + (0.847694 \cdot B \cdot H \cdot InvL) + (9.37930 \cdot InvL^2 \cdot H) \\
& - (1.08930 \cdot InvCO^2 \cdot L) + (4.04342 \cdot InvL) - (0.251627 \cdot InvL \cdot CO \cdot B) \\
& - (0.00783561 \cdot InvB \cdot lCO \cdot lE) + (0.523388 \cdot lL^2 \cdot InvCO) + (0.0947335 \cdot InvH \cdot lH \cdot L) \\
& + (46.8309 \cdot InvE \cdot H \cdot lDs) + (0.00764850 \cdot lH * B) + (0.000161108 \cdot lL \cdot L \cdot lE) \\
& - (20.5554 \cdot InvE \cdot CO \cdot Ds) - (0.00474725 \cdot InvL^2 \cdot InvDs) + (2.73101 \cdot InvL \cdot InvH \cdot CO) \\
& + (0.403996 \cdot InvCO \cdot lB \cdot L) - (0.0105914 \cdot lL \cdot L \cdot B) - (0.228100 \cdot lB^2 \cdot CO) \\
& + (0.00265642 \cdot InvL \cdot H^2) - (2.58386 \cdot InvB \cdot InvH \cdot CO) + (5.84142 \cdot InvCO \cdot H \cdot L) \\
& + (29.5168 \cdot InvCO \cdot H) + (0.849560 \cdot InvL \cdot lH \cdot CO) - (2.14776 \cdot InvB \cdot lH * lCO) \\
& + (1.34222 \cdot lB \cdot lH \cdot InvH) - (0.00333495 \cdot lE \cdot L \cdot InvH) - (2.64111 \cdot InvB^2 \cdot InvH) \\
& + (71.1358 \cdot InvH \cdot Ds \cdot InvE) - (17.9194 \cdot InvE \cdot lE \cdot lL) - 1.16636
\end{aligned}
\tag{13}
$$

where:

T is the fundamental period (s)

D_s is the depth of soil (m)

E is the soils Young's Modulus (kPa)

H is the building height (m)

L is the length of the building parallel to the oscillating direction (m)

B is the width of the building perpendicular to the oscillating direction (m)

CO is the orientation of the columns (either a 1 or 2)

$lParameter$ is $\ln(Parameter + 1)$ i.e., $lD_s = \ln(D_s + 1)$

$InvParameter$ is $\frac{1}{Parameter+1}$ i.e., $InvD_s = \frac{1}{D_s+1}$

3. Development of numerical models and the data set through HPC

In this study, the initial data set published in [1] is significantly extended by changing the maximum number of bays and stories to 25. It must be noted here that the columns assume a HEA (European H) section, where the beams are discretized with IPE (European I) sections for the entire framing system. The superstructure

[1] https://github.com/nbakas/nbml

was discretized through the use of Natural Beam-Column Force-Based Element (NBCFB) finite elements, where the raft slab was discretized through 8-noded isoparametric hexahedral elements.

The NBCFB element is a 2-noded 3D beam-column FE which is shown in Figure 2. The element has *12* degrees of freedom (dof), *6* per node, and assuming that *xyz* represents the global coordinate system, they are grouped in the vector:

$$\rho = [u_1 \quad v_1 \quad w_1 \quad \theta_1 \quad \varphi_1 \quad \psi_1 \quad u_2 \quad v_2 \quad w_2 \quad \theta_2 \quad \varphi_2 \quad \psi_2] \tag{14}$$

were *u*, *v* and *w* representing the translational dof, whereas θ, φ, and ψ denote the rotational dof. These dof can refer either to a global or to a local Cartesian coordinate system that are related through transformation matrices that contain directional cosines. A local Cartesian coordinate system *x'y'z'* is assigned to the element with the corresponding Cartesian dof:

$$\bar{\boldsymbol{\rho}} = \left[\bar{u}_1 \quad \bar{v}_1 \quad \bar{w}_1 \quad \bar{\theta}_1 \quad \bar{\varphi}_1 \quad \bar{\psi}_1 \quad \bar{u}_2 \quad \bar{v}_2 \quad \bar{w}_2 \quad \bar{\theta}_2 \quad \bar{\varphi}_2 \quad \bar{\psi}_2\right] \tag{15}$$

A natural coordinate α is adopted spanning the beam's axis which coincides with the local Cartesian axis *x'*. The local Cartesian dof are transformed into natural invariant rigid body and straining modes ρ_0 and ρ_N, respectively, so that a unique and reversible relation exists between the natural modes and the local and global dof:

$$\underset{(12x1)}{\boldsymbol{\rho}} \Leftrightarrow \underset{(12x1)}{\bar{\boldsymbol{\rho}}} \Leftrightarrow \underset{(6x1)}{\boldsymbol{\rho}_0}, \underset{(6x1)}{\boldsymbol{\rho}_N} \qquad \underset{(6x1)}{\boldsymbol{\rho}_N} = \underset{(12x1)}{\bar{\boldsymbol{\rho}}} - \underset{(6x1)}{\boldsymbol{\rho}_0} \tag{16}$$

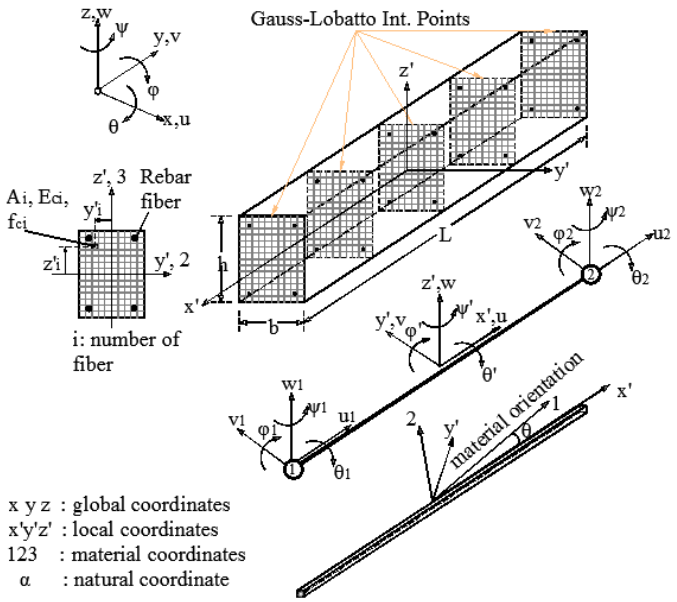


Figure 2: RC Fiber NBCFB in space [30]

In case of a fiber consideration along the cross section of the beam, an additional coordinate system is defined for every fiber (*k*), namely the *123* coordinate system with axis *1* along the principal reinforcement direction and axis *2* perpendicular to it. Note that material axis *3* is parallel to the local Cartesian axis *z'*. Then, for every fiber *k*, axis *1* forms an angle θ_k with the local axis *x'* (see Figure 2). Therefore, the NBCFB element comprises *12* Cartesian dof, and uses 4 different coordinate systems, where the actual number of straining modes that cause stress development is *6*. Furthermore, this flexibility-based FE uses an internal iterative procedure for the computation of the internal forces during the nonlinear analysis. For the complete formulation of this advanced beam-column FE see [30].

The kinematic connection of the NBCFB and 8-noded isoparametric hexahedral elements can be seen in Figure 3. The initial model developed consisted of a single bay with a length of 5 m and a width of 3 m. The building also had a single storey with a height of 3.5 m, and a raft slab foundation was used for supporting the base of the building. The initial model can be seen in Figure 4, while Figure 5 depicts the cases of frames with 2 and 3 bays. The mass of the slabs was modelled considering a 25 kg/m^3 bracing element as seen in Figure 3, which had a thickness of 220 mm to account for the dead and 30% of live loads that were applied on the slabs. It is also noteworthy to state here that each structural member was discretized through the use of 5 NBCFB elements for achieving maximum numerical accuracy [30].

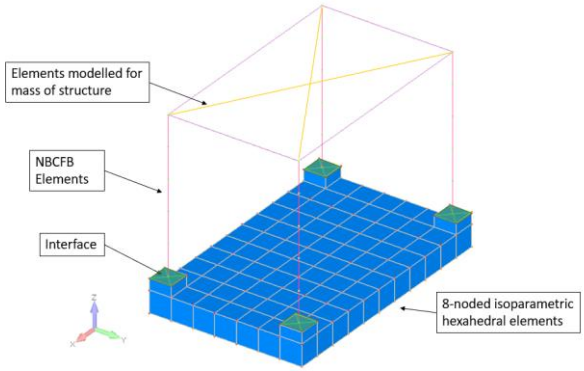


Figure 3: Initial model represented with elements used in models

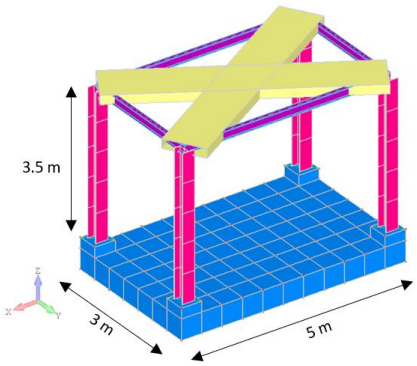
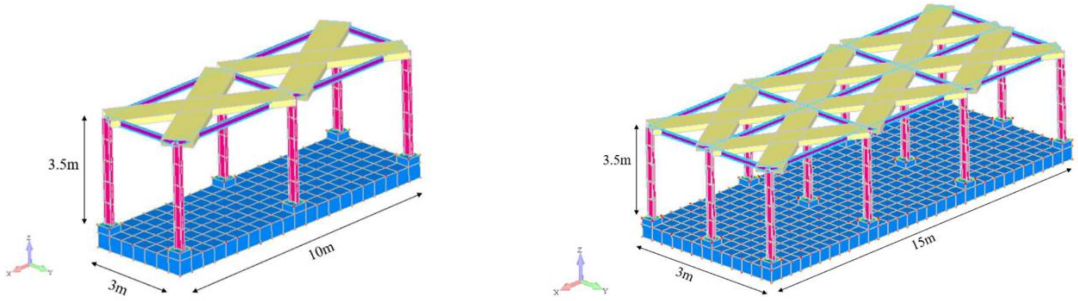


Figure 4: Initial model [1]



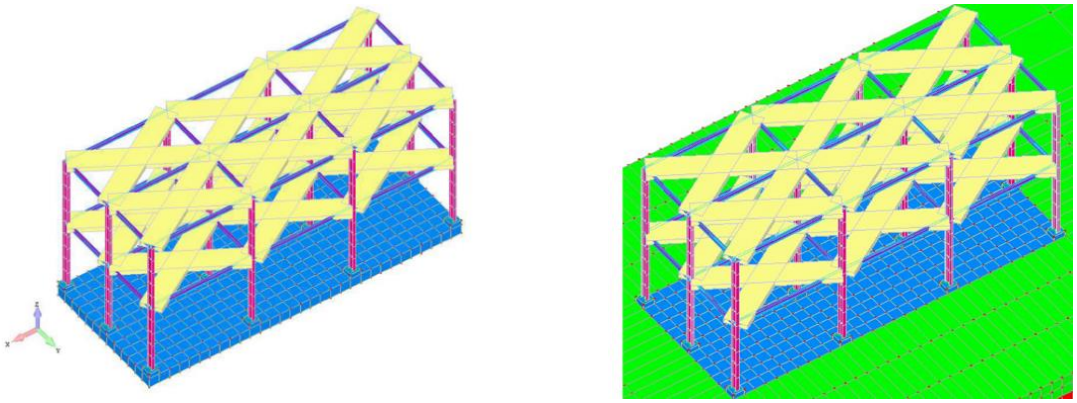
(a) (b)

Figure 5: Initial model modified to have (a) two bays and (b) three bays [1]

The initial model was expanded by altering various parameters such as the number of storeys, plan area, depth of soil, and orientation of the HEA-columns. It is important to note that every change in the number of bays in the x-direction results in one additional bay in the y-direction. According to the developed set of steel frames, the smallest plan view had a dimension of 5x3 m (1 bay along the x, and 1 along the y global axis), and the largest foresaw a 125x39 m (25 x 13 bays). All buildings, up to 25 bays, were expanded to include up to 25-storeys. The lowest building height that was used for the development of the data set was 3.5 m (1-storey) and the tallest was 87.5 m (25-storeys). With the completion of the fixed-based models, the data set was further expanded to account for the SSI effect by discretizing different soil domains as seen in Figure 6 A summary of the minimum and maximum parameter values used in developing the data set can be seen in Table 2. It has to be noted here that the dataset was developed by using 9 different Young moduli (65, 130, 200, 275, 300, 450 500, 600, and 700 MPa) to account for different soil conditions.

