ACADEMIA OAMENILOR
DE ȘTIINȚĂ DIN ROMÂNIA

Raport științific nr. 4 privind proiectul

Platformă DIGItală pentru esTimarea nivelului de dEgradaRe seismică a clădiRilor utilizând tehnici de mAchine learning

DIGITERRA

Domeniu științific: Științe Geonomice

Aplicație web bazată pe algoritmi de machine learning pentru estimarea nivelului de degradare seismică

Cuvinte cheie: algoritmi de machine learning; analiza dinamică neliniară; degradare structurală; digitalizare; reziliență

Conf.dr.ing CALOFIR Vasile - **Director de proiect**
Cpt.dr.ing. MUNTEANU Ruben-Iacob - **Membru în echipa de cercetare**
ȘL.dr.ing. SIMOIU Mircea-Stefan - **Membru în echipa de cercetare**

Universitatea Naţională de Ştiinţă şi Tehnologie
POLITEHNICA BUCUREŞTI
Facultatea de Automatică și Calculatoare

Introducere

Estimarea precisă a degradărilor structurale seismice este un aspect fundamental al ingineriei seismice, esențial pentru proiectarea clădirilor reziliente și protejarea comunităților. Consecințele potențiale ale cutremurelor — inclusiv colapsul structural, perturbările economice și pierderile de vieți omenești — subliniază necesitatea critică de a înțelege modul în care diverse structuri răspund la forțele seismice. În plus, estimarea daunelor seismice joacă un rol vital în managementul dezastrelor și dezvoltarea codurilor de construcție și a standardelor de siguranță. Prin cuantificarea pierderilor economice potențiale, aceste evaluări facilitează o mai bună alocare a resurselor pentru pregătirea și recuperarea în caz de urgență. De asemenea, sectorul asigurărilor se bazează pe evaluările seismice pentru a evalua riscurile și a informa prețurile polițelor, promovând astfel reziliența economică în regiunile cu risc ridicat.

Evaluarea degradărilor seismice a devenit un domeniu critic de cercetare, conducând la diverse abordări inovatoare pentru evaluarea riscurilor seismice. Studii recente au introdus metode precum analiza time history la scară regională [8] și proiectarea bazată pe performanță utilizând curbe de fragilitate [19], care îmbunătățesc acuratețea evaluărilor de vulnerabilitate seismică pentru clădirile din zonele cu risc ridicat. Tehnicile care integrează deep learining și augmentarea datelor au arătat potențial în îmbunătățirea capacităților predictive pentru degradările seismice [4], facilitând evaluări mai rapide. Mai mult, progresele recente în tehnicile de modelare numerică au îmbunătățit acuratețea simulărilor seismice prin integrarea unor factori specifici suplimentari, cum ar fi coroziunea armăturii [14].

Accentul global bazat pe evaluarea eficientă a riscurilor se reflectă în inițiative precum Cadrul Sendai pentru Reducerea Riscului de Dezastre, proiectul SHARE (Seismic Hazard Harmonization in Europe) și Global Earthquake Model (GEM). Împreună, aceste eforturi subliniază necesitatea unor instrumente robuste de evaluare a riscurilor seismice capabile să evalueze siguranța clădirilor și să reducă impactul cutremurelor asupra comunităților din întreaga lume.

Practicile actuale de proiectare structurală încă utilizează analiza elastică și metodele bazate pe forță, cum ar fi forța statică echivalentă și analizele spectrului de răspuns. Aceste abordări folosesc adesea factori de comportare care reduc forțele seismice în timpul proiectării, acceptând că unele degradări structurale sunt necesare pentru a disipa energia seismică în exces. Deși acest principiu este încorporat în toate codurile de proiectare seismică, mulți beneficiari ai clădirilor s-ar putea să nu înțeleagă pe deplin implicațiile sale. Ca rezultat, aceștia ar putea fi induși în eroare atunci când structurile lor, deși proiectate să se comporte adecvat în timpul evenimentelor seismice, suferă daune semnificative și devin nelocuibile. În acest context, analiza dinamică neliniară, care oferă o reprezentare mai precisă a răspunuslui structural sub încărcare seismică, este esențială pentru o estimare realistă a degradărilor. Cu toate acestea, implementarea analizei dinamice neliniare necesită de obicei software avansat și costisitor, cum ar fi SAP2000, ETABS sau TEKLA. Deși există alternative open-source precum OpenSees, STERA 3D, Code Aster și WARP3D, aceste instrumente necesită o expertiză tehnică substanțială, limitându-le utilizarea în principal la specialiștii din mediul academic și de cercetare. În consecință, analiza neliniară rămâne un proces complex și consumator de resurse, limitând aplicarea sa practică în industria ingineriei mai largi.

Pentru a aborda aceste provocări, acest raport propune o platformă web DIGITERRA, un instrument inovator, ușor de utilizat pentru estimarea degradărilor seismice, care extinde accesul la tehnici analitice avansate. DIGITERRA utilizează un algoritm de învățare automată antrenat pe un set de date de 120.000 de simulări dinamice neliniare pentru o mare varietate de tipuri de clădiri. Instrumentul propus simplifică evaluarea degradărilor structurale, permițând utilizatorilor să introducă parametrii esențiali ai clădirii — cum ar fi dimensiunile în plan, înălțimea totală, dimensiunile stâlpilor și caracteristicile mișcărilor seismice selectate — și să primească o estimare a stării de degradare bazată pe o evaluare a indicelui de degradare. Ca instrument de tip *open-access*, DIGITERRA permite atât inginerilor, cât și nespecialiștilor să efectueze evaluări ale degradărilor seismice fără a fi nevoie de software specializat sau de pregătire tehnică extinsă. Prin punerea la dispoziția unui public mai larg a analizei neliniare, platforma se aliniază cu obiectivele internaționale de reziliență la dezastre și deschide calea pentru cercetări și aplicații practice mai ample.

Stadiul actual al cercetării

Inițiativele recente în acest domeniu de cercetare au încorporat din ce în ce mai mult tehnici de învățare automată. De exemplu, Fundația Global Earthquake Model a dezvoltat platforma OpenQuake, un instrument open-source care integrează învățarea automată pentru a prezice performanța seismică a clădirilor și a standardiza evaluările riscurilor la nivel global. În mod similar, sistemul Prompt Assessment of Global Earthquakes for Response (PAGER) al USGS utilizează date istorice pentru a estima rapid impactul cutremurelor semnificative, oferind informații critice pentru respondenții de urgență și agențiile guvernamentale. În același timp, la Universitatea din California, Berkeley, platforma SimCenter oferă instrumente avansate de modelare și simulare pentru evaluarea hazardelor naturale. Aceasta îmbunătățește capacitatea de a evalua impacturile dezastrelor naturale, folosind învățarea automată pentru a rafina tehnicile de modelare utilizând date experimentale și istorice. În plus, proiectul Next Generation Attenuation (NGA) de la Pacific Earthquake Engineering Research Center dezvoltă modele predictive pentru atenuarea mișcării seismice, îmbunătățind înțelegerea răspunsurilor structurale în timpul cutremurelor.

În ciuda importanței acestor inițiative de cercetare, ele se bazează în mare parte pe seturi de date extinse fără a se concentra pe tipuri specifice de clădiri. Prin urmare, este necesar să se dezvolte metodologii care să estimeze indicii de degradare pe baza caracteristicilor individuale ale clădirilor. Mai mult, crearea unei platforme accesibile atât pentru specialiști, cât și pentru nespecialiști în sectorul construcțiilor facilitează o implicare publică mai mare și o înțelegere a riscului seismic și a practicilor de construcție rezidențială.

Platforma web dezvoltată în acest studiu cuantifică degradările structurale seismice folosind binecunoscutul indice de degradare Park-Ang [12], cu versiunea sa updatată [6]. Această abordare oferă un cadru robust și larg acceptat pentru estimarea degradărilor structurale. Platforma se bazează pe un set de date cuprinzător de analize dinamice neliniare care au fost realizate folosind o metodologie numerică eficientă dezvoltată anterior de autori [9–11]. Bazându-se pe lucrările anterioare [1], această lucrare îmbunătățește acuratețea și extinde aplicabilitatea algoritmilor pentru o accesibilitate mai largă.

Setul de date utilizat pentru antrenarea algoritmilor de învățare automată

Așa cum s-a discutat anterior, platforma web dezvoltată utilizează algoritmi de învățare automată pentru a estima daunele structurale. Prin urmare, antrenarea și validarea acestor algoritmi necesită un set de date extins care să cuprindă modele structurale și răspunsul structural corespunzător obținut din simulări dinamice neliniare.

Pentru a genera acest set de date, în primul rând, a fost dezvoltat un algoritm automat de generare a modelelor numerice ale uor clădiri cu caracteristici aleatorii. Acest algoritm de generare a modelelor construiește modele numerice unice prin atribuirea aleatorie a valorilor pentru principalii parametri care descriu clădirile. Aceștia includ: dimensiunile totale ale planului, numărul de deschideri și travee, numărul de etaje, înălțimea etajului, capacitățile de încovoiere ale grinzilor și stâlpilor, încărcarea aplicată, modulul lui Young , factorul de scalare al acțiunilor seismice și coeficientul de amortizare al materialuluit. Desigur, acești parametri sunt selectați în mod inerent în intervale predefinite derivate din practica ingineriei structurale pentru a asigura modele structurale realiste și plauzibile. În plus, modelele numerice sunt construite folosind elemente finite de tip bară cu două noduri, fiecare având șase grade de libertate (trei translații și trei rotații). Matricele de masă și rigiditate ale elementelor sunt asamblate în matricele globale de masă și rigiditate pentru întreaga structură 3D folosind o matrice topologică. De asemenea, matricea de amortizare este calculată folosind metoda Rayleigh, proporțională cu matricele globale de masă și rigiditate.

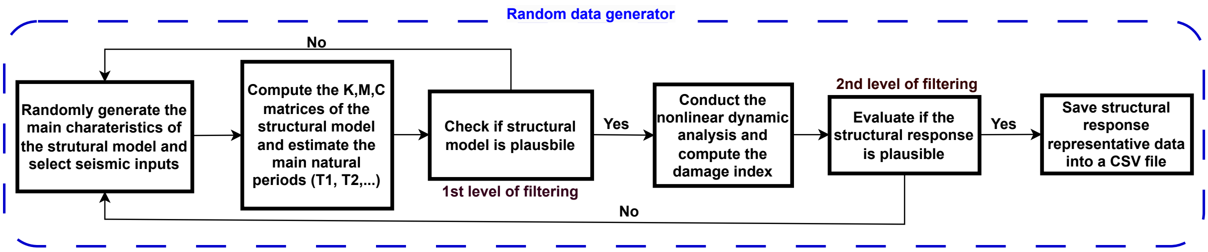


Figura 1: Principalele rutine pentru generarea datelor pentru algoritmul de învățare automată

După generarea modelului structural, se aplică un filtru inițial de plauzibilitate. Prin urmare, doar modelele cu o perioadă proprie a primului mod T_1 care se încadrează în intervalul $[0.07 \cdot no\_story, 0.2 \cdot no\_story]$ sunt acceptate ca plauzibile, unde no_story reprezintă numărul total de etaje. Dacă modelul este considerat acceptabil, se efectuează o simulare dinamică neliniară pentru a determina indicele de degradare; în caz contrar, se generează un nou model de clădire.

În continuare, simularea dinamică neliniară este realizată folosind un cod MATLAB dezvoltat personalizat care încorporează Formularea în Spațiul de Stări și Metoda Analogiei Forței (FAM) pentru a cuantifica răspunsul dinamic și deformațiile inelastice. Principalele etape ale procedurii de simulare sunt prezentate în Figura 1. Ideea centrală a FAM constă în augmentarea ecuației generale a mișcării cu un nou vector de necunoscute, $\boldsymbol{\theta}''$, care reprezintă rotațiile inelastice ale modelului structural, alături de o inegalitate care limitează momentele de încovoiere la capacitatea de încovoiere, așa cum este exprimat în Eq. 1. Această formulare asigură că matricea de rigiditate a structurii, calculată la începutul analizei, rămâne constantă pe parcursul calculelor inelastice. Este important de menționat că FAM utilizează plasticitate concentrată bazată pe articulații plastice pentru a modela deformațiile inelastice. În mod specific, în această lucrare, articulații plastice de moment încovoietor au fost utilizate pentru grinzi, în timp ce pentru stâlpi, interacțiunea dintre forța axială și încovoierea pe ambele direcții principale a fost considerată folosind un tip de articulație plastică cu suprafață de capacitate 3D ($n - m - m$). Prin combinarea FAM cu reprezentarea în spațiul de stări, această metodologie asigură o eficiență computațională ridicată și precizie în estimarea răspunsului dinamic neliniar al sistemelor structurale.

$$\mathbf{M} \cdot \ddot{\mathbf{u}}(t) + \mathbf{C} \cdot \dot{\mathbf{u}}(t) + \mathbf{K} \cdot \mathbf{u}(t) = -\mathbf{M} \cdot \mathbf{h} \cdot \ddot{\mathbf{a}}_g(t) + \mathbf{K}_p \cdot \boldsymbol{\theta}^{''}(t) \tag{1}$$

$$\mathbf{m}(t) \leq \mathbf{m}_{cap} \tag{2}$$

unde,

$$\mathbf{m}(t) = \mathbf{K}_P^T \cdot \mathbf{u}(t) - \mathbf{K}_R \cdot \boldsymbol{\theta}^{''}(t) \tag{3}$$

În primul rând, folosind ecuația de integrare explicită (Eq. 4), vectorul deplasărilor $\mathbf{u}_{k+1}$ este calculat la fiecare pas de timp pe baza matricelor modelului structural ($\mathbf{K}$,$\mathbf{M}$,$\mathbf{C}$), vectorulinput seismic $\ddot{\mathbf{a}}_g(t)$ și datele din pasul anterior. Ulterior, valorii de încercare (trial) sunt calculate pentru momentele de încovoiere $\mathbf{m}_t$ iar vectorul de incrementare a rotirii plastice este presupus $\boldsymbol{\Delta\theta}^{''} = \mathbf{0}$. În continuare, momentele de încovoiere de încercare sunt comparate cu capacitățile lor corespunzătoare de încovoiere. Dacă orice valoare de încercare depășește capacitățile de încovoiere, acestea sunt limitate la limita capacității, iar efortul de încovoiere în exces este redistribuit ca deformații plastice la în fiecare articulație plastică corespunzătoare, contribuind la valorile din vectorul $\boldsymbol{\Delta\theta}^{''}$. În plus, vectorul rotirilor plastice este calculat ca $\boldsymbol{\theta}^{''}_{k+1} = \boldsymbol{\theta}^{''}_k + \boldsymbol{\Delta\theta}^{''}$. Acest $\boldsymbol{\theta}^{''}_{k+1}$ actualizat este ulterior utilizat pentru a recalcula vectorul de momente de încovoiere $\mathbf{m}_t$ pentru a verifica dacă redistribuirea momentului de încovoiere are loc. Dacă nu se detectează nicio redistribuire, valorile de încercare ale $\boldsymbol{\Delta\theta}^{''}$ sunt confirmate ca valori finale, iar $\boldsymbol{\theta}^{''}_{k+1}$ este încorporat în schema de integrare pentru a calcula vectorul de deplasare pentru următorul pas de timp, $\mathbf{u}_{k+2}$. În caz contrar, $\mathbf{m}_t$ este trimis înapoi la secvența de limitare.