One of the parameters related to the geometry of the steel frames that was investigated herein was the orientation of the strong axis of the columns' section. All the models initially foresaw that the column sections' strong axis was parallel to the long direction of the building, where a 0 was used when summarising the results as

the input value within the data set. Thereafter, the columns were rotated by 90° and an input value of 1 was used instead, thus separating the two cases within the data set. An example of orientations 0 and 1 can be seen in Figure 7. According to the column section orientation in relation to the global axes of the building, the order of the building length, L, and width, B within the data set may differ depending on which mode is considered, while the different cases can be seen in Table 3.



(a) (b)

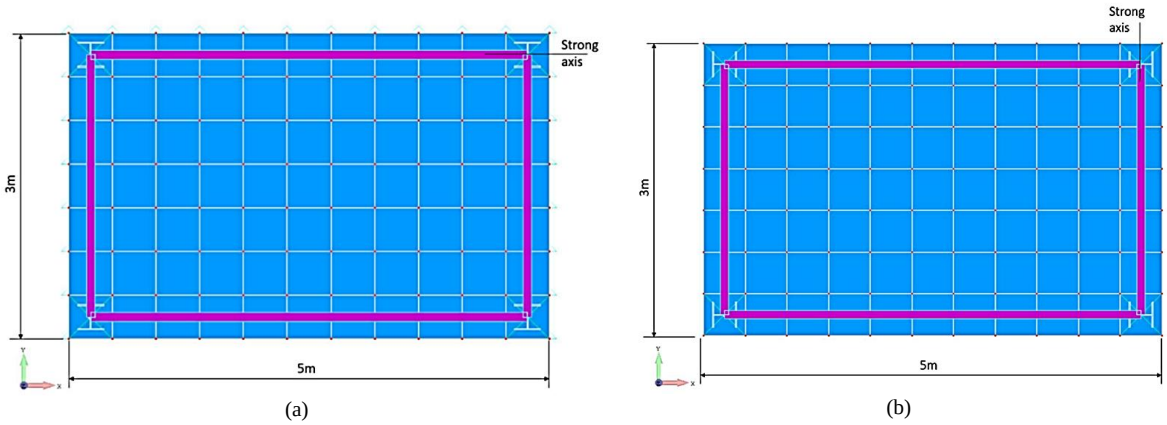
Figure 6: Triple span in the long direction, a double span in the short direction (a) fixes base with raft foundation (b) flexible base with hexahedral mesh.

Table 2: Minimum and maximum parameter values for model development

Parameter	Minimum	Maximum
Soil Depth, Ds [m]	1	37.5
Soil Young's modulus, E [kPa]	65 000	700 000
Height, H [m]	3.5	87.5
Length, L [m]	5	125
Width, B [m]	3	39

Table 3: Cases for plan inputs based on the mode considered.

Orientation of columns' strong axis (x or y)	Mode (1 or 2)	Order
x	1	L x B
x	2	B x L
y	1	B x L
y	2	L x B



(a) (b)

Figure 7: Cases input referring to (a) 0 for the strong axis of column parallel to the long direction of the building and (b) 1 for the strong axis of column sections parallel to the short direction of the building.

It is important to note at this point that the FE software, Femap, was used to graphically develop the models, whereas Reconan FEA [12] was used to perform the modal analyses. Due to the large data set, it was deemed necessary to utilize an HPC to perform the required number of modal analyses. For this reason, an HPC

was used in the analysis phase so that models could be analysed in parallel. The analysis of this many models demands a high computational effort that a standard desktop computer would not be able to handle. Finally, it must be noted that the total storage required for the input files was 800 GB, and the output files were 10 TB.

The solution of the numerical problem that computes the natural modes of each structure, with and without a soil domain assumes the detailed simulation of mass and stiffness of the structure. Therefore, there are several methods for solving the eigenvalue problem which is described by the equation below:

$$\mathbf{K}\boldsymbol{\varphi}_i = \lambda_i \mathbf{M}\boldsymbol{\varphi}_i \tag{17}$$

where **K** is the stiffness matrix, **M** is the mass matrix of the model, $\boldsymbol{\varphi}_i$ are the eigenvectors of the system and λ_i are the corresponding eigenvalues. For the needs of this research work, the subspace iteration algorithm [31] was adopted. One of the main advantages of this numerical technique is the ability to compute a specific number of eigenvalues and eigenvectors of a significantly demanding FE system.

The subspace iteration algorithm finds an orthogonal basis of vectors $\boldsymbol{E}_{K+1}$, and the required eigenvectors are calculated when $\boldsymbol{E}_{K+1}$ converges to $\boldsymbol{E}_{\infty}$: For k = 1, 2, …, iterate from $\boldsymbol{E}_k$ to $\boldsymbol{E}_{k+1}$:

$$\boldsymbol{K}\boldsymbol{X}_{k+1} = \boldsymbol{M}\boldsymbol{x}_k \tag{18}$$

then, finds the projections of matrices **K** and **M** onto $\boldsymbol{E}_{k+1}$

$$\boldsymbol{K}_{k+1} = \boldsymbol{X}_{k+1}^T \boldsymbol{K}\boldsymbol{X}_{k+1} \tag{19}$$

$$\boldsymbol{M}_{k+1} = \boldsymbol{X}_{k+1}^T \boldsymbol{M}\boldsymbol{X}_{k+1} \tag{20}$$

and solves for the eigensystem of the projected matrices:

$$\boldsymbol{K}_{k+1}\boldsymbol{Q}_{k+1} = \boldsymbol{M}_{k+1}\boldsymbol{Q}_{k+1}\boldsymbol{\Lambda}_{k+1} \tag{21}$$

Thereafter, find an improved approximation to the eigenvectors:

$$\boldsymbol{x}_{k+1} = \boldsymbol{X}_{k+1}\boldsymbol{Q}_{k+1} \tag{22}$$

Then, provided that the vectors $\boldsymbol{x}_1$ are not orthogonal to one of the equilibrium eigenvectors, $\boldsymbol{\Lambda}_{k+1} \rightarrow \boldsymbol{\Lambda}$ and $\boldsymbol{X}_{k+1} \rightarrow \boldsymbol{\Phi}$ as k → ∞.

It must be noted at this point that the convergence of this method assumes that within the iteration procedure the vectors in $\boldsymbol{X}_{k+1}$ are ordered in such a way that the i^{th} diagonal element in $\boldsymbol{\Lambda}_{k+1}$ is always larger than the previous i-1 element, i = 2, ..., p. This ensures that the i^{th} column in $\boldsymbol{X}_{k+1}$ converges to $\boldsymbol{\Phi}_i$.

It is noteworthy to state that the largest model that was fully developed and analyzed contained a total of $4.33 \text{x} 10^9$ dof, which required the inversion of the stiffness matrix with a size of 187,489 $\text{x}10^{14}$ elements. Since this demands a large computational time and effort, that a standard PC would not be able to handle, HPC was used to run the modal analysis to obtain the numerically predicted frequencies. It is important to keep in mind that this was only one of the 49,154 numerical models that required to be analyzed. Therefore, it was deemed necessary to analyze multiple models in parallel, thus using batch scripts was required in order to achieve this task as the structure of the operating HPC system does not offer any visual aids, where all commands given to perform any analysis were done through the use of script files or the command prompt-based environment.

The initial task was to copy the files to the HPC, which was achieved by splitting the results into sets that were analyzed on the HPC system. Within these main folders, subfolders were developed that each contained 10 numerical models. The subfolders were then compressed, where tar files were developed. A series of batch scripts were used to open and run each numerical model, named: Gen_tar_list.sh, Run.sh, Sub_tmp.sh, Submit_job.sh, Extract_data.sh.

Since the Reconan FEA input file has the generic name Input.neu, it was necessary to first change the name of the output file from Input_EIGENFREQ.dat to OriginalFileName_EIGENFREQ.dat so that the different output results could be correlated to a file. After this was done, the corresponding results were stored within the final data set, consisting of 98,308 fundamental periods, which is currently the largest of its kind. Figure 8 shows the flow chart of the procedure that was developed for the needs of this research work which uses the HPC to manage the execution of modal analyses in parallel and thereafter collect and same the output data.

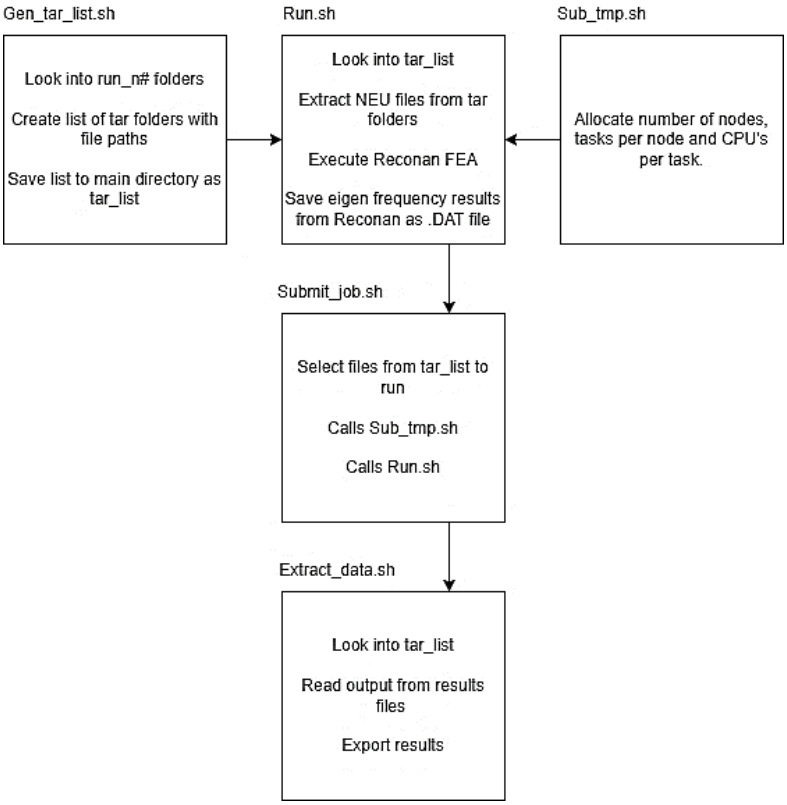


Figure 8: Flow-chart of the process of running and collecting FEA data through the HPC

After the successful completion of the data set analyses, the development of descriptive statistics was performed by the nbml code to analyze the input feature values as can be seen in Table 4. According to Table 4, D_S is the soil domain depth in m, E is the Young modulus of the soil in kPa, H is the height of the building in m, L is the length of the building in m, B is the width of the building in m and O represents the direction of the strong axis of columns as described in Table 4. Furthermore, Figure 9 shows the correlation matrix of all the input features, including the output T which is the fundamental period of the steel buildings in seconds.

Table 4: Descriptive statistics of the developed data set

	mean	**median**	**std**	**min**	**max**	**skewness**	**kurtosis**
Ds (m)	15.7	12.5	13.1	0	37.5	0.548	-1.061
E (kPa)	357 391	300 000	209 319	0	700 000	0.155	-1.225
H (m)	45.3	45.5	25.0	3.5	87.5	0.008	-1.182
L (m)	37.7	27.0	31.2	3.0	125.0	1.046	0.039
B (m)	38.3	27.0	31.5	3.0	125.0	1.003	-0.076
O	1.5	1.0	0.5	1.0	2.0	0.014	-2.000

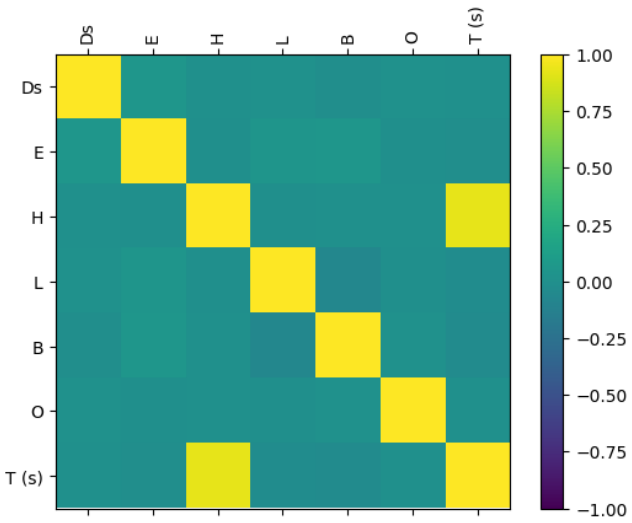


Figure 9: Correlation Matrix.

4. Validation of the finite element software

In this section, the validation procedure of the FE software Reconan FEA [12] will be presented through the use of experimental data found in the international literature. Two experiments [24, 25] were used to validate Reconan FEA [12] which was used to compute the fundamental period of all the steel-framed buildings. The 4-storey steel frame specimen presented in [24] can be seen in Figure 10. The framed specimen was discretized with NBCFB elements, the masses were applied at the level of the four levels according to the experimental setup, where a modal analysis was performed to compute the natural frequencies. According to the numerically obtained results, Figure 11 illustrates the fundamental modal shape of the numerical model, whereas Table 5 shows the comparison between the experimentally [24] and numerically obtained fundamental frequencies. It is easy to observe that the difference between the two values is 0.5% demonstrating the ability of the software to capture the dynamic feature of the specimen.

Table 5: Comparison between experimental [24] and numerical results

Frequency [Hz]		
Test	Numerical Analysis	Error
4.10	4.12	0.5%

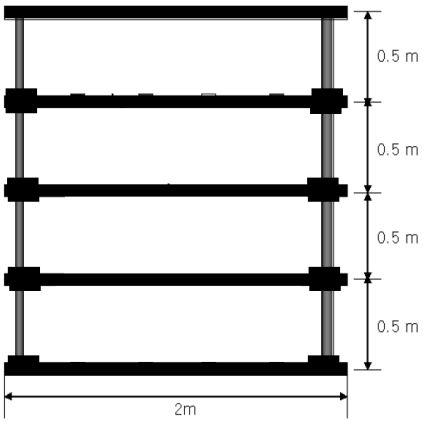


Figure 10: Dimensions of the steel framed specimen as constructed and tested in [24]

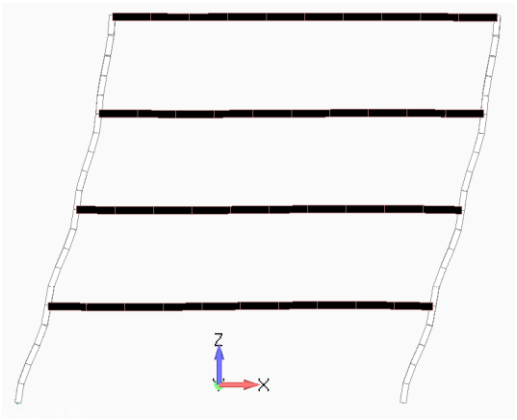


Figure 11: Numerically obtained fundamental modal shape through the use of Reconan FEA

A second, more recent seismic table experiment was performed and published in [25], where a 2-storey steel framed specimen was tested under dynamic loading conditions. The geometry and dimensions of the specimen can be seen in Figure 12. The fundamental modal shape obtained through the numerical analysis using Reconan FEA can be seen in Figure 13, where the numerically computed natural frequency was equal to 10.12 Hz. The modeling approach used to discretize and model the specimen was the same as described in the previous test. According to Table 6, the experimentally reported natural frequency [25] was 10.06 Hz, therefore, the difference between the experimentally and numerically obtained values is a mere 0.66%. This finding further highlights the ability of the adopted software [12] to capture this dynamic characteristic of steel structures. It is also important to note here that the ability of Reconan FEA [12] software to capture the behaviour of the SSI effect was presented in [26-28] where different specimens were numerically modelled in order to investigate the SSI phenomenon.

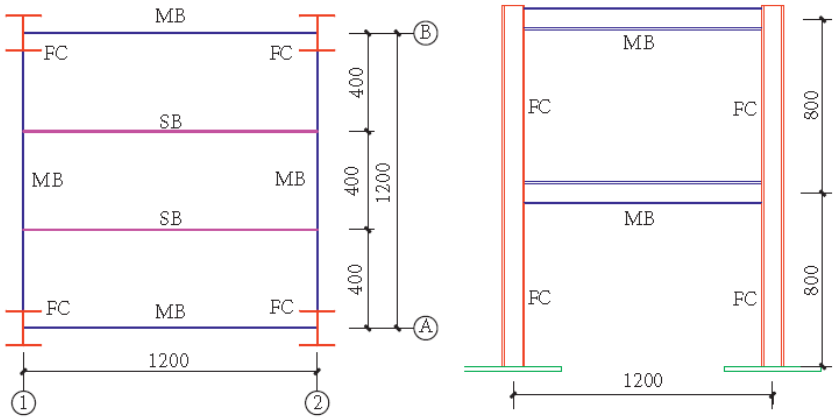


Figure 12: Dimensions of the steel framed specimen as presented in [25]. (Left) Plan view and (Right) side view schematics of the specimen

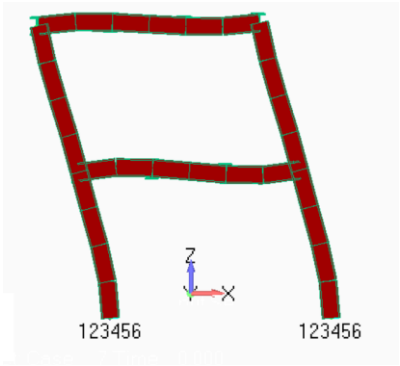


Figure 13: Numerically obtained fundamental modal shape. Side view of the modal shape

Table 6: Comparison between experimental results [25] and numerical method results

Frequency [Hz]		
Test	Numerical Analysis	Error
10.06	10.127	0.7%

5. Numerical results obtained from different ML algorithms

As previously stated, the open-source software used for the development of the ML-based predictive models was nbml[2], a software that has the ability not only to train and test the predictive models but also performs

[2] https://github.com/nbakas/nbml

an internal validation of the models through the check of the convergence trends [19] during the training and testing processes.

According to [19], iterative training is performed for all methods and for a partial random subset of the train set, starting from 25% up to 100% of the observations, with 20 intermediate new training iterations. Thus, by applying this process, the quality and adequacy of the data can be assessed by evaluating the shape of the curve depicting the performance of each individual training. It must be stated here that according to the findings of the parametric investigation performed in this research work related to steel structures, the numerically obtained curves were found to stabilize, which was an indication that the data within the training set were sufficient.

Furthermore, when using nbml a sensitivity analysis is performed to evaluate each feature's impact within the derived predictive model. Therefore, after the trained models are constructed, all features are kept constant at a specific value (i.e., 25%, Median, and 75% quantiles), allowing the permutation of only one feature around its assigned values. These values are set in training mode, and the corresponding sensitivity curves for each feature and each of the trained models are obtained. Thus, the corresponding impact of each one of the features on the target variable can be identified.

Finally, an error analysis is performed to further validate the obtained predictive models, which is one of the most critical issues [19], which has not been adequately addressed in previous studies. In this work, the difference between the prediction and target variable for each model is computed for the training and testing sets, and an error analysis was performed, comprising the:

1. The residual errors vs. target diagrams, and
2. The probability density functions and cumulative density functions of the errors.

This numerical procedure identifies specific patterns occurring in the prediction results that could impair both the generalization capability and reliability of the model. Conclusively, in addition to validation that is performed through the use of out-of-sample data, the developed ML algorithms [19] are integrated with several checks related to the performance of the ML-generated predictive models.

It is interesting to note at this point that there are numerous methods through which ML algorithms are implemented for the solution of finite element models [33, 34] and are not only used in the development of predictive models that derive from datasets that consist of analytically or experimentally obtained data. For the needs of this research work, standard finite element analysis was performed for developing the large dataset presented in section 3.

5.1 Polynomial Regression

The algorithm of the polynomial regression with hyperparameter tuning is described in this section, where the performance of the proposed algorithm is presented. The developed ML algorithm was initially adopted from [3] (see Algorithm 1 [3]) and can provide a closed-form solution. This involves the creation of several nonlinear terms that are formed with combinations of the independent variables with p, which represents the total number of polynomial features. Multiple polynomial regression models were applied to determine the effect, of changing the number of polynomial degrees (d), number of folds (f), and number of runs (r). The numbering of the polynomial regression models is given with the example *3d_5f_100r*, where 3 degrees was used with 5 folds and 100 runs assumed.

The first model developed was the *3d_5f_100r*. It is important to note here that the developed data set was divided into 85% and 15% training and testing data sets according to the findings presented in [34]. It was

derived from this analysis that the correlation achieved was 96.5%, and the MAPE was found to be 18%. This relatively high error could be attributed to the low number of folds used in the model. It was also found that as the periods increase, the capabilities of the model to predict the fundamental period decreases. An additional finding was that for higher frequencies, the model predicts negative periods which, in reality, is not possible.