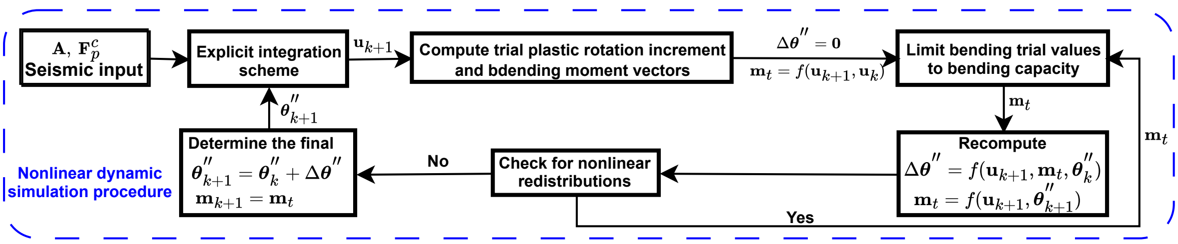


Figura 2: Schema de integrare explicită pentru FAM

Pe de altă parte, dacă valorile de încercare ale momentului de încovoiere rămân în limitele capacității, valorile inițiale de încercare sunt acceptate ca finale și sunt introduse direct în secvența de integrare explicită. În esență, în rutina numerică neliniară FAM, vectorul de rotație plastică $\boldsymbol{\theta}^{''}_{k+1}$ este determinat iterativ, cu valorile de încercare fiind rafinate sistematic până când condițiile de echilibru sunt satisfăcute.

$$\mathbf{z}_{k+1} = e^{\mathbf{A}\cdot\Delta t} \cdot \mathbf{z}_k + e^{\mathbf{A}\cdot\Delta t} \cdot \mathbf{H} \cdot \Delta t \cdot \ddot{\mathbf{a}}_{gk} + e^{\mathbf{A}\cdot\Delta t} \cdot \mathbf{F}_p^c \cdot \Delta t \cdot \boldsymbol{\theta}^{''}_k \tag{4}$$

unde

$$\mathbf{z}_{k+1} = \begin{bmatrix} \mathbf{u}_{k+1} \\ \dot{\mathbf{u}}_{k+1} \end{bmatrix} \ \mathbf{A} = \begin{bmatrix} \mathbf{O}_n & \mathbf{I}_n \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{C} \end{bmatrix}$$
$$\mathbf{z}_k = \begin{bmatrix} \mathbf{u}_k \\ \dot{\mathbf{u}}_k \end{bmatrix} \ \mathbf{F}_p^c = \begin{bmatrix} \mathbf{O}_n \\ -\mathbf{M}^{-1}\mathbf{K}_p \end{bmatrix} \tag{5}$$

De asemenea, este important de menționat că în această procedură numerică, încovoierea este considerată singura sursă de neliniaritate, în timp ce deformațiile din forfecare și axiale sunt presupuse a rămâne elastice. Această presupunere se aliniază cu principiul conform căruia în regiunile seismice, doar comportamentul inelastic ductil este acceptabil, în timp ce cedarea fragilă datorat forței tăietoare sau forței axiale nu este acceptată. Detalii suplimentare despre rutina de simulare pot fi găsite în lucrările anterioare ale autorilor [6, 9–11].

După ce răspunsul dinamic neliniar este simulat, indicele de daune Park și Ang este calculat [7,13] pe baza matricelor rezultate care conțin momentele de încovoiere și istoricul rotirilor plastice. În această fază, se aplică un criteriu secundar de filtrare: doar modelele cu valori ale indicelui de daune mai mici de 2.5 sunt considerate acceptabile. Dacă această valoare este depășită, rezultatele sunt eliminate. În final, un vector care conține parametrii de intrare care definesc modelul structural și rezultatele corespunzătoare ale simulării este generat. Aceste date sunt scrise într-un fișier CSV, fiecare rând reprezentând rezultatele unei singure simulări. Având în vedere cerințele substanțiale de stocare pentru păstrarea unui set complet de date de răspuns structural - estimat la aproximativ 1.5 GB per analiză - doar parametrii esențiali, cum ar fi indicele de daune, deplasarea maximă a ultimului etaj și raportul dintre energia histeretică și energia de intrare, au fost reținuți pentru a optimiza stocarea datelor. În plus, datele adunate în aceste fișiere CSV au fost utilizate pentru a antrena și testa algoritmii de învățare automată.

Trebuie menționat că acest studiu se concentrează pe două tipuri de clădiri: (1) clădiri regulate cu planuri rectangulare și (2) clădiri neregulate cu planuri în formă de L pe anumite etaje (Fig. 3) și că intrările seismice sunt selectate dintr-o bază de date de cutremure înregistrate în România din 1977 până în prezent.

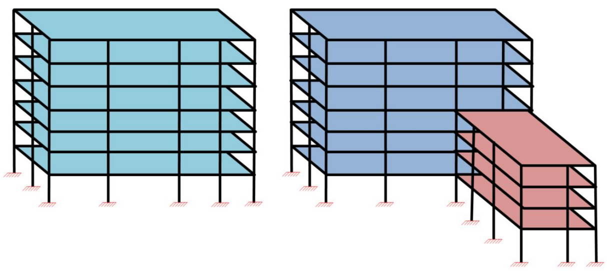


Figura 3: Tipuri de clădiri cu geometrie regulată și neregulată

Dezvoltarea modelului de învățare automată

Aplicarea învățării automate în ingineria structurală și seismică a crescut substanțial în ultimii ani. Datorită capacității lor de a procesa seturi mari de date, de a detecta tipare complexe și de a face predicții precise, algoritmii de învățare automată au devenit instrumente indispensabile în optimizarea performanței structurale, evaluarea riscurilor seismice și avansarea proiectării infrastructurii rezistente la cutremure.

Unul dintre cele mai semnificative studii recente [15] a aplicat tehnica de învățare automată Support Vector Regression pentru a estima raportul maxim de deplasare al clădirilor. Autorii au utilizat New Earthquake Data (NDE1.0), care constă în înregistrări reale de cutremure din 8520 de măsurători de mișcare puternică în 118 clădiri și 2737 de evenimente seismice. Constatările lor sugerează că modelul propus are un potențial promițător pentru evaluarea degradărilor seismice post-eveniment și poate îmbunătăți semnificativ capacitățile de răspuns de urgență în timp aproape real. Un alt studiu important [16] a utilizat algoritmul de învățare automată Random Forest pentru a evalua riscurile de construcție a podurilor. Autorii au demonstrat că acest algoritm permite predicții mai obiective și mai precise ale riscurilor prin identificarea eficientă a factorilor de risc critici, așa cum este ilustrat în analiza unui pod pietonal tipic. Într-un studiu de pionierat [18], modelele de învățare automată au fost utilizate pentru a prezice cutremurele în zona Los Angeles. Prin integrarea unor matrice de caracteristici cuprinzătoare cu algoritmi de învățare automată precum Random Forest, modelul a obținut o acuratețe ridicată în prezicerea categoriei maxime de cutremur pe o perioadă de 30 de zile. Rezultatele lor arată că acest model poate depăși metodele tradiționale și poate oferi o gestionare îmbunătățită a riscurilor seismice și strategii de pregătire mai eficiente. În plus, [17] a propus un algoritm inovator de control semi-activ bazat pe învățare automată pentru a reduce răspunsul structural în timpul evenimentelor seismice. Abordarea lor a încorporat evaluarea fiabilității seismice, permițând o evaluare mai cuprinzătoare a eficacității controlului pentru structurile de apeduct. Algoritmul a fost validat prin studii de caz, demonstrând o îmbunătățire semnificativă a fiabilității seismice comparativ cu metodele tradiționale.

Metodologia pentru dezvoltarea modelului de învățare automată prezentată în această lucrare este ilustrată în Fig. 4. Abordarea pune accent pe crearea unui model de estimare robust care să atingă un echilibru optim între acuratețe și simplitate.

În prima etapă a procedurii prezentate în Fig. 4, obiectivul este de a dezvolta un model capabil să estimeze cu acuratețe indicele de degradare prin luarea în considerare a unui set cuprinzător de caracteristici asociate unei clădiri și intrărilor seismice relevante, toate încapsulate într-un set de date adecvat. Un aspect important aici este echilibrarea numărului de caracteristici utilizate pentru estimare, ușurința obținerii acestor caracteristici pentru un utilizator mediu al platformei web și acuratețea generală a modelului. Acest echilibru este esențial pentru a asigura o platformă intuitivă și ușor de utilizat. Dacă setul de parametri necesar pentru estimare este dificil de obținut sau de gestionat pentru utilizatori, aceștia pot furniza valori inexacte sau pot fi descurajați să utilizeze platforma.

În urma generării setului de date prezentat în secțiunea anterioară, fiecare punct de date are următoarele componente:

- Caracteristici legate de accelerograma aleasă: Primele frecvențe naturale ale înregistrărilor - `Fund_Freq1` (direcția X) & - `Fund_Freq2` (direcția Y), PGA-ul înregistrării pe direcția X -`PGA_of_the_recording_scale_1`, PGA-ul înregistrării pe direcția Y - `PGA_of_the_recording_scale_2`.
- Caracteristici ale clădirii: Înălțimea clădirii - `b_Hieght`, Lungimea principală a clădirii - `dim_x`, Lățimea principală a clădirii - `dim_y`, Lățimea secțiunii coloanei - `b_st`, Înălțimea secțiunii coloanei - `h_st`, Numărul de etaje ale clădirii principale

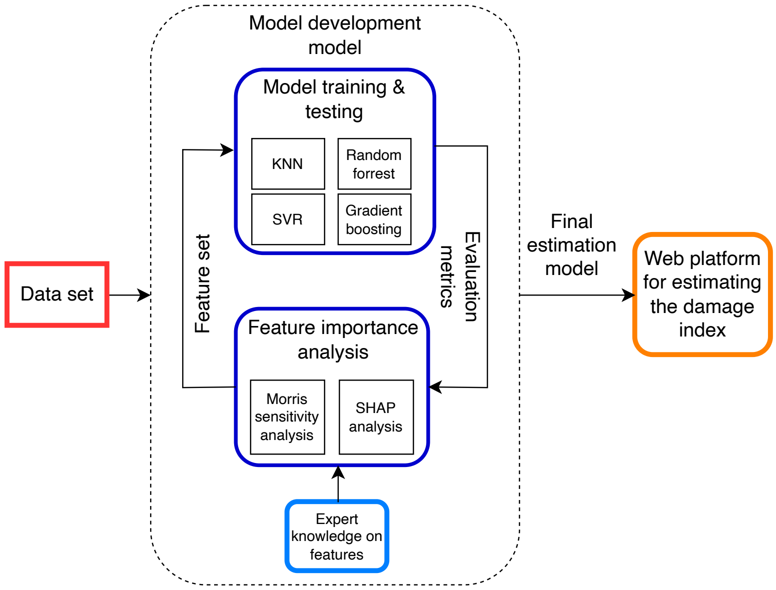


Figura 4: Metoda propusă pentru determinarea unui model de estimare optim

- `no_story`, Numărul de etaje ale părții neregulate a clădirii (cu roz în Fig.3)- `no_story2`, Înălțimea etajului - `Height`, Lățimea secțiunii grinzilor - `b_gr`, Înălțimea secțiunii grinzilor - `h_gr`, Modulul lui Young - `E`, Capacitatea de încovoiere a stâlpului pe direcția Y - `MstY`, Capacitatea de încovoiere a stâlpului pe direcția X - `MstX`, Capacitatea de încovoiere a grinzilor - `Mgr`, Dimensiunea deschiderii părții neregulate a clădirii (cu roz în Fig.3) - `bay2`, Numărul de travee ale părții neregulate a clădirii (cu roz în Fig.3) - `no_span2`, Numărul de deschideri ale părții neregulate a clădirii (cu roz în Fig.3)- `no_bay2`, Prima perioadă naturală de vibrație - `T1`, A doua perioadă naturală de vibrație - `T2`, A treia perioadă naturală de vibrație - `T3`, Tipul clădirii: geometrie regulată (0) sau neregulată (1)- `Lshape`

Variabila estimată este reprezentată de indicele de degradare. Această variabilă țintă, obținută cu modelul din MATLAB, servește drept adevăr de bază pentru metoda noastră de estimare. Deși încorporarea modelului MATLAB direct în platforma web este o opțiune, trebuie menționat că procesul de simulare este consumator de timp, iar parametrii de intrare necesari pentru a calcula indicele de degradare sunt extinși. Prin urmare, subliniem din nou necesitatea unui sistem de acces deschis destinat utilizatorilor non-experți în inginerie structurală, optimizând viteza de simulare și simplificând accesibilitatea parametrilor.