Algorithm 1: Polynomial feature selection algorithm [3]

Algorithm: Higher Order Regression
Input: *XX* (matrix of Independent Variables), *YY* (Vector of Dependent Variable),
*nlf (*number of nonlinear features to be kept in the model)
Output: Prediction Formulae
1. Create all nonlinear features* (*anlf*)
2. For *i* from 1 to *nlf* do
3. For *j* from 1 to *anlf* do
4. Add j^{th} feature to the model
5. Calculate Prediction Error, $MAPE_j$**
6. End
7. Keep in the model the j^{th} feature which yields the minimum prediction error
5. End
Return: Prediction Formula
*with all inter-items combinations up to the x^{th} degree, **Mean Absolute Percentage Error (MAPE).

Next, the number of folds was increased to 100, and a new model was developed. The *3d_100f_100r* model achieved an increased correlation equal to 97.7%, and the MAPE was reduced to 13%. A similar observation to the previous model resulted where the predictive capabilities decreased and increased in terms of fundamental period values and negative periods were also observed. Although lower errors were computed through this model, it results in a more computationally heavy and time-consuming model. In order to see the effect that the number of runs has, the initial model was altered to contain 1,000 runs, which was named *3d_5f_1000r*. In this case, the correlation was found to be 98.6%, and the MAPE was 9%, which is an improvement compared to the previous models but still, negative periods are predicted, which is not acceptable.

Thereafter, both the folds and runs were increased to develop the model *3d_100f_1000r.* When compared to the *3d_5f_1000r* model, it was observed that the correlation has not changed, while the MAPE has also remained at 9%. This shows that for 1,000 runs, increasing the number of folds does not influence the predictive capabilities of the model. Again, using a larger number of folds increases computational time and effort to train and test the predictive models. It is also noted that the model is not able to accurately predict the period with values over 3 seconds and predicts negative values for low periods, which is a numerical phenomenon that highlights the limitations of this ML algorithm. Finally, the number of degrees was altered to 5 to assess the above-mentioned influence. Four additional models were then developed, and similar trends were followed to that of the 3d models. It was found, however, that using 5 degrees did not improve the accuracy of the models.

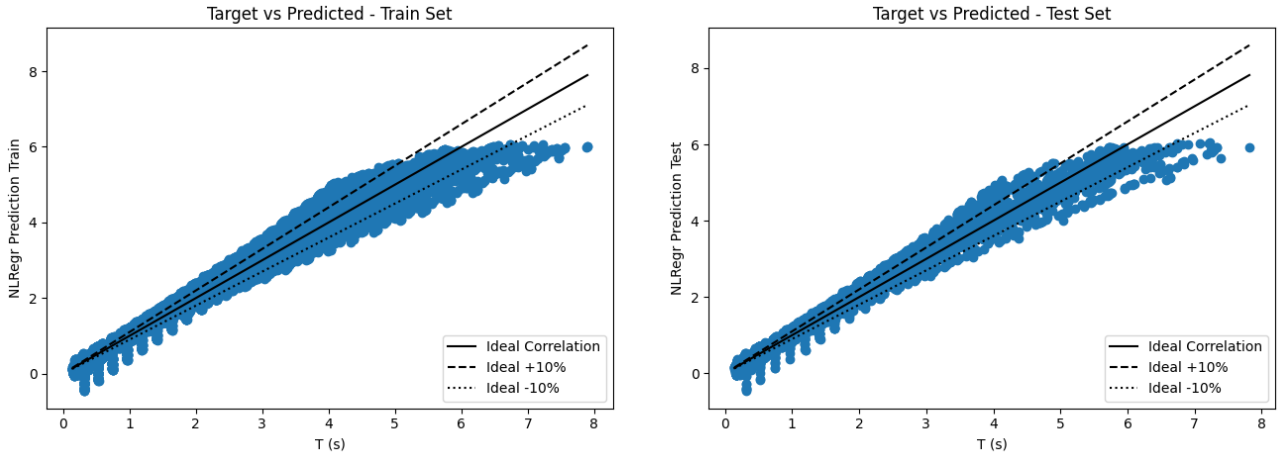
A new polynomial regression algorithm was proposed in [19], which was also used for the needs of this research work in developing a closed-form solution. According to the proposed polynomial feature selection algorithm shown in Algorithm 2, the proposed ML algorithm uses hypermeter tuning. Furthermore, based on the numerical investigation presented in [19], the authors claimed that the proposed POLYREG-HYT showed improved accuracy when trained on different data sets related to structural mechanics. Therefore, the proposed POLYREG-HYT is also used herein for the development of a closed-form formula, where it is compared to the best predictive model developed through the use of Algorithm 1 and the work suggested in [3].

Two new numerical analyses were performed through the use of the newly proposed ML algorithm POLYREG-HYT [19], where the first foresaw the use of 3rd-degree polynomial regression, 5 folds, and 1,000 tuning rounds (*3d_5f_1000r**) and the second analysis assumed the same values except from the degree which was set to 5 (*5d_5f_1000r**).

Figure 14 and Figure 15 show the comparison between the numerically predicted fundamental periods and the numerically computed ones according to the train and test data sets. It is evident that the newly proposed algorithm [19] outperforms the older algorithm that was proposed in [3]. The complete error metrics that resulted from the new POLYREG-HYT can be seen in Table 7 and Table 8 for the case of 3rd and 5th-degree polynomials, respectively. As can be seen, the 5th-degree model (*5d_5f_1000r**) outperforms the *3d_5f_1000r** model, where the numerically obtained MAPE on the test data set was found to be 8.65% and 4.84%, respectively. It is evident that the new POLYREG-HYT outperforms the older polynomial regression algorithm, thus, the findings reported in [19] are also verified herein.

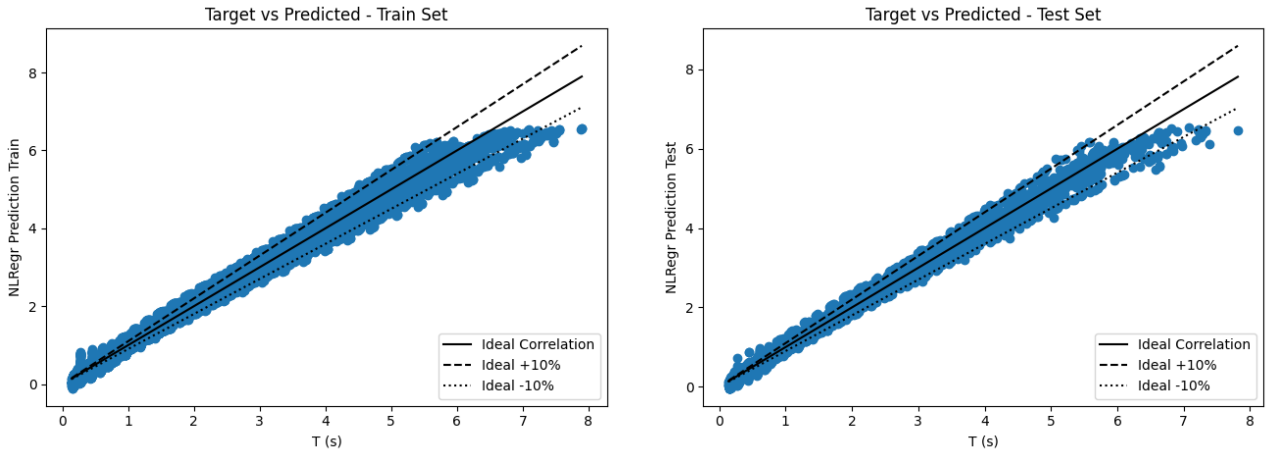
Algorithm 2: Polynomial feature selection algorithm [19]

Data: $\mathbf{X}, \mathbf{y}, m_f$ (maximum number of features)
Result: Initialize $[o] = 1$ with the constant term $\in [p]$
Solve Linear System $\mathbf{X}' \times \mathbf{a} = \mathbf{y}$, where $\mathbf{X}' \subset \mathbf{X}$, with $[o]$ columns.
Compute regression errors e_1.
Set as optimal error $\hat{e} \leftarrow e_1$.
Set as optimal indices $[\hat{o}] \leftarrow [o]$.
for $i \in [1, 2, \ldots, l]$ **do**
  repeat
    Select an index $d \in [p]$ randomly.
    if $d \in [o]$ **then**
      $r \leftarrow \mathcal{U}(0, 1)$
      if $r < \frac{1}{2}$ **then**
        Select randomly $o_d \in [p] : o_d \notin [o]$
        $[o] \leftarrow ([o] \setminus d) \cup o_d$;
      else
        $[o] \leftarrow [o] \setminus d$;
      end
    else
      if $o < m_f$ **then**
        $[o] \leftarrow [o] \cup d$;
      else
        Select randomly $o_d \in [o]$
        $[o] \leftarrow ([o] \setminus o_d) \cup d$;
      end
    end
  until $rank(\mathbf{X}') \equiv o$;
  Solve Linear System $\mathbf{X}' \times \mathbf{a} = \mathbf{y}$.
  Compute regression error e_i.
  if $e_i < \hat{e}$ **then**
    $\hat{e} \leftarrow e_i$
    $[\hat{o}] \leftarrow [o]$
  else
    $[o] \leftarrow [\hat{o}]$
  end
end



a. Train set b. Test set

Figure 14: Comparison between numerically obtained fundamental periods and 3*d_5f_1000r** formula-predicted results for the POLYREG-HYT predictive model.



a. Train set b. Test set

Figure 15: Comparison between numerically obtained fundamental periods and *5d_5f_1000r** formula-predicted results for the POLYREG-HYT predictive model.

Table 7: POLYREG-HYT *3d_5f_1000r** model's error metrics on the training and test data sets

	R	**MAPE**	**MAMPE**	**MAE**	**RMSE**
Train	99.267%	8.603%	4.885%	12.127%	18.359%
Test	99.260%	8.651%	4.896%	12.114%	18.519%

Table 8: POLYREG-HYT *5d_5f_1000r** model's error metrics on the training and test data sets

	R	**MAPE**	**MAMPE**	**MAE**	**RMSE**
Train	99.754%	4.813%	2.550%	6.331%	10.649%
Test	99.746%	4.841%	2.584%	6.394%	10.870%

5.2 XGBoost with Hyperparameter Tuning and Cross-Validation

XGBoost with hyperparameter tuning and cross-validation (XGBoost-HYT-CV [19]) is used for the needs of this research work. Hyperparameters are settings that can be used to control the behaviour of the learning algorithms and can significantly improve the numerical response of the derived predictive models. According to [19], the proposed tuning is performed for the following significant training parameters,

1. Maximum Number of XGBoost Rounds $\in$ [10, 20, . . . , 1000],

2. Maximum tree depth ∈ [1, 7, 15],

3. Learning Rate Eta ∈ [0.05, 0.2, 0.5],

4. Colsample_bytree ∈ [0.5, 1],

5. Subsample ∈ [0.5, 1].

It is noteworthy to state here that the most common method is the k-fold cross-validation procedure, in which the data set is split into k non-overlapping subsets by keeping each time the i^{th} subset as a test set and the rest as a training set [20]. The nbml user is able to, within the algorithm, define the number of folds. The method works as follows [21]:

- the model is trained using *k-1* of the folds as training data and
- the resulting model is validated on the remaining part of the data.
- The performance measure reported by k-fold cross-validation is then the average of the values computed in the loop.

In general, a higher number of folds means that each model is trained on a larger training set and tested on a smaller test fold. In theory, this should lead to a lower prediction error as the models see more of the available data. In contrast, a lower k means that the model is trained on a smaller training set and tested on a larger test fold. Here, the potential for the data distribution in the test fold to differ from the training set is bigger, and on average it should be expected to obtain a higher prediction error [22]. A comparison between the numerically obtained fundamental periods and those obtained using the XGBoost-HYT-CV-generated model as described above can be seen in Figure 16.

The correlation was found to be equal to 99.999%, and the MAPE on the test set was found to be 0.243% as can be seen in Table 9. As anticipated, due to the large data set, this model is able to accurately predict the fundamental period of the understudy structures and produces better results than the linear and all the polynomial regression models. The model accurately predicts the fundamental period for all period results within the test data set and no negative periods are predicted, which is a significant improvement compared to the polynomial regression models.

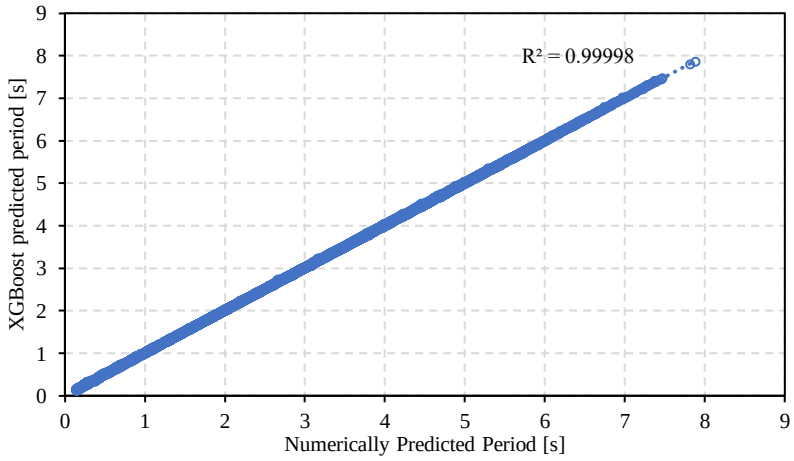


Figure 16: Comparison between numerically obtained fundamental periods and predicted results for the XGBoost-HYT-CV model on the test data set.

Table 9: XGBoost-HYT-CV model's error metrics on the training and test data sets

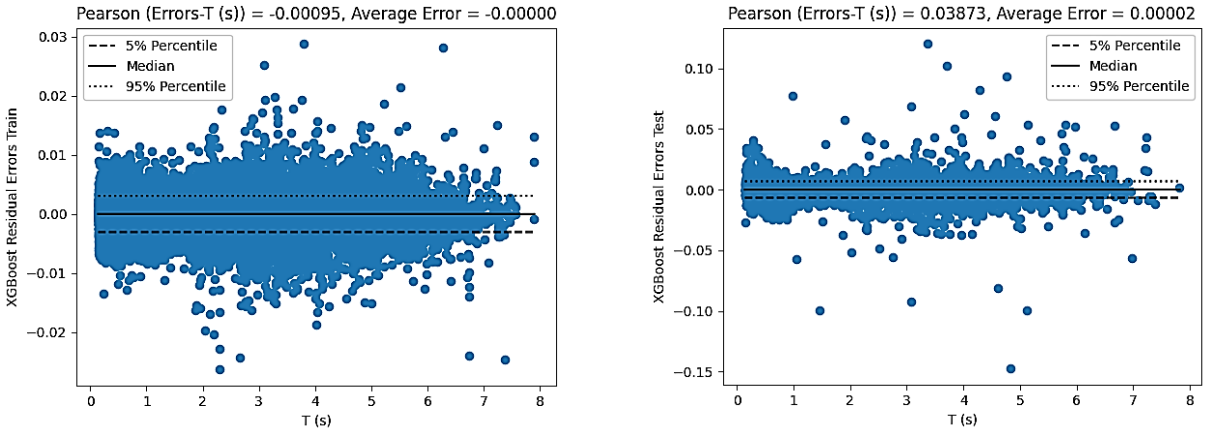
	R	MAPE	MAMPE	MAE	RMSE
Train	99.999%	0.108%	0.058%	0.144%	0.204%
Test	99.999%	0.243%	0.128%	0.317%	0.588%

The XGBoost-HYT-CV model was initially computationally demanding, so the number of folds and the number of runs were not increased. The model was also found to be accurate enough for all period values, therefore, it was not deemed necessary to run different XGBoost-HYT-CV models in an attempt to reduce the error. This highlights the superiority of this method when dealing with this type of problem.

Before moving to the next section, it is important to stress at this stage the importance of investigating the error analysis that was numerically obtained by the nbml code. As described in [19], the difference between the prediction and target variable for each model has to be computed for the training and testing sets, where the error analysis is performed. This comprises the residual errors vs target diagrams, the probability density functions, and the cumulative density functions of the errors.

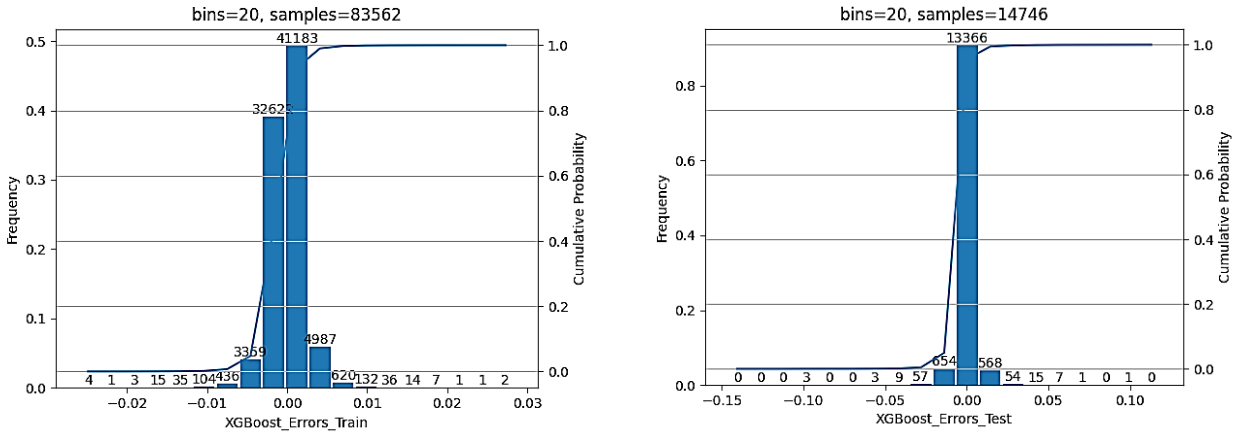
Figure 17a and b show the residual errors of the XGBoost-HYT-CV model vs the given fundamental period T, where it can be seen that for both train and test data sets the ML algorithm manages to develop an extremely accurate predictive model. In addition, this can be depicted in Figure 18, where the errors and cumulative distribution functions for training and testing can be seen. Finally, Figure 19 shows the history of the coefficient of determination during the training testing and validation [19], where it is seen that the ML algorithm reaches a plateau during the 100 tuning cycles that were performed for the needs of developing the predictive model.

Before discussing the results derived from the ANN method, it is important to note that a sensitivity analysis was performed and presented in Figure 20. It is easy to observe that the graph finds the height of the buildings to be the most important input feature with the width and length being second and third, respectively. This is in line with and also confirms the reason why design codes use the height of the buildings in computing the fundamental period of our structures. Additionally, it can be observed that the depth of the soil and its Youngs modulus have the lowest importance out of all the input features that were investigated in this research work.



a. Train set b. Test set

Figure 17: Residual errors of XGBoost-HYT-CV model vs given fundamental period T.



a. Train set b. Test set

Figure 18: Errors and cumulative distribution functions for training and testing obtained from the XGBoost-HYT-CV

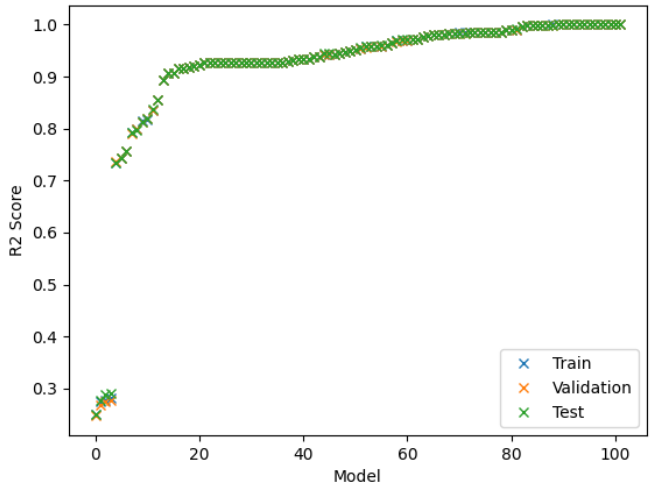


Figure 19: XGBoost-HYT-CV coefficient of determination R^2 history.

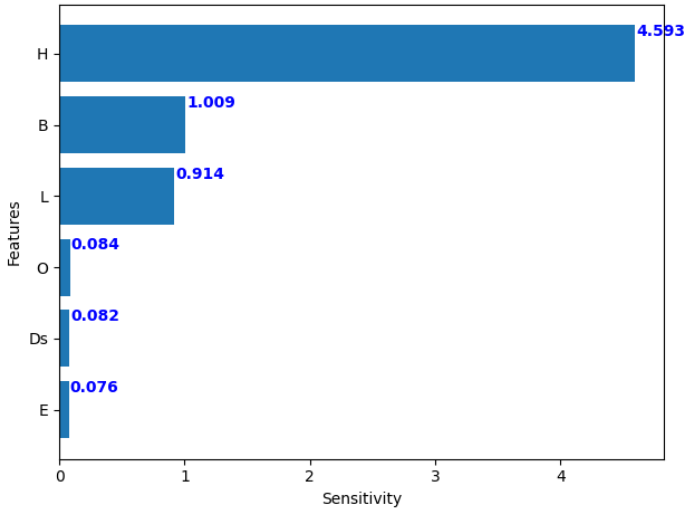


Figure 20: Sensitivity graph of the input parameters according to the XGBoost-HYT-CV predictive model.

5.3 Artificial Neural Networks Method

The final model applied was an ML model that foresaw the use of an ANN. The model was trained with the algorithm presented by [23] which is used to construct an ANN. Figure 21 shows the structure of the ANN used to train, test, and validate the corresponding predictive model [3, 23]. The architecture of ANNs consists of interconnected processing elements known as neurons, which are used to mimic how the human brain learns. The

input variables are represented by $x_{i1}, x_{i2}, .. x_{in}$ and v_k, the sought weights of the output layer. The sigmoid function, σ, is applied in the hidden layer for all input neurons and the output is a linear combination of nonlinear neurons. A comparison between the numerically obtained fundamental periods and those obtained using the ANN model can be seen in Figure 22. One hidden layer and 13,846 neurons were used for the development of the predictive model.