În ceea ce privește algoritmii de estimare utilizați, pornim de la un studiu anterior [1] și considerăm cel mai bun algoritm obținut acolo (un model K Nearest Neighbours antrenat cu hiperparametri optimizați) ca punct de plecare pentru identificarea unor alternative mai bune printre opțiuni precum Random Forest sau Gradient Boosting. Continuăm, de asemenea, acest studiu prin utilizarea a două metode de analiză a modelului black-box pentru a cuantifica mai bine importanța fiecărei caracteristici în procesul general de estimare: analiza de sensibilitate Morris și analiza caracteristicilor SHAP. În final, folosim aceste cunoștințe combinate cu expertiza în inginerie structurală pentru a selecta un model care să atingă un echilibru adecvat între acuratețea estimării, măsurată prin scorul R2, eroarea medie absolută și eroarea medie pătratică și simplitatea în ceea ce privește colectarea datelor caracteristicilor.

În final, cel mai bun model este adăugat pe platforma web pentru a putea fi utilizat în continuare în estimarea indicelui de daune pe baza caracteristicilor introduse de utilizatori. În cele din urmă, indicele de degradare estimat va fi generat într-o perioadă foarte scurtă de timp, împreună cu un raport și o interpretare a indicelui corespunzător clădirii date și cutremurului ales.

Algoritmi de Estimare

Având în vedere natura extrem de neliniară a impactului cutremurelor asupra clădirilor, următorii algoritmi sunt utilizați în procesul de estimare:

- **N-Nearest Neighbors (KNN)**: Acest algoritm prezice variabila țintă a unui punct de date prin analizarea celor mai similari k vecini din spațiul caracteristicilor. Dacă vorbim de un model de tip regresie, predicția este de obicei media sau mediana valorilor țintă ale vecinilor, în timp ce pentru o problemă de clasificare, este determinată de clasa majoritară. Asigurarea unui set de date bine distribuit și selectarea unei valori optime pentru k sunt critice pentru succesul algoritmului.

- **Random Forest**: Acest algoritm de învățare prin ansamblu construiește multiple arbori de decizie folosind subseturi aleatorii de date și caracteristici (bagging). Pentru regresie, predicția finală este media tuturor ieșirilor arborilor, în timp ce pentru clasificare, este determinată prin vot majoritar. Random Forest este robust la supraînvățare, gestionează eficient seturi mari de date și oferă predicții stabile.

- **Gradient Boosting**: Acest algoritm de învățare prin ansamblu construiește arbori de decizie secvențial, fiecare arbore corectând erorile predecesorilor săi. Prin minimizarea unei funcții de pierdere specifice (de exemplu, eroarea pătratică medie pentru regresie), algoritmul îmbunătățește iterativ predicțiile. Gradient Boosting combină ieșirile învățătorilor slabi pentru a forma un model puternic și precis, deși necesită regularizare atentă pentru a preveni supraînvățarea și este costisitor din punct de vedere computațional.

- **Support Vector Regression (SVR)**: SVR este un algoritm de regresie care prezice o variabilă țintă continuă prin găsirea unei funcții într-o marjă de toleranță specificată, ϵ. Utilizează funcții kernel, cum ar fi funcția linieară, polinomială sau funcția de bază radială (RBF), pentru a captura relațiile liniare și neliniare. Parametrii cheie includ ϵ (marja de toleranță), C (regularizarea) și setările specifice kernelului.

Algoritmii propuși au fost selectați pentru simplitatea și eficacitatea lor dovedită în sarcini similare de estimare [5]. Aceștia sunt bine adaptați pentru estimarea indicelui de daune, având în vedere dimensiunea setului de date.

Indicatori pentru evaluarea performanței

Performanța modelelor de estimare propuse este evaluată folosind următorii indicatori:

- **Coeficientul de Determinare** (R^2): Măsoară proporția variației în variabila țintă care este estimată din variabilele independente:

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

  unde y_i sunt valorile reale, $\hat{y}_i$ sunt valorile prezise și $\bar{y}$ este media valorilor reale.

- **Eroarea Medie Absolută (MAE)**: Reprezintă media diferențelor absolute dintre valorile estimate și cele reale:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Eroarea Pătratică Medie (MSE)**: Reprezintă media diferențelor pătratice dintre valorile estimate și cele reale:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Analiză Preliminară

Având în vedere cadrul de estimare definit anterior, putem acum efectua o analiză preliminară a performanței algoritmilor de estimare. Setul de date a fost împărțit în seturi de antrenament și testare cu un raport de 80-20%. Rezultatele sunt prezentate în Fig. 5.

În această figură, putem observa că cei mai buni algoritmi se dovedesc a fi algoritmul Random Forest și algoritmul Gradient Boosting, ambele având un scor R^2 foarte bun și un MAE și MSE scăzut. Este, de asemenea, important de remarcat că ambii algoritmi depășesc algoritmul KNN, ceea ce nu este surprinzător având în vedere natura neliniară a problemei. Aceasta este o îmbunătățire față de rezultatele obținute în studiul anterior [1], unde algoritmul KNN era cel mai performant.

Algoritm	**Scor R²**	**MAE**	**MSE**
K Nearest Neighbours	0.81	0.09	0.05
K Nearest Neighbours Optimizat	0.88	0.05	0.03
Gradient Boosting	0.73	0.17	0.07
Gradient Boosting Optimizat	0.97	0.04	0.01
Random Forest	0.96	0.04	0.01
Random Forest Optimizat	0.96	0.04	0.01
Support Vector Regression	0.51	0.22	0.12
Support Vector Regression Optimizat	0.51	0.22	0.12

	Cel mai bun rezultat
	Al doilea cel mai bun rezultat
	Al treilea cel mai bun rezultat

Tabela 1: Evaluarea algoritmilor de învățare automată

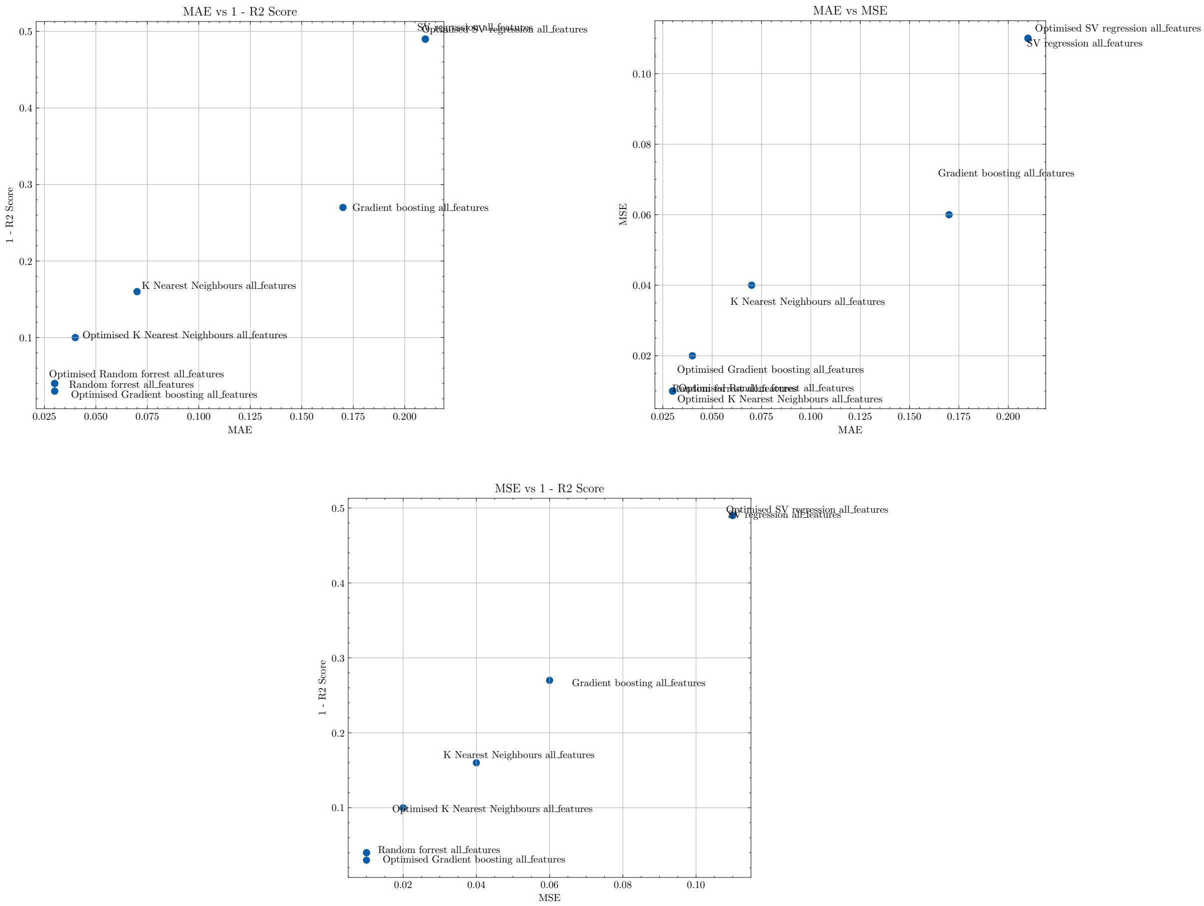


Figura 5: Analiza performanței algoritmilor de estimare

Selecția caracteristicilor

Pentru a dezvolta un model robust și accesibil pentru estimarea indicelui de degradare Park și Ang, s-a efectuat o analiză de selecție a caracteristicilor pentru a evalua dacă numărul de parametri de intrare ar putea fi minimizat fără a compromite acuratețea predictivă a modelului. Obiectivul a fost simplificarea modelului pentru aplicarea practică atât de către specialiști, cât și de nespecialiști, prioritizând parametrii care pot fi identificați în mod fiabil fie din documentația clădirii, fie prin estimări directe sau măsurători pe teren. Această concentrare pe intrări esențiale dar ușor de obținut este deosebit de benefică pentru evaluarea structurilor mai vechi, unde datele detaliate pot fi indisponibile. Pentru aceasta, folosim două metode: o analiză de sensibilitate și o analiză SHAP (SHapley Additive exPlanations).

Analiza de sensibilitate

Această analiză a început cu o analiză de senzitivitate, o metodă care cuantifică efectul fiecărei intrări asupra ieșirii modelului, clasificând caracteristicile în funcție de contribuția lor la acuratețea predictivă și robustețe. Această metodă este utilizată de obicei la evaluarea modelelor de tip *black-box*, unde importanța parametrilor individuali nu poate fi evaluată direct folosind metrici de performanță aplicate ieșirii.

În esență, analiza de senzitivitate implică aplicarea modelului la un set de parametri de intrare care variază de la o simulare la alta, ajustând câte un parametru pe rând. Traiectoria de-a lungul căreia variază parametrii este concepută pentru a acoperi eficient spațiul posibil de intrare. Modificarea rezultată în ieșire (efectul elementar) este caracterizată folosind două metrici cheie:

- **Media efectelor elementare** (μ) : indică importanța generală a unui parametru de intrare asupra variației ieșirii.
- **Deviația standard a efectelor elementare** (σ) : evidențiază impactul neliniar al unui parametru sau efectul interacțiunii sale cu alți parametri.

Astfel, pentru a identifica cele mai relevante caracteristici în ceea ce privește impactul lor asupra indicelui de degradare, am folosit modelul optimizat Gradient Boosting în cadrul analizei de sensibilitate. Am utilizat instrumentul implementat în biblioteca SALib pentru a genera 1000 de traiectorii peste spațiul posibil al variabilelor de intrare și pentru a evalua ulterior rezultatele. Limitele spațiului variabilelor de intrare au fost stabilite conform expertizei inginerești structurale care a condus la dezvoltarea procedurii de generare a clădirilor.

Caracteristicile utilizate în modelul de învățare automată sunt prezentate în Tabelul 2. În plus, limitele spațiului de intrare sunt, de asemenea, prezentate în tabel, oferind o mai bună înțelegere a spațiului de intrare pentru analiza de seinzitivitate.

Cod Caracteristică	**Nume Complet**	**Limite**
Fund_Freq1	Frecvența Fundamentală a Primei Intrări Seismice	[0.394, 3.794]
Fund_Freq2	Frecvența Fundamentală a Celei de-a Doua Intrări Seismice	[0.413, 3.774]
PGA of the recording scale_1	Accelerația Maximă la Sol a Primei Înregistrări Seismice	[0.717, 10.833]
PGA of the recording scale_2	Accelerația Maximă la Sol a Celei de-a Doua Înregistrări Seismice	[0.717, 11.872]
b_Height	Înălțimea Clădirii	[12, 52]
dim_x	Dimensiunea Clădirii pe Direcția X	[6, 35]
dim_y	Dimensiunea Clădirii pe Direcția Y	[6, 35]
b_st	Lățimea Stâlpilor de Etaj	[0.35, 1.4]
h_st	Înălțimea Stâlpilor de Etaj	[0.35, 1.4]
b_gr	Lățimea Stâlpilor de la Parter	[0.25, 0.35]
h_gr	Înălțimea Stâlpilor de la Parter	[0.2, 0.65]
E	Modulul de Elasticitate al Materialului	[13000000, 20000000]
MstY	Capacitatea de Încovoiere a Stâlpilor pe Direcția Y	[630, 2800]
MstX	Capacitatea de Încovoiere a Stâlpilor pe Direcția X	[535.5, 3220]
Mgr	Capacitatea de Încovoiere a grinzilor	[30.1, 509]
bay2	Lățimea Celei de-a Doua Travee	[5, 6]
no_span_2	Numărul de Deschideri în Secțiunea Neregulată	[1, 4]
no_bay_2	Numărul de Travee în Secțiunea Neregulată	[1, 5]
no_story_2	Numărul de Etaje în Secțiunea Neregulată	[2, 11]
T1	Prima Perioadă Naturală	[0.329, 2.540]
T2	A Doua Perioadă Naturală	[0.305, 2.330]
T3	A Treia Perioadă Naturală	[0.214, 1.794]
Lshape	Indicator pentru Configurația în Formă de L	[0, 1]

Tabela 2: Descrierea Caracteristicilor Utilizate în Modelul de Învățare Automată. Definirea limitelor pentru spațiul de intrare pentru analiza de sensibilitate.