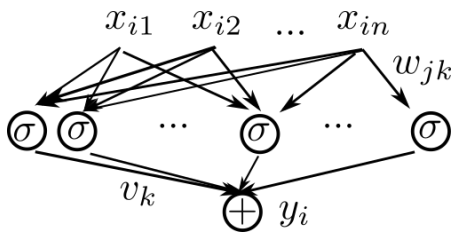


Figure 21: ANN architecture [3, 23].

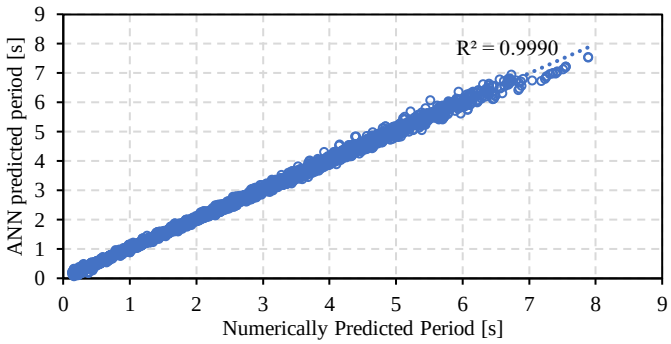


Figure 22: Comparison between the numerically obtained fundamental period results and the results from the ANN model.

The correlation was found to be 99.9%, and the MAPE was approximately equal to 0.7%. Compared to the XGBoost-HYT-CV method, the results do not fall on a straight line as closely as in Figure 16. For periods between 0 and 4.5 seconds, the predictions are slightly larger than the numerically obtained results, and for periods above 7 seconds, the predictions are slightly smaller than the numerically obtained results. Similar to the XGBoost-HYT-CV method, this model is simply numerical, and no closed-form formula is produced.

It is safe to conclude that the XGBoost-HYT-CV and ANN, were able to produce extremely accurate results for all periods in the test data set, whereas it is necessary to assess their predictive capability in determining the fundamental period of steel structures that had parameters that differ from those used to train and test the ML models. This is also known as out-of-sample data, which will be discussed in the following section.

6. Validation of predictive capabilities through out-of-sample data

Upon successful completion of the different closed-form solutions and ML models developed with XGBoost-HYT-CV and ANNs, the final phase of the research foresaw the validation of the predictive capabilities of the developed models presented thus far through the use of out-of-sample data. To further validate the models, an additional data set of 255 models (510 new numerically predicted periods) was developed, where the bay length in the x- and y-directions were modified in such a way that the models did not include geometrical characteristics nor soil domains depths and Young moduli values used during training and testing.

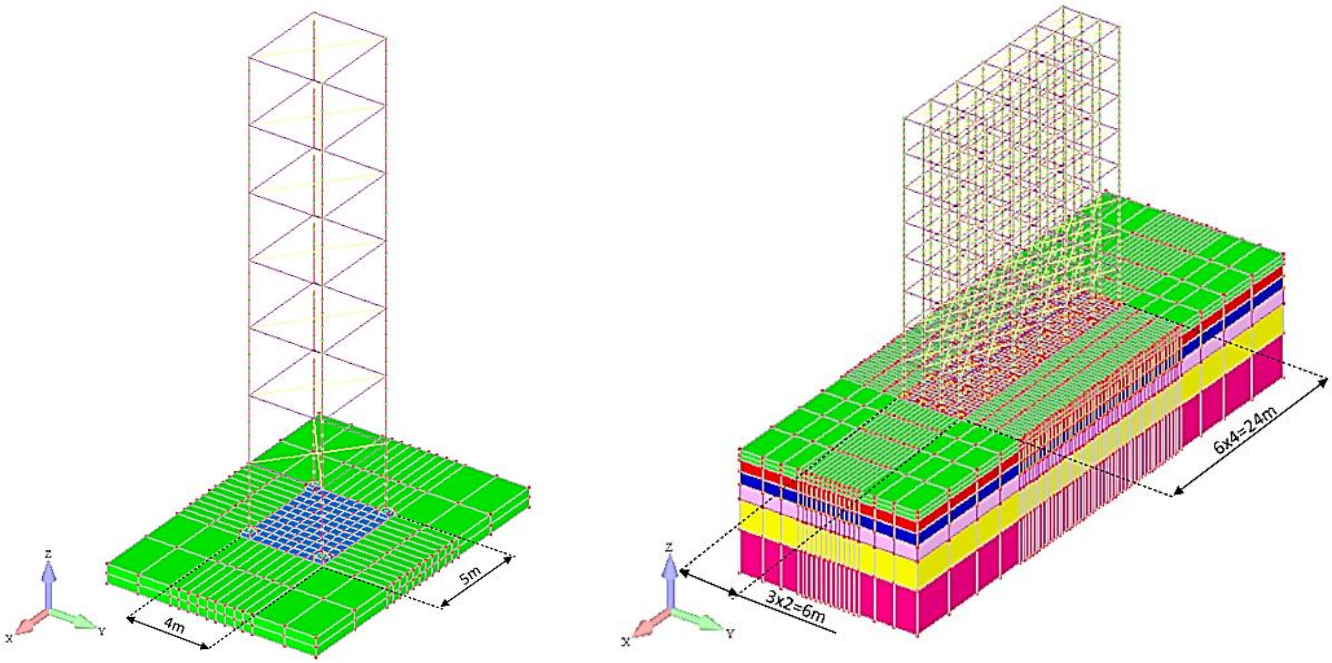


Figure 23: (Left) 110 MPa, 1 bay, 7-storey, (Right) 210 MPa, 6 bay, 8-storey models developed for validation stage.

The validation data set involved the change in plan area where the length along the long direction of the building was changed to 4 m for some models and 6 m for others. The length in the short direction was also altered to include a length of 2 and 4 m. Some models also included a change in the soils' Young modulus. Figure 23 shows two models that were developed and foresaw a 5x4 m plan area for a 1-bay, 7-storey building founded on 1 m soil with E =110 MPa and a 24x6 m plan area for a 6-bay, 8-storey building founded on 5 m soil with E = 210 MPa. A comparison between the numerically predicted periods and those predicted using the *3d_100f_1000r* model derived from the polynomial regression algorithm can be seen in Figure 24.

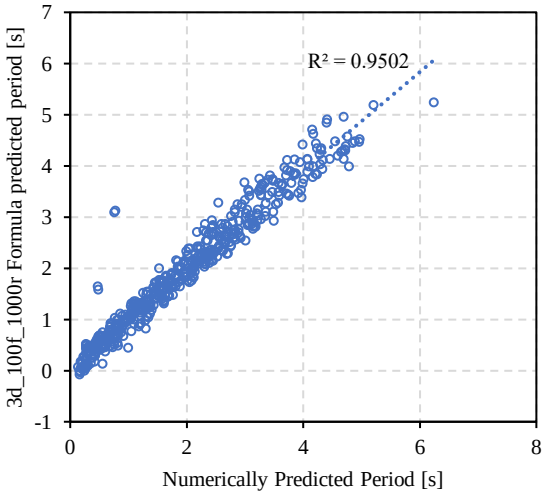


Figure 24: Polynomial regression predictive model. Comparison between the numerically obtained periods and formula obtained periods for the first validation data set for the *3d_100f_1000r* model.

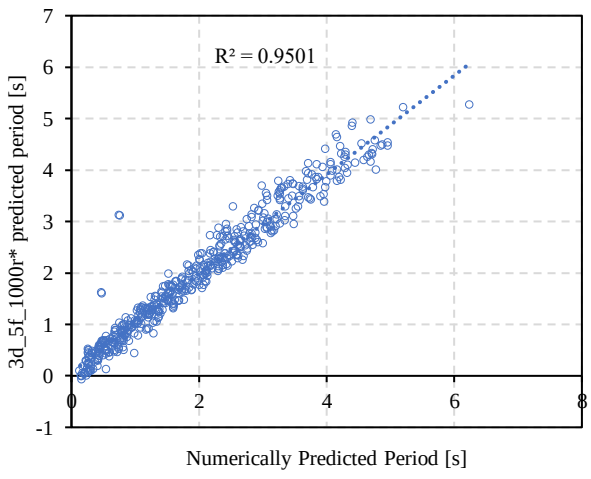


Figure 25: Polynomial regression predictive model. Comparison between the numerically obtained periods and formula obtained periods for the first validation data set for the *3d_5f_1000r** model.

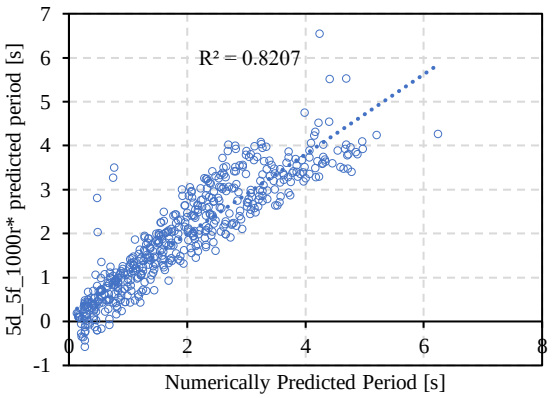


Figure 26: Polynomial regression predictive model. Comparison between the numerically obtained periods and formula obtained periods for the first validation data set for the 5*d_5f_1000r** model.

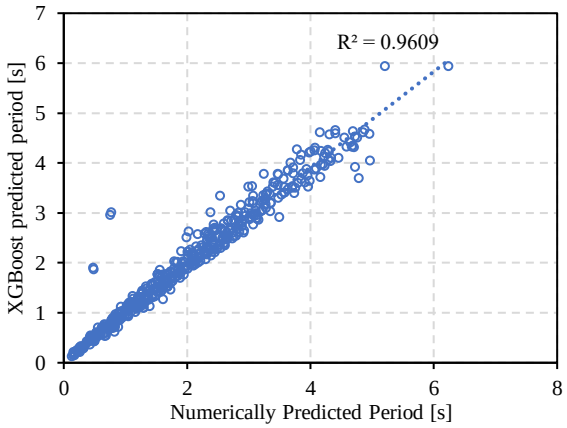


Figure 27: Comparison between the numerically obtained periods and XGBoost-HYT-CV obtained periods for the first validation data set.

It can be seen that the correlation is 95% and the MAPE was found to be 16%. The predictive capabilities for out-of-sample data are almost as good as that of the training data set. For periods of less than 3 seconds, there is a more accurate prediction for the developed formula as the results are closer to the straight line, while some of the predictions for low-rise buildings are found to be negative indicating the need for further improvement. On the other hand, the *3d_5f_1000r** (see Figure 25) is found to have a higher correlation (95%) compared to the *5d_5f_1000r** predictive model that resulted in 82% on the validation data set (see Figure 26). This is attributed to the overfitting phenomenon that results when high polynomial degrees are used during the training and testing of the closed-form formulae.

Next, the comparison between the numerically predicted periods and those predicted using the XGBoost-HYT-CV-generated model was performed as can be seen in Figure 27. It can be easily observed that the correlation between the numerically obtained results and those obtained using the XGBoost-HYT-CV-generated predictive model for the out-of-sample data, is 96%, where the computed MAPE was 9%. The predictive capabilities of the XGBoost-HYT-CV predictive model are better than that of the polynomial regression model when validating their ability to predict fundamental periods for out-of-sample results, whereas the XGBoost-HYT-CV models did not generate any negative fundamental periods. From the results, it can be concluded that the models are able to accurately predict the fundamental period even for out-of-sample models, therefore avoiding the overfitting phenomenon, with the XGBoost-HYT-CV predictive model being the best in terms of accuracy.