Fig. 6 arată impactul fiecărui parametru în termeni de medie și deviație standard a efectelor elementare.

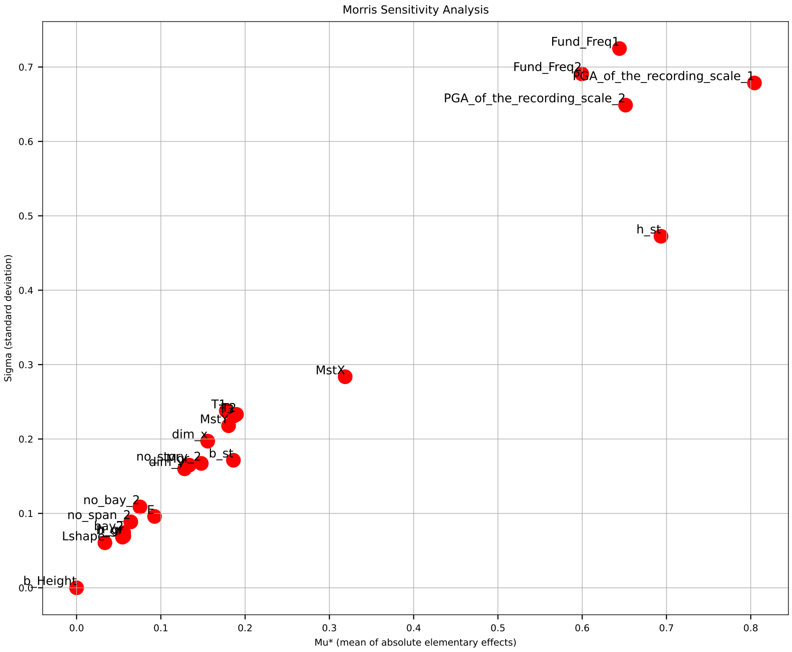


Figura 6: Rezultatele analizei de sensibilitate

Se poate observa în Fig. 6 că, în mod tipic, parametrii sunt împărțiți în două zone: colțul stânga jos, indicând un cluster de variabile care au puțină influență în estimarea indicelui de degradare, precum și colțul dreapta sus, unde observăm parametri care au o influență mare, nelineară asupra calculului variabilei țintă. Din această concluzie, putem vedea că frecvențele fundamentale și accelerațiile maxime ale miscării seismice au cel mai mare impact, un aspect care era într-un fel așteptat. Cu toate acestea, se observă că în clusterul parametrilor cu impact ridicat se află parametrul înălțimii etajului și, în mod notabil, capacitatea de încovoiere a stâlpilor *MstX* apare ușor în afara clusterului de importanță redusă. Aceste observații sugerează că acești parametri ar putea forma un subset potențial optimal, mai mic, pentru estimarea eficientă a indicelui de degradare a unei clădiri. În plus, mai mulți parametri au demonstrat un impact moderat dar notabil, îmbunătățind precizia modelului în predicția deteriorării structurale. Acestea includ primele trei perioade ale clădirii, capacitățile de încovoiere ale stâlpilor, dimensiunile generale ale clădirii, lățimea stâlpilor și numărul de etaje în secțiunile neregulate ale clădirii.

Analiza SHAP

În continuare, s-a efectuat și o analiză SHAP pentru a valida suplimentar parametrii cheie identificați prin analiza de senzitivitate Figura 7. Analiza SHAP, o metodă pentru interpretarea modelelor complexe de învățare automată, cuantifică contribuția fiecărei caracteristici de intrare la predicțiile modelului prin estimarea impactului fiecărei caracteristici asupra ieșirii individual și în combinație cu alte caracteristici. Această

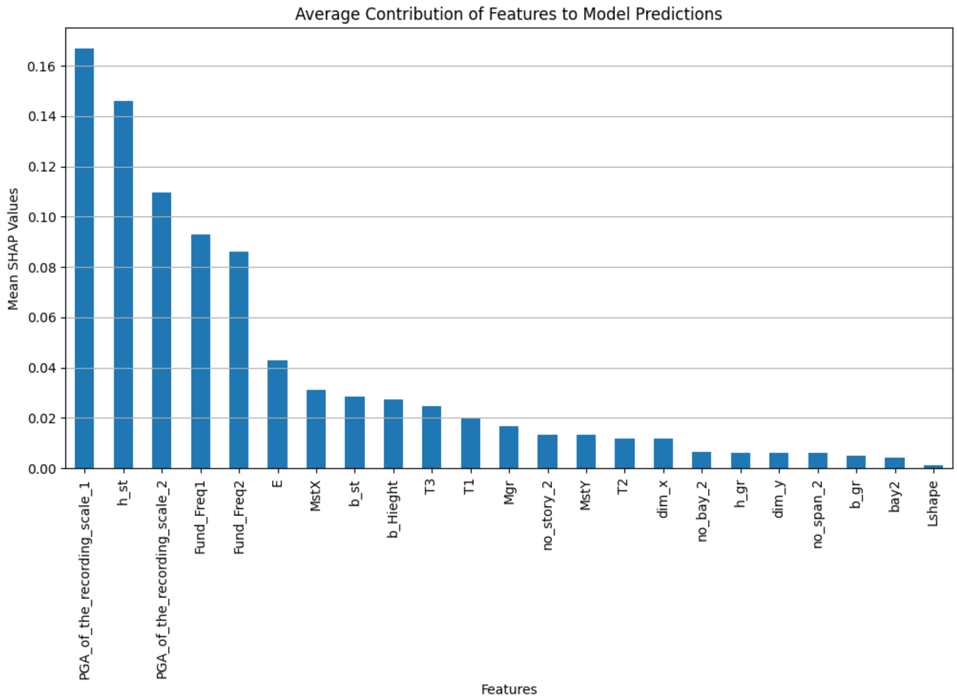


Figura 7: Contribuția Medie a Caracteristicilor la Predicțiile Modelului

abordare oferă informații atât despre direcția, cât și despre magnitudinea influenței fiecărui parametru, îmbunătățind interpretabilitatea modelului. Analiza SHAP a confirmat parametrii primari evidențiați de analiza de senzitivitate, inclusiv frecvențele mișcării seismice, accelerațiile maxime ale terenului și înălțimea stâlpilor. În plus, SHAP a identificat o caracteristică influentă suplimentară—modulul de elasticitate al materialului—care reflectă rigiditatea materialului și importanța sa în evaluarea răspunsului structural sub încărcare seismică. Parametrii cu impact moderat dar relevanți identificați de SHAP includ înălțimea totală a clădirii, perioadele sale naturale, numărul de etaje și travee în secțiunile neregulate, capacitățile de încovoiere ale stâlpilor, lățimea stâlpilor și dimensiunile generale ale clădirii pe direcția X. Această abordare cuprinzătoare nu doar validează setul inițial de caracteristici, dar evidențiază și parametri suplimentari critici pentru predicția precisă a indicelui de degradare Park și Ang în contexte seismice.

Setul final de caracteristici și concluzii

În consecință, pentru a testa acuratețea obținută cu caracteristicile selectate, algoritmul optimizat Gradient Boosting a fost antrenat folosind doar parametrii cei mai influenți identificați prin analizele de Sensibilitate și SHAP, și anume, frecvențele fundamentale ale intrării seismice, accelerațiile maxime, înălțimea stâlpilor și modulul de elasticitate. Folosind aceste intrări, modelul a atins metrici de performanță foarte bune, cu rezultate de test arătând o Eroare Medie Absolută (MAE) = 0.05, o Eroare Medie Pătratică (MSE) = 0.01 și un R^2 = 0.96. Aceste valori confirmă acuratețea și fiabilitatea ridicată a modelului când este aplicat în domeniul său de aplicabilitate. Cu toate acestea, din perspectivă inginerească, modelul presupune că utilizatorii furnizează valori de intrare realiste și structural adecvate pentru parametri precum înălțimea stâlpilor și modulul de elasticitate. Aceste intrări ar trebui să fie consistente cu proporțiile tipice rezultate din clădiri proiectate corespunzător din punct de vedere seismic. De exemplu, o structură în cadre cu 8 etaje ar avea în general înălțimi ale stâlpilor între 0.85 și 1.15 metri și un modul de elasticitate depășind 33 GPa. Astfel, în timp ce modelul demonstrează performanță predictivă excepțională, fiabilitatea sa în aplicații practice depinde de calitatea și validitatea datelor de intrare. Structurile care deviază semnificativ de la standardele de proiectare seismică sau proporțiile structurale realiste pot produce rezultate nereliabile. Prin urmare, prin restrângerea intervalului de intrare la caracteristici esențiale, măsurabile, modelul echilibrează acuratețea cu utilizabilitatea, dar datele de intrare trebuie să fie aliniate cu principiile inginerești reale.

Având în vedere că modelul de învățare automată este destinat implementării într-o aplicație web accesibilă atât specialiștilor, cât și nespecialiștilor, au fost necesare considerații suplimentare. În consecință, algoritmul a fost antrenat folosind caracteristicile cu impact moderat dar încă relevant, în plus față de caracteristicile primare identificate anterior. Acest set extins de caracteristici a rezultat într-o ușoară creștere a acurateței, cu metrici de performanță de Eroare Medie Absolută (MAE) = 0.03, Eroare Medie Pătratică (MSE) = 0.01 și $R^2 = 0.97$. Cu toate acestea, includerea anumitor parametri specializați, precum capacitățile de încovoiere ale stâlpilor, prezintă provocări practice. Aceste valori sunt rareori documentate explicit, chiar și în înregistrările detaliate ale clădirilor, făcându-le dificil de obținut. Similar, parametri precum perioadele naturale ale clădirii și modulul de elasticitate pot fi dificil de identificat din documentația tehnică pentru nespecialiști. Pentru a aborda aceste provocări, au fost luate în considerare aproximări și metode practice. De exemplu, prima și a doua perioadă naturală pot fi estimate rezonabil ca $T_1 = 0.1 \cdot (\text{numărul de etaje})$ și $T_2 = 0.8 \cdot T_1$. Modulul de elasticitate, deși mai puțin direct disponibil, poate fi dedus din valori standard sau măsurat folosind dispozitive portabile. În ciuda acestor soluții, capacitățile de încovoiere și a treia perioadă naturală au fost excluse din setul de caracteristici din cauza accesibilității lor limitate și impactului neglijabil asupra acurateței generale. Această abordare asigură că modelul rămâne atât precis, cât și practic pentru utilizare largă de către diverse tipuri de utilizatori.

În consecință, setul final de caracteristici este prezentat în Tabelul 3.

Set de Caracteristici	**R^2**	**MAE**	**MSE**
Set Complet: Fund_Freq1, Fund_Freq2, PGA_scale_1, PGA_scale_2, b_Height, dim_x, dim_y, b_st, h_st, b_gr, h_gr, E, MstY, MstX, Mgr, bay2, no_span_2, no_bay_2, no_story_2, T1, T2, T3, Lshape	0.97	0.03	0.01
Caracteristici Esențiale: Fund_Freq1, Fund_Freq2, PGA_scale_1, PGA_scale_2, h_st, E	0.96	0.05	0.01
Set Optimizat: Fund_Freq1, Fund_Freq2, PGA_scale_1, PGA_scale_2, b_Height, dim_x, dim_y, h_st, b_st, E, no_bay_2, no_story_2, T1, T2	0.97	0.03	0.01

Tabela 3: Comparația Performanței Diferitelor Seturi de Caracteristici în Modelul Gradient Boosting. Setul de date folosit pentru comparația performanței este setul de test.

Setul complet de caracteristici (23 caracteristici) a atins performanță excelentă pe setul de test cu $R^2 = 0.97$, MAE $= 0.03$ și MSE $= 0.01$. Cu toate acestea, acest set include parametri care sunt dificil de obținut în practică, precum capacitățile de încovoiere și a treia perioadă naturală.

Setul de caracteristici esențiale, identificat prin analizele de sensibilitate și SHAP, menține putere predictivă puternică ($R^2 = 0.96$) folosind doar șase caracteristici cheie. Aceasta reprezintă o configurație minimală care încă oferă estimări fiabile.

Setul optimizat de caracteristici (14 caracteristici) atinge aceeași performanță ridicată ca setul complet ($R^2 = 0.97$) excluzând în același timp parametrii cei mai dificil de obținut. Acest set reprezintă configurația noastră recomandată, echilibrând acuratețea cu utilizabilitatea practică.

Aplicație web

Arhitectura software

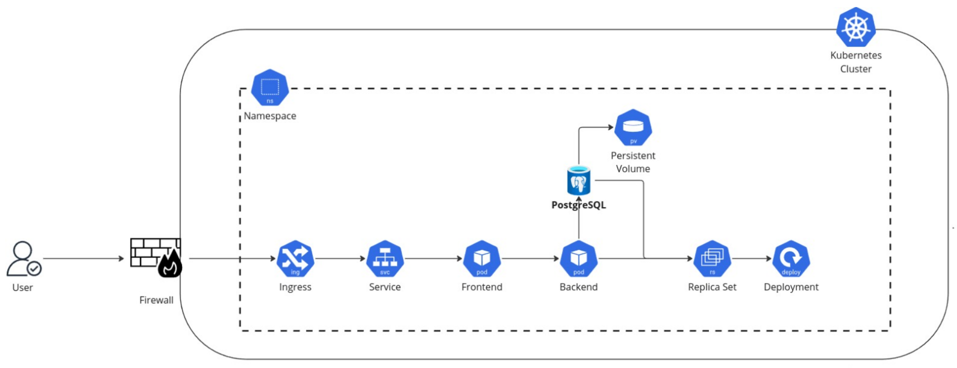


Figura 8: Arhitectura software

Fig. 8 prezintă arhitectura software a aplicației web concepute pentru determinarea indicelui de degradare structurală a unei clădiri supuse acțiunilor seismice. Această aplicație este dezvoltată într-un *cluster Kubernetes*, care oferă caracteristici cheie precum reziliență, scalabilitate automată, orchestrare de containere, disponibilitate ridicată și gestionarea stocării persistente.

Funcționalitatea aplicației este împărțită în două părți principale: backend și frontend.

Backend-ul gestionează funcționalitatea de bază a întregii aplicații, începând cu managementul caracteristicilor clădirilor, continuând cu stocarea și procesarea accelerogramelor și culminând cu estimarea comportamentului unei clădiri sub acțiuni seismice folosind algoritmul de învățare automată selectat.

Frontend-ul găzduiește interfețele prin care aplicația interacționează cu utilizatorii, incluzând formulare pentru introducerea parametrilor clădirilor și configurarea simulărilor, pagini pentru vizualizarea clădirilor, accelerogramelor și rezultatelor simulărilor, precum și o hartă care afișează clădirile simulate împreună cu valorile indicelui lor de degradare și interpretările corespunzătoare.

Scalabilitatea și reziliența aplicației sunt asigurate în cadrul clusterului Kubernetes de către componenta *Replica Set*. Această componentă este responsabilă pentru crearea unei noi instanțe al frontend-ului sau backend-ului ori de câte ori una dintre instanțele existente eșuează, menținând astfel numărul specificat de replici. Cu alte cuvinte, rolul său principal este de a implementa redundanța pentru instanțele frontend și backend, asigurând disponibilitatea continuă și toleranța la defecte.

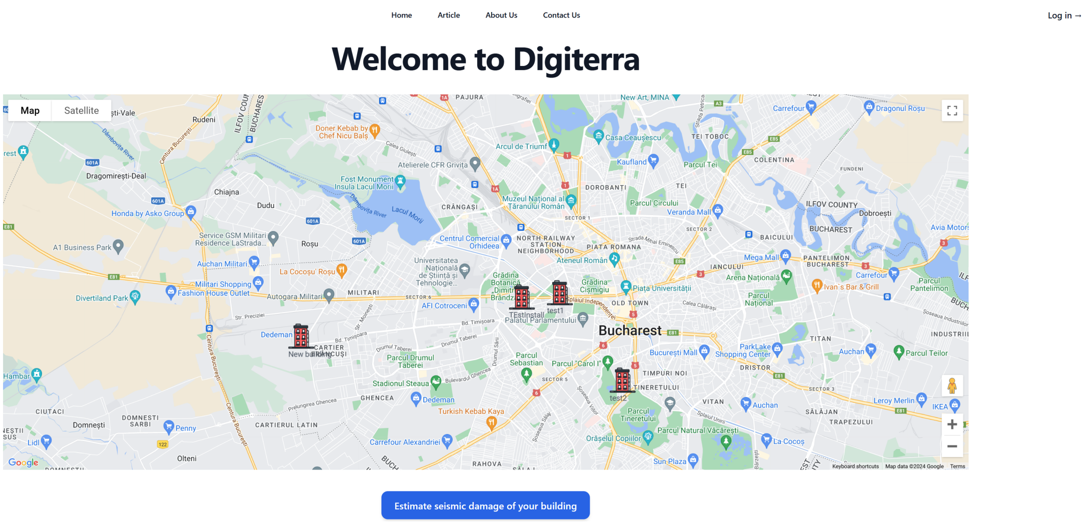
Pentru a gestiona scenariile cu trafic ridicat de utilizatori, componenta *Service (SVC)* acționează ca un distribuitor intern de sarcină, distribuind eficient traficul între instanțele replicate de frontend și backend pentru a asigura performanță și fiabilitate optimă. În timpul acestui proces, fiecărei instanțe îi este atribuit un nume unic (alias) distinct de cel expus pe internet. Interfața *Ingress* gestionează aceste nume și le mapează la adrese accesibile public, facilitând accesul extern fără probleme la aplicație.

Clădirile și datele simulărilor sunt stocate într-o bază de date *PostgreSQL*, accesibilă tuturor instanțelor backend. Această bază de date rezidă într-un *volum persistent*, asigurând că datele rămân disponibile pentru instanțele backend chiar și în timpul scalării sau redeployării.

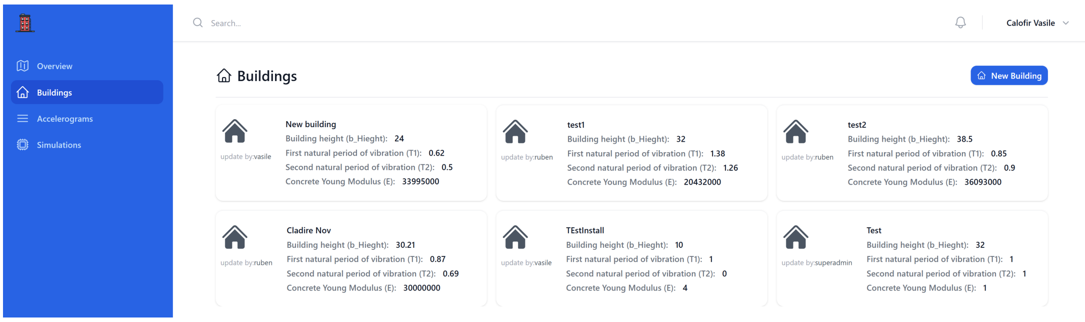
Componenta *Deployment* gestionează starea dorită a aplicației prin automatizarea proceselor de creare, actualizare și ștergere a pod-urilor, gestionarea scalabilității aplicației prin adăugarea sau eliminarea replicilor instanțelor frontend și backend și implementarea noilor versiuni ale aplicației fără întreruperea funcționalității acesteia (actualizări progresive). În plus, asigură că aplicația rămâne aliniată cu starea sa dorită chiar și în fața eșecurilor neașteptate.

În final, componenta *Firewall* protejează aplicația prin prevenirea accesului neautorizat al utilizatorilor, asigurând un mediu sigur pentru date și servicii.

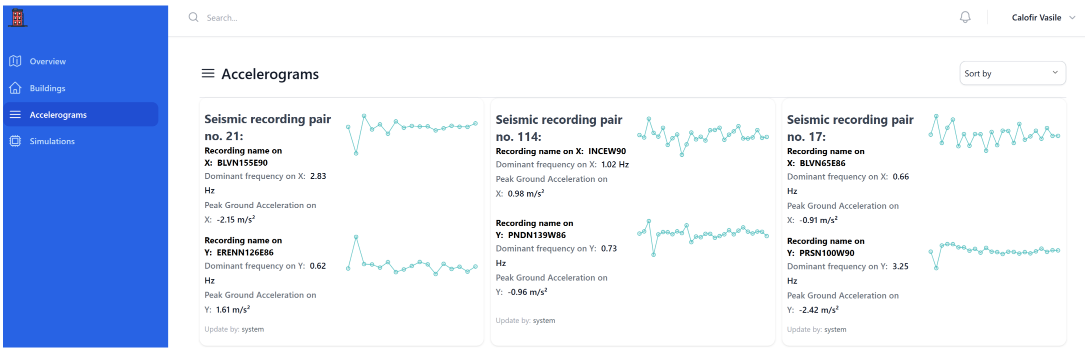
Aplicația este disponibilă pentru utilizare publică [2] și câteva capturi de ecran sunt prezentate în Fig. 9. De exemplu, în Fig. 9a putem vedea pagina de prezentare generală a aplicației, unde utilizatorii pot observa harta cu clădirile simulate și valorile corespunzătoare ale indicelui de degradare. În Fig. 9b putem vedea pagina clădirilor, unde utilizatorii pot vedea lista clădirilor și unele informații despre acestea. Aici, utilizatorii pot proiecta propriile clădiri și le pot simula prin includerea parametrilor doriți. Parametrii sunt validați intern de aplicația web în raport cu anumite limite predefinite. În Fig. 9c putem vedea pagina accelerogramelor, unde utilizatorii pot vedea lista accelerogramelor și unele informații despre acestea, împreună cu un profil al accelerogramei selectate. Utilizatorii nu pot, totuși, să încarce propriile accelerograme încă, deoarece accelerograma reprezintă informații sensibile care trebuie gestionate cu atenție. Codul sursă al aplicației este disponibil pe Zenodo [3].



(a) Pagina de prezentare generală - harta cu clădirile simulate și valorile corespunzătoare ale indicelui de degradare



(b) Pagina clădirilor - lista clădirilor și informații despre acestea



(c) Pagina accelerogramelor - lista accelerogramelor și informații despre acestea

Figura 9: Aplicația web DIGITERRA

Concluzii

În acest raport, am prezentat contribuția noastră în dezvoltarea unei soluții complete pentru evaluarea degradării structurale induse de cutremur utilizând tehnici de învățare automată.

În primul rând, lucrarea propune o metodologie robustă pentru selecția caracteristicilor și optimizarea modelului, identificând parametrii cei mai influenți pentru predicția degradării structurale. Frecvențele fundamentale ale cutremurului , accelerațiile maxime, înălțimea stâlpilor și modulul de elasticitate s-au dovedit ca parametrii de bazaă care permit estimări precise cu un scor R^2 de 0,96.

În al doilea rând, studiul a abordat provocările practice ale implementării învățării automate în ingineria structurală prin echilibrarea acurateței modelului cu utilizabilitatea în lumea reală. Prin considerarea accesibilității parametrilor de intrare și furnizarea unei estimări precise a indicelui de degradare, cercetarea asigură că instrumentul propus rămâne valoros atât pentru specialiști, cât și pentru nespecialiști. Modelul final, încorporând atât caracteristici primare, cât și caracteristici cu influență moderată, a atins o valoare R^2 îmbunătățită de 0,97, menținând în același timp aplicabilitatea practică.

În cele din urmă, cercetarea a culminat cu dezvoltarea unei aplicații web scalabile și reziliente, folosind principii moderne de arhitectură software prin containerizare Kubernetes. Această implementare asigură disponibilitate ridicată, scalabilitate automată și acces securizat la instrumentul de predicție, valorificând acest instrument pentru comunitatea inginerească. Arhitectura aplicației relevă o legătură eficientă între algoritmii complecși de învățare automată și nevoile practice de inginerie, oferind o interfață accesibilă pentru evaluarea degradării structurale.

Limitările implementării actuale se referă în principal doar la anumite tipuri specifice de clădiri (modele structurale în cadre regulate și în formă de L). În plus, au fost folosite doar înregistrări seismice din România pentru antrenarea modelelor de învățare automată, care ar putea să nu fie complet reprezentative pentru alte regiuni seismice. În lucrările viitoare, ne propunem să includem diferite sisteme structurale dincolo de cadrele regulate, iar procesul de selecție a caracteristicilor ar putea fi adaptat pentru alte tipuri de analiză structurală.

Bibliografie

[1] Vasile Calofir, Ruben-Iacob Munteanu, Mircea-Stefan Simoiu, and Karol-Cristian Lemnaru. Innovative approach to estimate structural damage using linear regression and K-nearest neighbors machine learning algorithms. *Results in Engineering*, 22:102250, June 2024.

[2] Vasile Calofir, Mircea Stefan Simoiu, and Ruben Munteanu. DIGITERRA. `https://digiterra.acs.pub.ro/`.

[3] Vasile Calofir, Mircea Stefan Simoiu, and Ruben Munteanu. DIGITERRA - Digital platform for estimating the building degradation index with machine learning techniques, December 2024. 10.5281/ZENODO.14254187.

[4] Qingle Cheng, Aiqun Li, Haotian Ren, Chea Por, Wenjie Liao, and Linlin Xie. Rapid seismic-damage assessment method for buildings on a regional scale based on spectrum-compatible data augmentation and deep learning. *Soil Dynamics and Earthquake Engineering*, 01 2024.

[5] Claudia Chiţu, Grigore Stamatescu, and Alberto Cerpa. Building occupancy estimation using supervised learning techniques. In *2019 23rd International Conference on System Theory, Control and Computing (ICSTCC)*, pages 167–172, 2019.

[6] S. K. Kunnath, a. M. Reinhorn, and R. F. Lobo. *IDARC Version 3.0: A Program for the Inelastic Damage Analysis of Reinforced Concrete Structures. Technical Report NCEER-92-0022.* U.S. National Center for Earthquake Engineering Research, 1992.

[7] S. K. Kunnath, a. M. Reinhorn, and R. F. Lobo. *IDARC Version 3.0: A Program for the Inelastic Damage Analysis of Reinforced Concrete Structures. Technical Report NCEER-92-0022.* U.S. National Center for Earthquake Engineering Research, 1992.

[8] Xuchuan Lin, Xueyan Liu, Jiang Hui, and Wenchen Shan. Assessment on detailed regional seismic damage risk of buildings based on time-history dynamic analyses. *Bulletin of Earthquake Engineering*, 22:1–21, 03 2024.

[9] Munteanu Ruben Iacob, Florin Moţa, Vasile Calofir, and Cătălin Baciu. New approach to nonlinear dynamic analysis of reinforced concrete 3d frames; an accurate and computational efficient mathematical model. *Applied Sciences*, 12(3), 2022.

[10] Munteanu Ruben Iacob, Enache Ruxandra, Baciu Cătălin, and Calofir Vasile. A new perspective into torsional inelastic response of actively controlled irregular multistorey buildings. *Alexandria Engineering Journal*, 71:691–706, 2023.

[11] George Bogdan Nica, Munteanu Ruben Iacob, Vasile Calofir, and Mihail Iancovici. Modelling nonlinear behavior of 3d frames using the force analogy method. *Structures*, 35:1162–1174, 2022.

[12] Young-Ji Park and Alfredo H.-S. Ang. Mechanistic seismic damage model for reinforced concrete. *Journal of Structural Engineering*, 111, 04 1985.

[13] Young-Ji Park and Alfredo H.-S. Ang. Mechanistic seismic damage model for reinforced concrete. *Journal of Structural Engineering*, 111, 04 1985.

[14] Xiaoyan Song, Xiaowei Cheng, Yi Li, Guo Ruijie, Haoyou Zhang, Zihan Liang, and Senna Wang. A numerical model database for rapid seismic damage assessment of typical regular reinforced concrete frame structures in urban building clusters. *Journal of Building Engineering*, 90:109392, 08 2024.

[15] Dongwang Tao, Shizhe Fang, Haixu Liu, Jianqi Lu, Jiang Wang, and Qiang Ma. Support vector regression model for the prediction of buildings' maximum seismic response based on real monitoring data. *Scientific Reports*, 14(1):29874, 2024.