Table 10: Comparison between the MAPE of the ML models and current design code formulae (First validation data set).

Model	MAPE [%]
3d_100f_1000r	16
*3d_5f_1000r**	15.7
XGBoost-HYT-CV	**9**
Eurocode 8	33
ASCE	32
Cinitha [7]	52

Table 10 shows the MAPE comparison between the developed predictive models and the design codes including the equation proposed in [7]. It is easy to observe the XGBoost-HYT-CV model results in the lowest error metric when all the models are tested on the validation data set, where it results in approximately a 3.6 times

lower error compared to Eurocode 8 and ASCE. It is important to note that the formula presented in [7] only applies to buildings up to 30 m in height, so only 40% of the validation results contributed to the calculation of the MAPE seen in Table 10, which could explain an additional reason for deriving a high MAPE.

7. Validation of Predictive Capabilities for Irregular Buildings

Since the predictive models performed well on the first validation data set, a second out-of-sample data set, consisting of 50 additional models (100 modal analysis results), was developed. This second data set included structures with irregularities to test the predictive capability of the developed models in cases of irregular structures, thus exploring the proposed predictive models' limitations in capturing the fundamental period of steel-framed buildings. These models foresaw the reduction of the number of bays by a maximum of 30% of the plan view as seen in Figure 28.

The numerically predicted periods were plotted against those obtained from the *3d_100f_1000r* formula, as seen in Figure 29. The correlation was found to be equal to 86%, and the MAPE was found to be 16% as it was obtained from the *3d_100f_1000r* formula without resulting negative results. It is important to note here that the polynomial regression performs ideally when the data sets are smaller [3, 15], whereas in this case where the data set is larger, it is found that it does not provide optimal accuracy. The *3d_5f_1000r** predictive model that resulted from the newly proposed POLYREG-HYT, was found to result in a corresponding 8.65% MAPE on the test data set, outperforming the *3d_100f_1000r* formula, where the numerical findings reported in [19] are once more verified in this research work. In addition to that, the MAPE that resulted from the *3d_5f_1000r** (see Figure 30) was also 16% on the first validation data set, which was found to be the best predictive model out of all the polynomial regression-derived predictive models. The respective MAPE on the second validation data set for the case of the *3d_5f_1000r** model was 15.7%, which is lower than the *3d_100f_1000r* which resulted in a 16% when used to predict the fundamental period of irregular buildings.

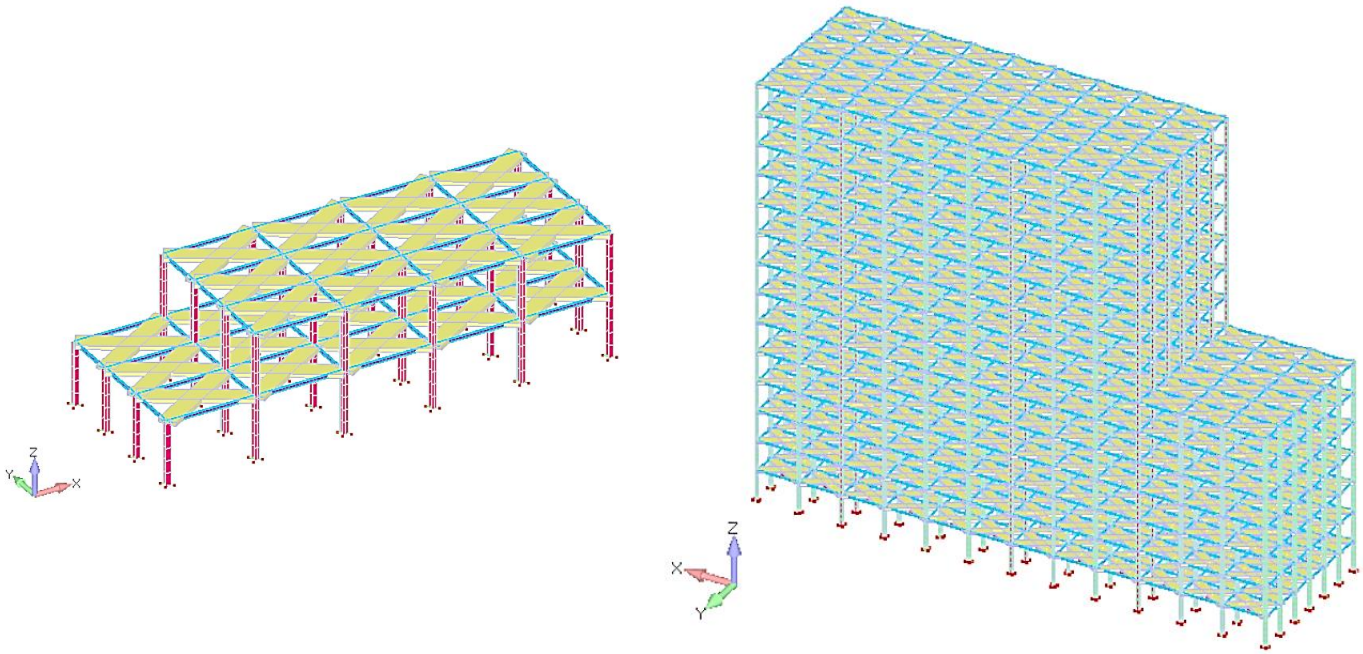


Figure 28: Second set of validation models. (Left) 5-bay, 2-storey, and (Right) 12-bay, 13-storey irregular steel framed buildings.

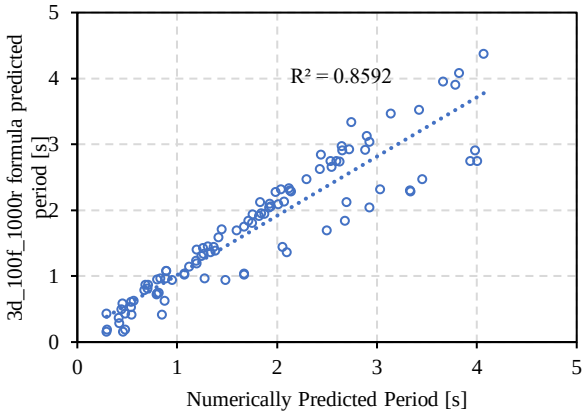


Figure 29: Comparison between the numerically predicted periods and formula-predicted results for the *3d_100f_1000r* model for the second validation data set with irregular frames.

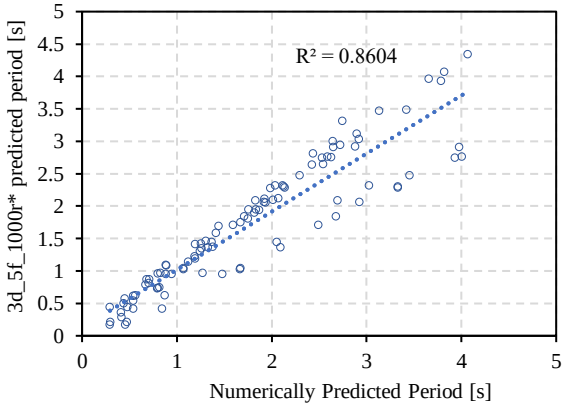


Figure 30:Comparison between the numerically predicted periods and 3d_5f_1000r* results for the second validation data set with irregular frames.

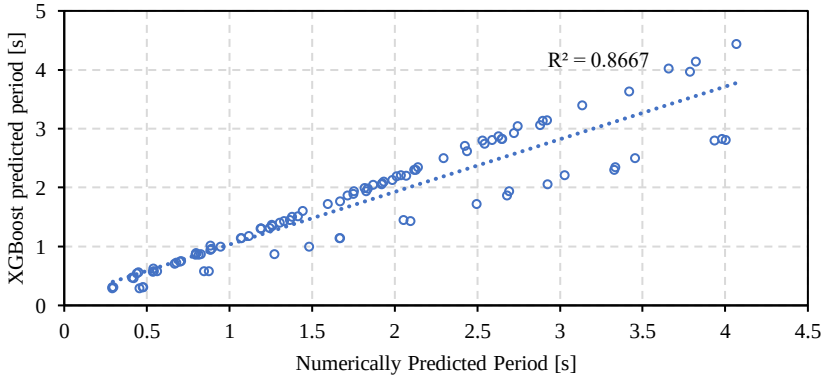


Figure 31:Comparison between the numerically predicted periods and XGBoost-HYT-CV results for the second validation data set with irregular frames.

On the other hand, for the case of the XGBoost-HYT-CV model (see Figure 31), the correlation was found to be equal to 87%, with a corresponding MAPE of 9%, outperforming all the polynomial regression formulae. Overall, the correlations do not appear to be high since the frame irregularity was not considered in the development process of the proposed predictive models. The fundamental periods, however, are both over- and under-estimated for all period values, a numerical finding that shows the need for future research that will foresee the development of models that will account for the frame's irregularities, whereas this will be accounted for as an input feature within the data set.

8. Conclusions and recommendations

Initially, a data set was developed that was large enough to train with, thus utilizing as many numerical modal results as possible. Approximately 50 000 models were analysed on HPC, resulting in the largest to-date data set on modal results related to steel-framed buildings with and without SSI effects. The exact number of data found in the generated data set is 98 308 fundamental period results.

The linear regression model was found to have developed a formula that resulted in a correlation of 85% and a MAPE of 23%. Although this correlation was relatively high and the MAPE was quite low (compared to the current formulae found in the literature), it was found that the model both under- and over-estimates the period for all values. This model had the worst numerical response compared to all the ML-generated predictive models

presented in this study. This was expected since the linear regression method is limited when it comes to establishing a connection between input and output that exhibit nonlinear behaviour.

The next ML model that was applied was the polynomial regression model. In this case, eight formulae were developed with varying folds, runs, and degrees. The *3d_5f_1000r** model was the optimum out of the generated polynomial regression formulae with a correlation of 99.3% and a MAPE of 8.65% on the test data set. In this model, it was found that its predictive capabilities reduced for higher values of the fundamental period. The disadvantage of these models was the calculation of negative fundamental period values for high-frequency structures. This numerical phenomenon is attributed to the large data set that was used to train, possibly pushing the methodology's limitations. This also illustrates the well-known fact that there are no fit-all solutions when training for different data sets.

The next method that was utilized herein was the XGBoost-HYT-CV, which was found to outperform all other ML algorithms that were investigated in this research work. This method derived a significantly high correlation of 99.99%, where the fundamental periods were predicted to almost the exact results found through the numerical analysis. The corresponding MAPE was also found to be optimal with a value of 0.5%, which further validates the accuracy of the proposed predictive model and its numerical superiority when dealing with this type of problem.

Finally, an ANN was implemented, which resulted in a correlation of 99.9% and a corresponding MAPE of 0.7% on the test data set. Although these results were optimal, some deviations were still found in the prediction of fundamental periods between 0 and 4.5 seconds and above 7 seconds. In general, the ANN approach is not a fit-all solution, where the optimum structure of the network is not known at the beginning of the training. Nevertheless, the network developed for the needs of this research work was found to be ideal, deriving a minimal error, slightly larger than that obtained from the XGBoost-HYT-CV.