[16] Ying Wu, Yigang Wang, Hongbing Liu, Liping Xie, Lili Jiao, and Pengzhen Lu. Risk assessment of bridge construction investigated using random forest algorithm. *Scientific Reports*, 14(1):20964, 2024.

[17] Zhongyuan Xiao, Jianguo Xu, Li Wang, and Liang Huang. Research on intelligent semi-active control algorithms and seismic reliability based on machine learning. *Scientific Reports*, 14(1):29487, 2024.

[18] Cemil Emre Yavas, Lei Chen, Christopher Kadlec, and Yiming Ji. Improving earthquake prediction accuracy in los angeles with machine learning. *Scientific Reports*, 14(1):24440, 2024.

[19] M. Zuher, Ade Nasution, Zairah Sidiq, Masrilayanti Masrilayanti, and Jafril Tanjung. Fragility assesment of mid-rise rc building using hazus method in high seismic zone. *Jurnal Bangunan: Konstruksi & Desain*, 1:79–89, 08 2023.

Results in Engineering 22 (2024) 102250

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Research paper

Innovative approach to estimate structural damage using linear regression and K-nearest neighbors machine learning algorithms

Vasile Calofir [a], Ruben-Iacob Munteanu [a], Mircea-Stefan Simoiu [a,*], Karol-Cristian Lemnaru [b]

[a] Department of Automatic Control and Industrial Informatics, Politehnica University of Bucharest, Bucharest, RO 060042, Romania
[b] Department of Electrotechnics and Measurements, Tehnical University of Cluj-Napoca, Cluj-Napoca, RO 400114, Romania

ARTICLE INFO

Keywords:
Machine learning algorithms
Nonlinear dynamic analysis
Seismic structural damage
Time efficient seismic simulations

ABSTRACT

Conventional structural design methodologies often utilize elastic analysis techniques, such as the equivalent static force method and the response spectrum method. While these methods are known for their simplicity and computational efficiency, they prove inadequate in capturing the extent of structural damage caused by seismic forces. Additionally, employing nonlinear dynamic analysis to estimate structural damage represents a challenging and intricate task, posing difficulties for many structural designers. Consequently, the objective of this paper is to present an innovative methodology for evaluating seismic structural damage of moment-resisting frame structures. This involves the utilization of machine learning algorithms, which have been trained and tested on a large data set generated using a newly developed and numerically efficient simulation procedure. The machine learning algorithms employ both linear regression and K-nearest neighbors approaches to accurately replicate the Park-Ang structural damage index.

1. Introduction and scope

The importance of designing resilient buildings in regions prone to seismic activities cannot be overstated. Such structures must be designed to do more than just withstand the immediate impact of earthquakes; they must also ensure minimized structural damage, guarantee continuous functionality and facilitate quick recovery, in order to reduce both human and economic losses.

A key component in achieving seismic resilience of structures is the use of nonlinear dynamic analysis. These type simulations offer a more accurate estimation of how structures will respond to seismic forces, taking into account the complexities of real-world conditions. Unlike linear analysis, nonlinear dynamic simulations are able to capture the inelastic behavior of the structural elements, providing a comprehensive view of the expected structural damage.

Traditional structural design strategies often employ elastic analysis methods like the equivalent static force approach and the response spectrum method. While these methods are straightforward and computationally less demanding, they can fall short in capturing the complex behavior of structures under seismic forces. Furthermore, although nonlinear dynamic analysis can prove more accurate, it turns out to be a very complex and intricate task that poses challenges for most structural designers.

This is where the power of machine learning algorithms becomes evident. Trained on comprehensive data sets that include damage assessments for a diverse range of structural models, these algorithms can learn to accurately predict structural damage. This not only simplifies the task of conducting accurate structural analysis but also provides invaluable insights that can significantly aid structural engineers and companies in designing more resilient buildings.

The main objective of this paper is to showcase an innovative approach for estimating structural damage, using machine learning algorithms trained on data generated by a numerically efficient simulation routine previously developed by the authors [1] [2]. The machine learning algorithm is particularly noteworthy for its ability to accurately replicate the level of structural damage, requiring only a limited set of input parameters related to the structural model and seismic excitation. From a practical perspective, the aim of this paper aim to showcase a tool dedicated especially for practitioners in the civil engineering domain for estimating structural damage of regular and irregular moment resisting frame structures. Comparing to the complexity and the computation effort emphasised by the traditional nonlinear dynamic analysis,

* Corresponding author.
E-mail address: mircea_stefan.simoiu@upb.ro (M.-S. Simoiu).

https://doi.org/10.1016/j.rineng.2024.102250
Received 22 January 2024; Received in revised form 26 April 2024; Accepted 8 May 2024
Available online 14 May 2024

this methodology offers a simplified, reliable approach. For example, one can use the tool to estimate the seismic damage of a given structure (quantified through the Park & Ang damage index) by providing some specific characteristics of the building such as the dimensions of the structural elements or the story height and a particular earthquake scenario. Using this data, the proposed machine learning approach can offer a reliable estimation of the structural damage level.

2. State of the art review

In recent years, the field of structural dynamics has undergone a revolutionary changes, due to the advent of machine learning technologies. Once dominated by traditional computational models and analytical methods, the field is now increasingly enriched by machine learning algorithms [3] capable of handling complex, nonlinear systems with excellent accuracy and efficiency.

At first, several machine learning applications have focused on assessing the fragility and robustness of structures under various forces, an important aim in structural engineering. For example, in [4], the authors use data from masonry walls to determine the fragility curves using learning algorithms. Another example concerning the resistance of steel tube columns is presented in [5], where the authors estimate internal parameters of steel tubes while adopting a distinct machine learning approach instead of the traditional analytical one. In [6], the authors describe a method for estimating the natural period of reinforced concrete structures with infill walls using machine learning algorithms. These examples illustrate interesting and novel approaches for using machine learning in structural engineering, while also concluding that this type of tool can reduce time and resources in comparison with traditional structural engineering analysis.

Regarding seismic structural damage estimation, one of the first studies that can be found in the literature [7] presents a general method for predicting seismic-induced damage using Artificial Neural Networks (ANNs). A set 2200 $2D$ reinforced concrete (RC) frames that varied in topology, stiffness, strength and damping were use for training ANNs while 619 models were employed for the testing phase. The dynamic structural responses were simulated using nonlinear FEM analysis while the Park and Ang indices were used to estimate the extent of structural damage. Consequently, using an ANN, a mapping between the structural and ground motion properties and the damage indices was established. The results showed that the ANN was able to accurately predict the level of damage. Another important paper [8], aimed at developing an artificial intelligence-based method to be used in analyzing the seismic performances of 3D reinforced concrete (RC) buildings in Turkey. A total of 66 RC structures with 4–10 storeys, representing the typical RC buildings in Turkey, were selected for the study. The level of seismic performance of these buildings was determined using elastic analysis and the 4-grade performance levels specified in Turkish Earthquake Code-2007 ($TEC-2007$). Thus, an ANN was trained to replicate the performance level of each building using as input 19 parameters describing the earthquake motion and the building characteristics. The results show that the algorithm can reach up to 80% accuracy. Another approach to investigate if the ANNs can be effectively used to predict the seismic damage state of RC buildings was proposed by [9]. Their study treats the problem both as function approximation and pattern recognition, using Multilayer Feedforward Perceptron networks. The training data consisted of 30 RC buildings subjected to 65 real-world ground motions while the structural damage is assessed conducting nonlinear time history analyses and based on maximum interstorey drift ratio. The study also explores the impact of various ANN configuration parameters on prediction reliability and tests the model's generalization capabilities. The most significant conclusion of the paper is that the ANNs can reliably estimate the seismic damage state of RC buildings after an earthquake. Moreover, these authors also investigated the number and the combination of seismic parameters through which an optimum prediction for the damage state of RC buildings can be achieved [10]. The main conclusion was that ANNs can adequately predict seismic damage if at least 5 seismic parameters are used as inputs. In a more recent paper, [11] investigates the potential of using an alternative method to estimate seismic damage in masonry buildings conducting a comparative analysis of two different approaches: vulnerability index method and an innovative approach based on the use of ANNs. The analytical results obtained through the referred strategies was compared to real data sets of structural damage collected after the 1998 Azores earthquake. The results show that the ANN was able to provide much better estimations than those achieved with the original vulnerability approach. At the same time, a new analytical expression for damage estimation was derived from the ANN results.

While these studies are highly complex, there are some issues that have not been addressed yet. These include the use of more expansive data sets that capture a broader spectrum of structural models having both plan and elevation irregularities, while also considering in the nonlinear time history analyses accelerograms from a specific geographic area in order to account for the unique soil characteristics of that region. Moreover, using more detailed data sets, an interesting aspect to investigate is what would be an adequate feature set for estimating the damage index and whether simpler models such as linear regression or K-nearest neighbors (KNN) can be used efficiently used in such a problem, considering the aforementioned feature sets. Additionally, several review studies [12,13] indeed indicate that data requirements continue to be a significant limitation for machine learning techniques in structural engineering. These reviews also emphasize the utility of such techniques in reducing risks and resource expenditure during the design stages of structures.

This paper seeks to carry on the research efforts made so far by introducing a new perspective on how machine learning techniques can be used to estimate structural damage of reinforce concrete moment resisting frame structures. The authors' main contribution is the creation of a complex numerical tool, enabling researchers and structural engineers to develop highly accurate machine learning-based routines for estimating structural damage of buildings subjected to seismic motions from specific geographic areas. The key features of the numerical tool are the use of a very fast nonlinear dynamic response simulation procedure and the implementation of an efficient machine learning technique that considers both a feature selection procedure and a machine learning algorithm hyper-parameter optimization.

3. Proposed method

As stated before, the main objective o the paper is to provide a numerical tool that facilitates the estimation of structural damage using machine learning based algorithms. This tool is composed of two main parts. Firstly, a numerical routine was developed to generate a wide range of diverse structural models and subsequently, to estimate seismic damage through nonlinear dynamic simulations. The purpose of this part is to compile an extensive data set containing representative information about the structural models, seismic motion, and structural response. Next, leveraging the data set previously generated, two advanced machine learning algorithms are developed and trained in order to accurately estimate structural damage.

3.1. Generating structural models and extracting relevant data

In order to generate the structural models the Finite Element Method was used. Therefore, to accurately represent the three-dimensional behavior of beams and columns, the specific beam type finite element with 2 nodes and 6 degrees of freedom per node was employed. Consequently, local mass and stiffness matrices were associated to every structural element and were subsequently assembled to describe the mass and stiffness of the entire model. The damping matrix $\mathbf{C}$ was constructed using the Reylegh approach, namely proportional to the mass $\mathbf{M}$ and stiffness $\mathbf{K}$ matrix.

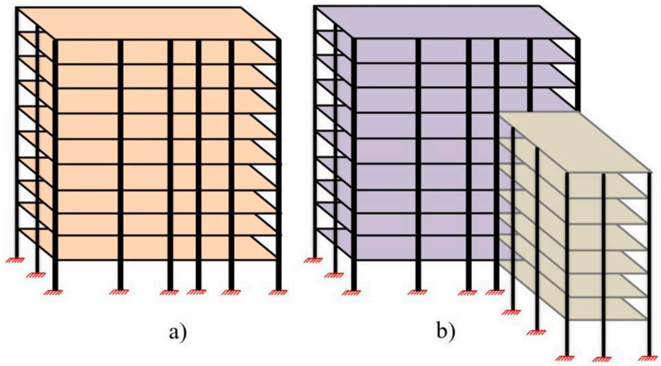


Fig. 1. Regular and irregular geometry type models.

Furthermore, in order to cover a large number of building typologies for the data set, an automatic structural model generation subroutine was developed. Consequently, in each nonlinear dynamic analysis an unique model is constructed by randomly choosing: the total dimension of the layouts, the number of bays, the number of spans and the number of storeys, the story height, the bending capacities for the beams and columns, the live load, the Yong modulus of the structural system and the damping ratio of the reinforced concrete. Certainly, these values were selected arbitrarily, but within carefully chosen intervals, in order to derive structural models with plausible characteristics. Moreover, it is important to note that the generated models fall into two distinct categories: those with regular geometry and those with irregular geometry. Regular models feature a rectangular shape on each floor Fig. 1 a), whereas irregular models exhibit an L-shaped design, with a partial rectangular configuration on higher stories Fig. 1 b).

At the same time, to ensure that the randomly generated models are rationally constructed, an additional two-level filtering process was implemented. The first level, is employed before the nonlinear dynamic analysis starts and is related to the linear behavior of the model. It entails accepting as plausible only those models for which the first mode natural period $T1$ falls within the range of $0.07 \cdot no\_story < T1 < 0.2 \cdot no\_story$ where no_story is the model number of storeys. The second level of filtering is applied only after the dynamic simulation ends and is linked to nonlinear behavior of the model. It involves accepting the structural response as plausible only if the Park and Ang the damage index is below the 2.5 value. The procedural sequence is illustrated in Fig. 2.

After each simulation, the structural response is generated by aggregating data obtained at each discrete time step during the nonlinear dynamic analysis. In this regard, it must be mentioned that preserving the complete time history of structural response variables, such as displacements, velocities, bending moments, and related parameters, necessitates approximately 1.5 gigabytes of hard drive memory. Given the extensive number of dynamic analyses performed, it becomes very important to reduce the memory utilization. Consequently, post-simulation, we selectively retain only the essential parameters describing the structural model and nonlinear response. This data set is used for both training and testing of the machine learning algorithm and is presented in Tables 3, 4 and 5.