The ML algorithms all produced acceptable results on the training data set, but the best models were further validated using a data set composed of out-of-sample results. The models used in the validation phase were the *3d_100f_1000r*, *3d_5f_1000r**, and XGBoost-HYT-CV. All models were able to accurately predict the period value of the out-of-sample results with a maximum MAPE of 16% derived from the polynomial regression model *3d_100f_1000r*, and the same 16% for the case of *3d_5f_1000r**, where the *3d_100f_1000r* model derived once more some negative periods for high-frequency structures, indicating that the model requires further improvements. On the other hand, the improved feature selection algorithm proposed in [19] was used, where the model *3d_5f_1000r** was developed. This model demonstrated higher accuracy and advanced extendibility when tested on the large data set and the two validation data sets. On the other hand, the XGBoost-HYT-CV outperformed the polynomial regression models by deriving a 9% MAPE on the second validation data set, highlighting its numerical superiority.

Given that the proposed models' numerical response was deemed to be accurate when implemented on the out-of-sample validation data set, a second validation data set was developed to test the predictive capabilities of the models when used to predict the fundamental periods of irregular structures. It was found that the models had a relatively good performance, with the polynomial model *3d_100f_1000r* deriving a 16% MAPE, the *3d_5f_1000r** model a 15.7%, whereas the XGBoost-HYT-CV gave a 13%. None of the proposed predictive models derived negative fundamental period values. According to these numerical results, it is deemed necessary to initiate a new research project that will foresee the numerical investigation of irregular steel framed structures where the irregularity will be accounted as an input feature.

It is also recommended that the proposed formulae only be applied to steel structures that have geometrical and soil properties that form part of the main data set presented herein. Furthermore, it is recommended that the XGBoost-HYT-CV model would be used for real practical implementations since it was found to be the most accurate and robust predictive model.

Finally, future work is required where the natural periods of buildings with irregularities have to be computed; therefore, this frame feature should be included in a future data set that is going to be used to train and test the ML algorithms, so formulae can be developed to determine the fundamental period of such structures. Furthermore, X-bracing and infill walls should be considered in the development of fundamental period formulae and predictive models, with and without accounting for the SSI phenomenon. Finally, different column sections have to be considered, as presented in the pilot project by Duan et al., [29].

ACKNOWLEDGMENTS

The financial support from the EuroCC project (GA 951732) and EuroCC 2 project (101101903) of the European Commission is acknowledged. Parts of the runs were performed on the MeluXina (https://docs.lxp.lu/) as well as Cyclone (https://hpcf.cyi.ac.cy/) Supercomputers.

REFERENCES

[1] van der Westhuizen, A. M. Markou, G. and Bakas, N. 2022. Development of a New Fundamental Period Formula for Steel Structures Considering the Soil-structure Interaction with the Use of Machine Learning Algorithms. *14th International Conference on Agents and Artificial Intelligence 2022*, Vol3, pp 952-957. https://doi.org/10.5220/0010978400003116.

[2] Saadi, D. 2018. *Nonlinear FEA of Soil-Structure-Interaction Effects on RC Shear-Wall Structures*. MSc thesis. American University of Sharjah.

[3] Gravett, D. Z. Mourlas, C. Taljaard, V.L. Bakas, N. Markou, G. and Papadrakakis, M. 2021. New fundamental period formulae for soil-reinforced concrete structures interaction using machine learning algorithms and ANNs. *Soil Dynamics and Earthquake Engineering*, Vol 144, pp 106656. https://doi.org/10.1016/j.soildyn.2021.106656.

[4] Eurocode. 2004. CEN-8: *Design of structures for earthquake resistance. Part 1: general rules, seismic actions and rules for buildings*. European Standard EN 1998-1:2004, Comit'e Europ'een de Normalisation, Brussels, Belgium.

[5] South African National Standard (SANS). 2009. *Seismic actions and general requirements for buildings.* SANS 10160-4, Pretoria, South Africa.

[6] American Society of Civil Engineers. 2010. *Minimum Design Loads for Buildings and Other Structures.* ASCE/SEI 7-05, -10, Reston, Virginia.

[7] Cinitha, A. 2012. A rational approach for fundamental period of low and medium rise steel building frames. *International Journal of Modern Engineering Research*, Vol 2, No 5, pp 3340-3346.

[8] Young, K. and Adeli, H. 2014. Fundamental period of irregular moment-resisting steel frame structures. *The structural design of tall and special buildings*, Vol 23, No 15, pp 1141-1157. https://doi.org/10.1002/tal.1112.

[9] Khalil, L. Sadek, M. and Shahrour, I. 2007. Influence of the soil–structure interaction on the fundamental period of buildings. *Earthquake engineering & structural dynamics*, Vol 36, No 15, pp 2445-2453. https://doi.org/10.1002/eqe.738.

[10] Mourlas, C. Khabele, N. Bark, H. A. Karamitros, D. Taddei, F. Markou, G. and Papadrakakis, M. 2020. Effect of Soil–Structure Interaction on Nonlinear Dynamic Response of Reinforced Concrete Structures. *International Journal of Structural Stability and Dynamics*, Vol 20, No 13, pp 2041013. https://doi.org/10.1142/s0219455420410138.

[11] Taljaard, V.-L. Gravett, D. Z. Mourlas, C. Bakas, N. Markou, G. and Papadrakakis, M. 2021. Development of a New Fundamental Period Formula by Considering Soil-Structure Interaction with the Use of Machine Learning Algorithms. *COMPDYN 2021*. 27-30 June 2021. Athens, Greece. https://doi.org/10.7712/120121.8748.18534.

[12] Reconan FEA v2.00, 2020, *User's Manual*. 10.13140/RG.2.2.28378.77761. https://www.researchgate.net/publication/342361609_ReConAn_v200_Finite_Element_Analysis_Software_User's_Manual

[13] Bakas, N. P. 2019. Numerical Solution for the Extrapolation Problem of Analytic Functions. *Research*, Vol 2019. https://doi.org/10.34133/2019/3903187.

[14] Markou, G. and Bakas, N. P. 2021a. Prediction of the Shear Capacity of Reinforced Concrete Slender Beams without Stirrups by Applying Artificial Intelligence Algorithms in a Big Database of Beams Generated by 3D Nonlinear Finite Element Analysis. *Computers and Concrete*, 28(6), 433-447. https://doi.org/10.12989/cac.2021.28.6.533.

[15] Markou, G. and Bakas N.P. 2021b, Developing reinforced concrete structures through AI algorithms and large-scale modelling. *Innovate*, 16, 38-40.

[16] Thai, H-T. 2022. Machine learning for structural engineering: A state-of-the-art review. *Structures*, Vol38, pp 448-491. https://doi.org/10.1016/j.istruc.2022.02.003

[17] Ghunaim, F. 2022. *AI vs Machine Learning vs Deep Learning: What's the Difference*. Available at: https://www.sitech.me/blog/ai-vs-machine-learning-vs-deep-learning.

[18] Dutta, N. Subramaniam, U. and Padmanaban, S. 2020. Mathematical models of classification algorithm of Machine learning. In International Meeting on Advanced Technologies in Energy and Electrical Engineering. *Hamad bin Khalifa University Press (HBKU Press)*, Vol 2020, pp 3. https://doi.org/10.5339/qproc.2019.imat3e2018.3.

[19] Markou, G., Bakas, N.P., Chatzichristofis, S.A. and Papadrakakis, M. (2024), General Framework of High-Performance Machine Learning Algorithms: Application in Structural Mechanics, *Computational Mechanics*, 2024, 73, 705–729.

[20] Bengio, Y. Goodfellow, I. and Courville, A. 2017. *Deep learning*. MIT press Cambridge, MA, USA.

[21] Pedregosa, F. Varoquaux, G. Gramfort, A. Michel, V. Thirion, B. Grisel, O. Blondel, M. Prettenhofer, P. Weiss, R. and Dubourg, V. 2011. Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, Vol 12, pp 2825-2830. https://doi.org/ https://arxiv.org/abs/1201.0490.

[22] Olsen, L.R. 2022. Multiple-k: Picking the number of folds for cross-validation. Available at: https://cran.r-project.org/web/packages/cvms/vignettes/picking_the_number_of_folds_for_cross-validation.html.

[23] Bakas, N. P. Langousis, A. Nicolaou, M. and Chatzichristofis, S. A. 2019. A Gradient Free Neural Network Framework Based on Universal Approximation Theorem. arXiv preprint arXiv:1909.13563.

[24] In-Kll, C. Kyu, K. M. Young-Sun, C. and Jeong-Moon, S. 2005. Shaking table test of steel frame structures subjected to scenario earthquakes. Nuclear Engineering and Technology, Vol 37, No 2, pp 191-200.

[25] Wang, T. Shao, J. Zhao, C. Liu, W. and Wang, Z. 2021. Shaking table test for evaluating the seismic performance of steel frame retrofitted by buckling-restrained braces. Shock and vibration, Vol 2021, pp 1-17.

[26] Braun, K.T., Bakas, N.P., Markou, G. and Jacobsz, S.W. (2023), Advanced Numerical Modelling of the Nonlinear Mechanical Behaviour of a Laterally Loaded Pile Embedded in Stiff Unsaturated Clay, SAICE, In Press.

[27] Gravett, Z.D. and Markou, G. (2021), State-of-the-art Investigation of Wind Turbine Structures Founded on Soft Clay by Considering the Soil-Foundation-Structure Interaction Phenomenon – Optimization of Battered RC Piles, Engineering Structures, 235, 112013.

[28] Markou, G., AlHamaydeh, M. and Saadi, D., "Effects of the Soil-Structure-Interaction Phenomenon on RC Structures with Pile Foundations", 9th GRACM International Congress on Computational Mechanics, Chania, Greece, 4-6 June 2018, pp. 338-345.

[29] Duan Calitz, George Markou, Nikolaos Bakas and Manolis Papadrakakis, Developing Fundamental Period Formulae for Steel Framed Structures Through Machine Learning and Automated Algorithms, COMPDYN 2023, 12-14 June 2023, Athens, Greece.

[30] Markou G., 2011, Detailed Three-Dimensional Nonlinear Hybrid Simulation for the Analysis of Large-Scale Reinforced Concrete Structures, National Technical University of Athens, Ph.D. Thesis, Greece.

[31] Bathe K.J., Solution Methods for Large Generalized Eigenvalue Problems in Structural Engineering. Report UCSESM 71-20, Department of Civil Engineering, University of California, Berkeley: 1971.

[32] Samaniego, E., Anitescu, C., Goswami, S., Nguyen-Thanh, V.M., Guo, H., Hamdia, K., Zhuang, X., and Rabczuk, T. (2020), "An energy approach to the solution of partial differential equations in computational mechanics via machine learning: Concepts, implementation and applications", Computer Methods in Applied Mechanics and Engineering, Volume 362:112790.

[33] Goswami, S., Anitescu, C., Chakraborty, S., and. Rabczuk, T, (2020), "Transfer learning enhanced physics informed neural network for phase-field modeling of fracture", Theoretical and Applied Fracture Mechanics, Volume 106:102447.

[34] Daniel Rademan and Geroge Markou, A Parametric Investigation of the Train-Test Ratio for Machine Learning Algorithms in Structural Mechanics Applications, ECCOMAS 2024, Portugal, 3-7 June 2024.