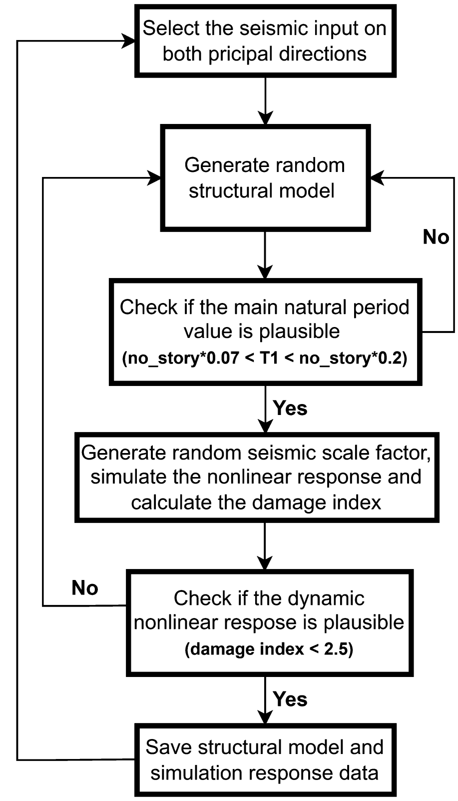


Fig. 2. The overall procedural steps to generate structural models and extract relevant data.

3.2. The nonlinear dynamic simulation procedure

The common method for modeling nonlinear behavior in civil engineering involves employing simplified approaches. For instance, in the case of multi-story moment-resisting frame models, plasticity is concentrated at the level of each element within plastic hinges, which model the nonlinear behavior through force-displacement relationships. Although many sophisticated models employing distributed plasticity to capture the inelastic behavior have been proposed in scientific literature, simplified models have proven to be more practical due to their relatively concise formulations and numerical efficiency. These simplified models have shown that, in general, they can effectively capture the relevant characteristics of the structural response with an identical or only marginally reduced level of precision.

Despite the various possible numerical approaches to model nonlinear behavior using plastic hinges, in the present paper the authors adopted a recently developed method, namely, the Force Analogy Method (FAM).

The first work referring to the FAM, published by [14], presents the fundamental concepts and the application principles in the field of simulating the nonlinear dynamic behavior of structures. Over the years, the method has evolved and has been extensively discussed, including in comprehensive books by [15] and [16], offering detailed principles and illustrative examples. Furthermore, the authors of this paper have also made significant contributions by extending the application of FAM from two-dimensional to three-dimensional models with complex histeretic behavior, as detailed in references [2] and [17].

More recently, [18] employed FAM for conducting 9600 3D unidirectional nonlinear dynamic analyses in order to develop an automated procedure to assess seismic fragility of $3D$ reinforced concrete frame structures. The total duration for the simulation of the 9600 model amounted to almost 128 hours, yielding an average processing time of about 40 seconds per analysis. Therefore, they demonstrated that FAM can be successfully used to conduct large sets of nonlinear dynamic analysis in very reasonable amount of time. Consequently, building upon the numerical framework developed by these authors, the present work introduces enhancements that involve accounting for seismic effects on both principal directions and utilizing structural models with a wider range of geometric characteristics. Additionally, while the prior study conducted 9600 analyses using 24 structural models, the present research involved dynamic nonlinear analyses for 59.569 distinct structural models, which served as the training and testing data for the machine learning algorithm.

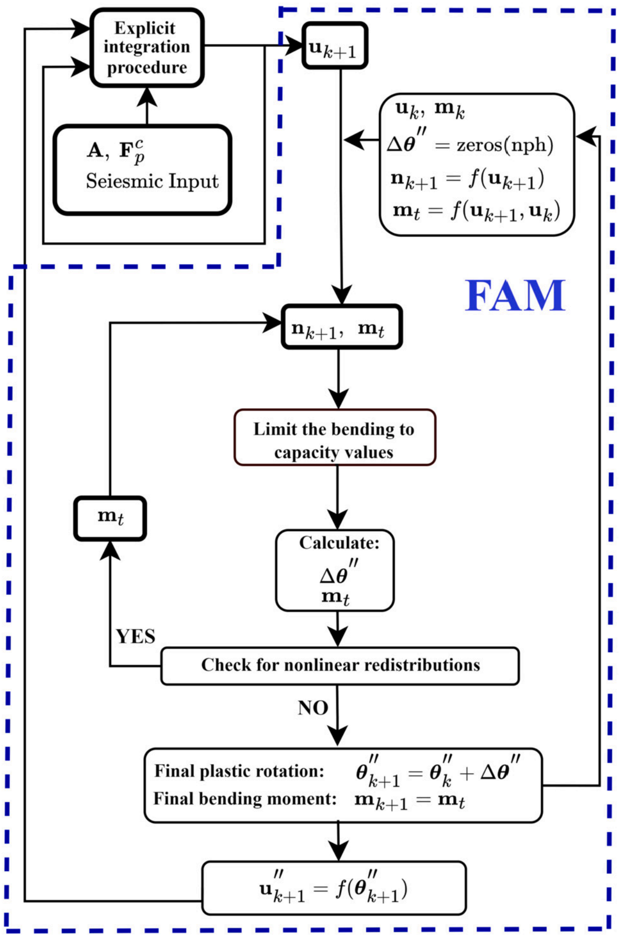


Fig. 3. The nonlinear dynamic structural response calculation procedure.

The FAM key idea is to add in the equation of motion a new vector of unknowns values that accounts for the inelastic displacements $\mathbf{u}''(t)$ of the structural model Eq. (1). This ensures that the stiffness matrix of the structure, determined at the beginning of the analysis, remains constant throughout the inelastic calculation. When combined with the state space representation, FAM becomes a robust, accurate and efficient approach for simulating the dynamic nonlinear response.

$$\mathbf{M} \cdot \ddot{\mathbf{u}}(t) + \mathbf{C} \cdot \dot{\mathbf{u}}(t) + \mathbf{K} \cdot \mathbf{u}(t) = -\mathbf{M} \cdot \mathbf{h} \cdot \ddot{\mathbf{a}}_g(t) + \mathbf{K} \cdot \mathbf{u}''(t) \tag{1}$$

The seismic simulation procedure, as depicted in Fig. 3, consists of two primary operations: Firstly, using previous step data, the displacement vector $\mathbf{u}_{k+1}$ is calculated using the explicit integration procedure in Eq. (2), developed by [16]. Secondly, in order to compute the plastic displacement vector $\mathbf{u}''(t)$, trial values are calculated for the bending moment $\mathbf{m}_t$ and inelastic rotation increment vector $\boldsymbol{\Delta\theta}''$. These quantities describe bending and plastic deformation levels at each plastic hinge location. At the same time, the axial force vector $\mathbf{n}_{k+1}$ is computed as a function of current step displacement vector $\mathbf{u}_{k+1}$. Next, the values within the $\mathbf{m}_t$ vector are individually compared to the bending capacities of the plastic hinges. It must be mentioned that in case of beams a bending moment plastic hinge was used, while for columns the interaction between axial force and bending on both principal direction was considered using a $3D$ capacity surface $(n - m_y - m_z)$. If the bending exceeds the elastic limit, the values are capped to capacity, the exceeding effort being considered as an unknown plastic deformation increment to be determined. Subsequently, a system of equations with *nph* (total number of plastic hinges) equations and *nph* unknowns is used to determine updated values within the $\mathbf{m}_t$ and $\boldsymbol{\Delta\theta}''$ vector. Next, the values in $\mathbf{m}_t$ are once again compared to capacity in order to assess if any inelastic bending redistribution occurred. If there are new values exceeding the elastic limit, the $\mathbf{m}_t$ is sent back to the capping routine, otherwise the final bending moments $\mathbf{m}_{k+1}$ and plastic rotations θ'' are calculated. Ultimately, the plastic displacement vector is computed and sent into the explicit integration formula in order to determine the next step displacement vector. In essence, the calculation of plastic displacement vector is based on a assuming a series of trial values that are iteratively adjusted until equilibrium is reached.

$$\mathbf{z}_{k+1} = e^{\mathbf{A}\cdot\Delta t} \cdot \mathbf{z}_k + e^{\mathbf{A}\cdot\Delta t} \cdot \mathbf{H} \cdot \Delta t \cdot \ddot{\mathbf{a}}_{gk} + e^{\mathbf{A}\cdot\Delta t} \cdot \mathbf{F}_p^c \cdot \Delta t \cdot \mathbf{u}''_k \tag{2}$$

where

$$\mathbf{z}_{k+1} = \begin{bmatrix} \mathbf{u}_{k+1} \\ \dot{\mathbf{u}}_{k+1} \end{bmatrix} \; \mathbf{A} = \begin{bmatrix} \mathbf{O}_n & \mathbf{I}_n \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{C} \end{bmatrix}$$
$$\mathbf{z}_k = \begin{bmatrix} \mathbf{u}_k \\ \dot{\mathbf{u}}_k \end{bmatrix} \; \mathbf{F}_p^c = \begin{bmatrix} \mathbf{O}_n \\ -\mathbf{M}^{-1}\mathbf{K} \end{bmatrix} \tag{3}$$

The dynamic analysis is performed using a database of seismic motions. In this process, the numerical tool selects certain dynamic inputs which are used for each analysis.

3.3. Proposed estimation framework

Given that the main objective is to estimate the damage index based on an extensive set of features, the global estimation framework is represented by Fig. 4.

During the initial phase, we consider that we have a complete data set, characterized by multiple features. Specifically, each data point is represented by a building, with its characteristics and the main parameters of the seismic input. Considering this data set, a first objective is to choose the features that have the most significant impact in the variation of the damage index. To achieve this, we implement a feature selection process starting from a comprehensive set of features properly selected based on specialized structural engineering knowledge. This step is designed to pinpoint several sets of key attributes based on adequately chosen criteria, thereby offering multiple perspectives regarding the optimal methodology for feature selection.

The next step is to split the data between a train data set and the test data set. Once the machine learning model is calibrated on the train data set, it can be effectively used to estimate the damage index from the structural models and seismic input data in the test set. Given the ample data set generated, as described in the previous section, there are no restrictions whatsoever on selecting the proportionality between the training and testing data sets.

As estimation methods, we consider two algorithms:a typical linear regression from the linear models spectrum of methods and K-nearest-neighbors, a model that is rather suitable for estimation more complex and non-linear relationships. The reasoning behind this choice resides in one of the secondary objectives of this paper, which is to present a comparative analysis between the performances of typical linear models in relation to their more complex counterparts.

Eventually, considering typical evaluation indicators such as the R2 score or the Mean Absolute Error, predictions on the test data using the chosen models are further compared. The objective of this concluding evaluation is to furnish a comprehensive explanation of the manner in which various feature sets, when analyzed in conjunction with differing models, affect the aforementioned metrics. This, in turn, determines the efficacy of the models in accurately estimating the damage index, as well as the minimum key elements that must be taken into account as features.

3.4. Estimation algorithms

The first algorithm that we use is linear regression. A linear regression model represents a widely used instrument for estimating a dependent variable y based on an independent variable (or feature) x,

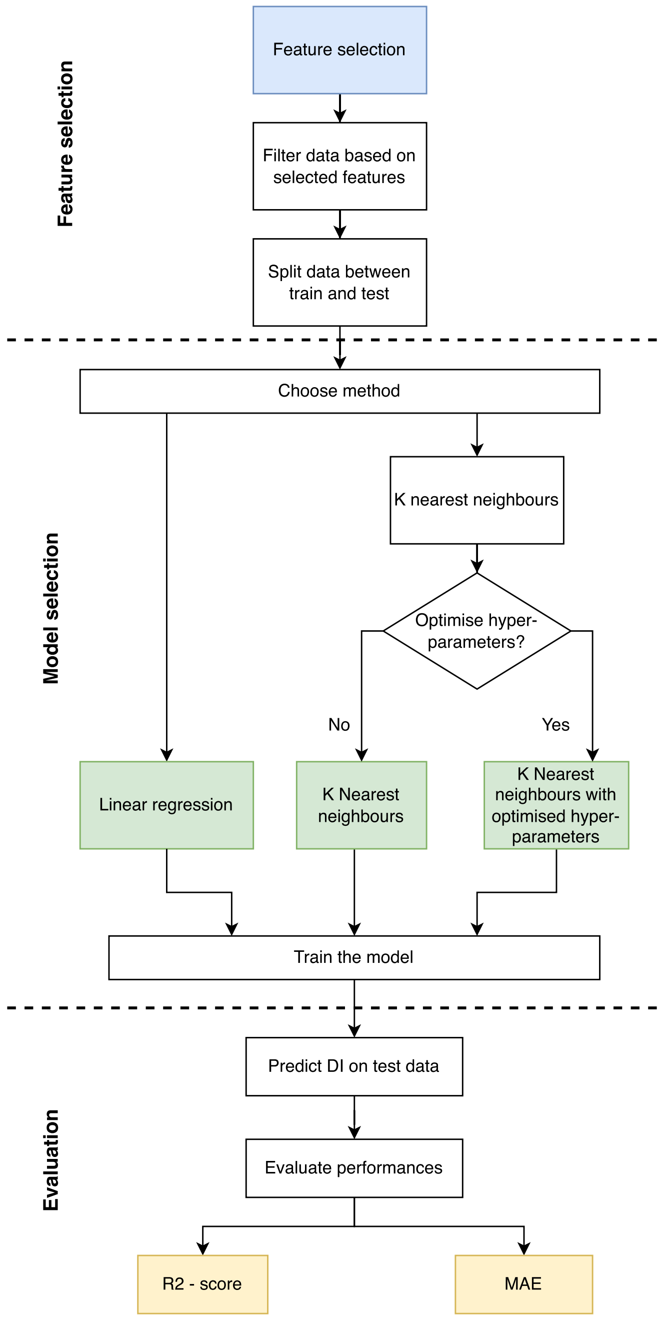


Fig. 4. Proposed framework, emphasising the sequential manner in which features selection, model selection and evaluation of a chosen configuration can be conducted to efficiently estimate the degradation index.

given the assumption that a linear relationship exists between the dependent and independent variables.

The general form of the model is:

$$y = xa + b \tag{4}$$

where a is known as the intercept and b is known as the slope.

This simple model is suitable for simple estimation applications, where a single feature is enough to provide subsequent information for further estimating the dependent variable. However, given the context of this paper where the aim is to estimate the damage index based on multiple building features and a seismic component, the following model represents a more suitable option:

$$y = XA + b \tag{5}$$

where X is a set of independent variables (or features) and A represents a vector of coefficients corresponding to the intercept. This kind of model represents a first step in estimation problems and represents a good solution, due to its simplicity.

Nonetheless, considering obvious nonlinear relationship between the features and the damage index, an alternative algorithm is proposed: the K-nearest neighbors (KNN) algorithm [19]. Although the KNN algorithm is traditionally utilized for classification issues, where it is suitable in clustering analogous entities (neighbors) according to the Euclidean distances among their features, it also offers estimation capabilities. This capability allows the algorithm to be applied to novel feature sets in order to determine the most appropriate and proximate category considering the Euclidian distance.

In the context of the problem presented within this paper, the challenge may not be strictly characterized as a classification dilemma but rather as a regression one. The adaptation of the KNN algorithm for regression purposes is designed to work through the following sequential steps:

1. Choose the number 'K' of neighbors to employ in the estimation process (a decision also fundamental in the classification variant of the algorithm).
2. For a new set of structural model attributes and seismic parameters, for which the aim is to estimate the damage index, compute the Euclidean distance between the new feature vector and all other instances within the selected data set.
3. Identify the 'K' nearest data points in proximity to the new feature vector.
4. Calculate the mean damage index value of these 'K' neighbors and assign this average as the estimated damage index for the new feature vector.

The choice of this algorithm for estimating the damage index resides specifically in its simplicity in solving estimation problems that include several non-linear components.

3.5. Evaluation procedure

Two main indicators are used for the assessment of a machine learning evaluation process: mean absolute error and the R2 score.

The mean absolute error (MAE) represents the average error between the estimated damage index values for multiple buildings and the actual, simulated values:

$$\mathrm{MAE} = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{6}$$

In this case, low MAE error shows that the estimated damage index is relatively close to the true one, considering the scale used for the quantifying the index.

The R2 score is a popular indicator that shows how much of the dependent variable's variance comes as a consequence of the variance of an independent variable. Usually, an R2 score of 1 shows that the estimation of the real data is very good, while an R2 score of 0.4 or lower may show that the choice of model or features might be improved.

As the damage index under consideration typically oscillates between 0 and 2.5 for several pairs of buildings and seismic components, we consider that these two indicators may represent an adequate platform for evaluating an estimation model. Although each indicator possesses distinct characteristics (for instance, the absolute may be susceptible to outliers) it is imperative to also take into account the number of features as a potential indicator. This consideration underscores the significance of data volume required for each estimation scenario. Through such an approach, there may be an opportunity to pinpoint sets of features that yield satisfactory performance while utilizing merely several independent variables. This finding would be particularly advantageous

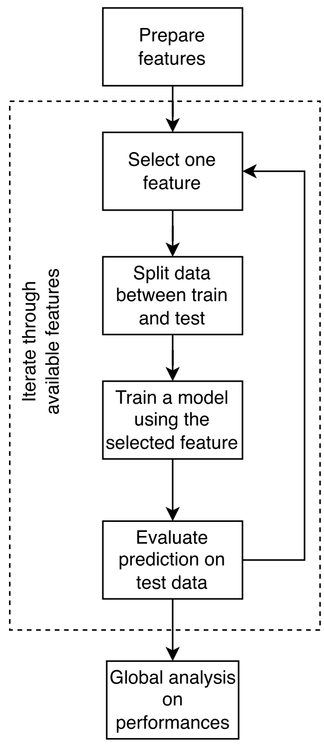


Fig. 5. Feature selection process designed based on a machine learning analysis.

within the scope of this paper's context, as not all building parameters may be readily measurable.

3.6. Feature selection process

The process of feature selection is predicated on a singular analysis of each attribute, in relation to the two predictive models previously mentioned. Specifically, as it is described in Fig. 5, we pick each feature from the feature set, we apply an algorithm to estimate the damage index based solely on that isolated feature, and finally we compare the algorithm's predicted value of the damage index with the actual observed value. This comparative analysis employs two metrics as indicators of the performance: the coefficient of determination ($R2$ score) and the mean absolute error (MAE)."

4. Case study

This section is dedicated to showcasing the effectiveness of the proposed numerical tool. The process consists of two key steps. Firstly, nonlinear dynamic simulations are conducted to generate a comprehensive database containing data about the structural models and their seismic response. Subsequently, the linear regression and K-nearest neighbors machine learning algorithms are trained and tested using this database in order to provide accurate estimations of the damage index for the analyzed structural models.

4.1. Database compilation: structural models and seismic response

In order to generate the database describing the structural models and their seismic response, three computers form the laboratories of National University of Science and Technology POLITEHNICA BUCHAREST were utilized. The simulations spanned a three-month period, involving a total of 59.569 analyses. The dynamic nonlinear simulations were conducted employing a ground motion data set consisting of 20 recordings obtained from three Vrancea (Romania) intermediate-depth earthquakes occurring in 1977, 1986, and 1990. This data set was

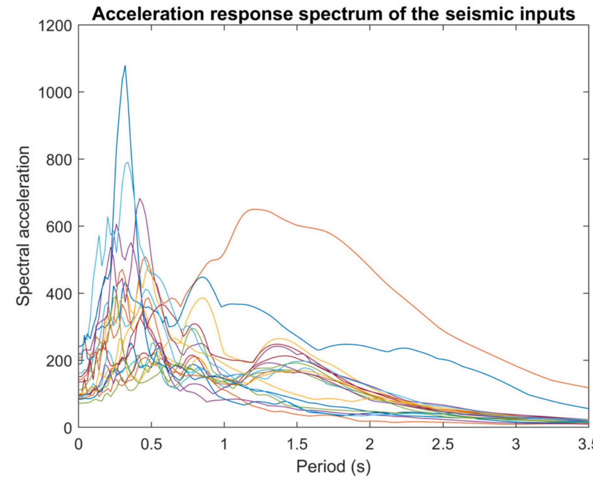


Fig. 6. The acceleration response spectrum of the 20 recorded seismic inputs used.

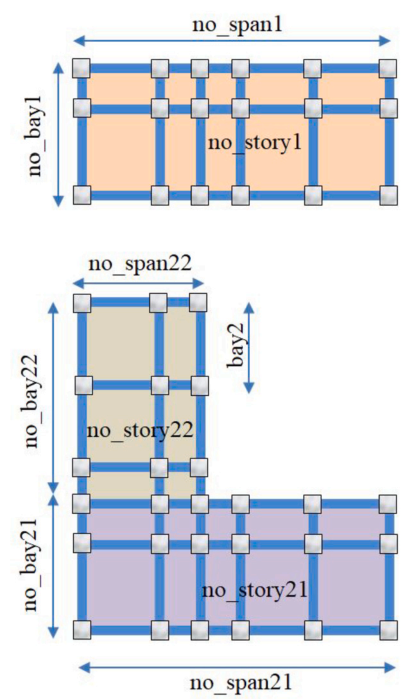


Fig. 7. Regular and irregular geometry type layouts.

specifically chosen to represent the local soil conditions in Bucharest, city located in southern Romania [20]. Fig. 6 displays the acceleration response spectrum for each of these recordings. It can be observed that the most pronounced spectral amplifications are concentrated within the medium period range. Therefore, in each simulation a pair of accelerograms is selected as seismic input along the model's principal directions, Y and X. Moreover, before the start of the analysis, these accelerograms undergo a preprocessing stage, wherein a scale factor is randomly assigned from the range of 1 to 4. The peak ground accelerations of the seismic motions selected to conduct the simulations are presented in Figs. 10 and 11, for each of the principal directions.

Furthermore, a brief presentation of the structural models and the simulation results is illustrated in Figs. 8 to 13, in correlation with Fig. 7. For clarity, the data collected from the total of 59569 simulations is divided into two groups: one comprising results from regular geometry models (29997 simulations) and the other from irregular geometry models (29572 simulations).

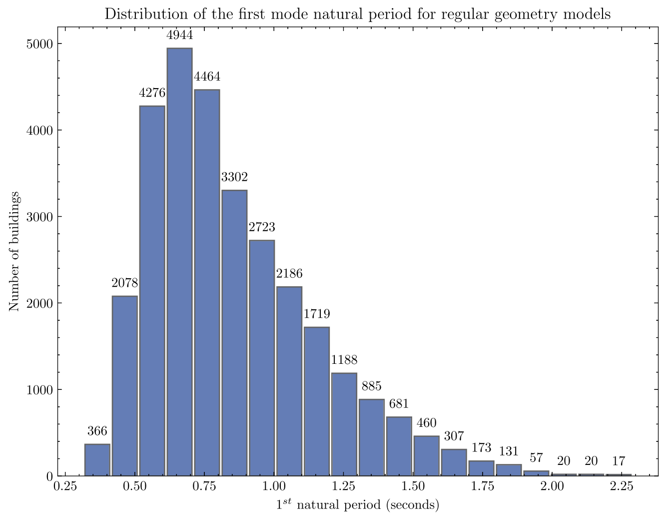


Fig. 8. Distribution of the 1^{st} mode natural period for regular geometry models.

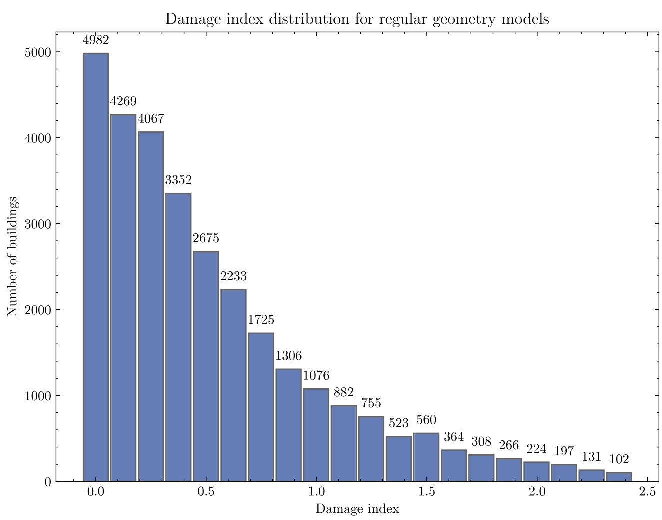


Fig. 9. Damage index distribution for regular geometry models.

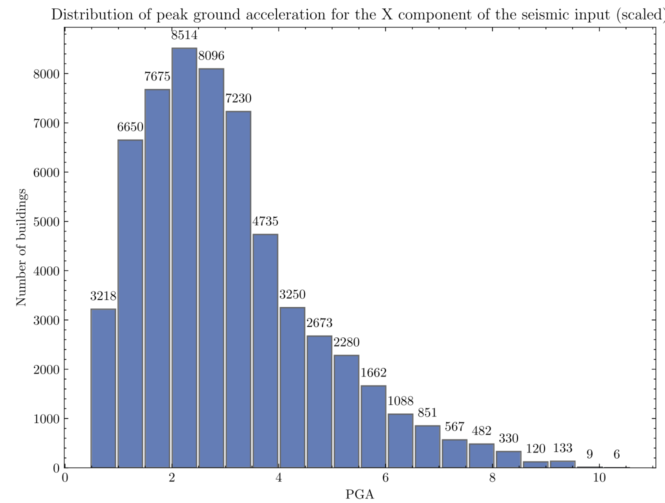


Fig. 10. Distribution of peak ground acceleration for the X component of the seismic input (scaled).

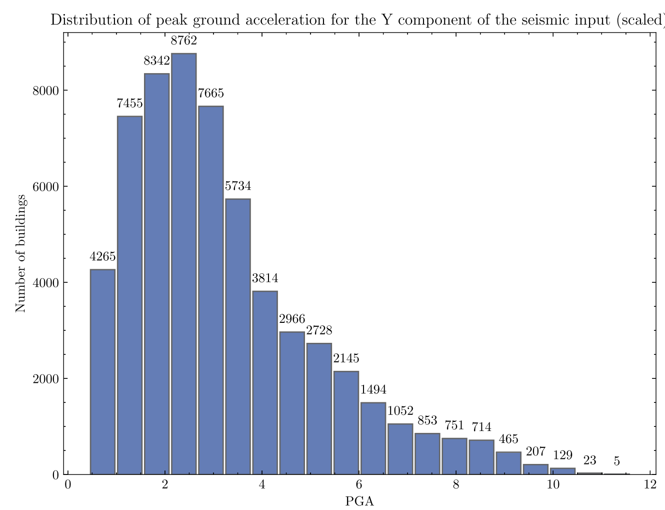


Fig. 11. Distribution of peak ground acceleration for the Y component of the seismic input (scaled).

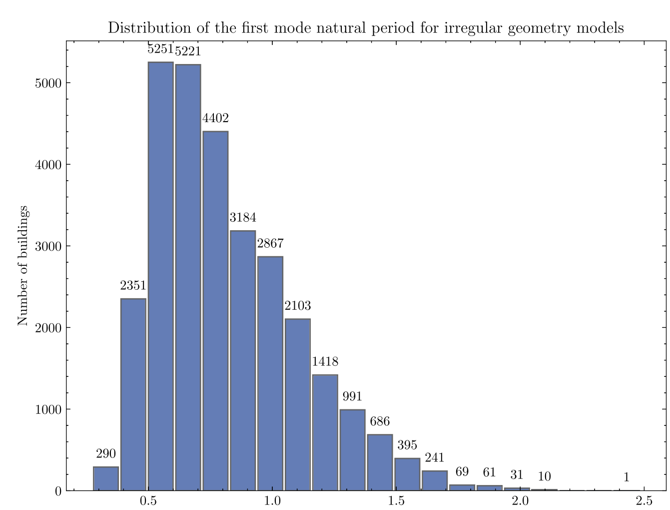


Fig. 12. Distribution of the 1^{st} mode natural period for irregular geometry models.

Regarding the regular geometry layouts, Fig. 8 illustrates that the majority of the models (92.5%) have the first mode natural period between $0.4s$ and $1.3s$, which is consistent with number of story range considered in the random generation phase. Moreover, in Fig. 9 the Park Ang damage index value distribution is presented. It can be observed that 55.57% of the models are below 0.4 value, which is considered the reparable damage limit, while 14.37% exceed the 1 value corresponding to collapse. Furthermore, in Fig. 12 the first mode natural period distribution for the irregular models is presented. It can be seen that these structures are less flexible than the regular ones, almost all the values being situated below $2s$. Regarding the damage index distribution, it can be noted that the results are similar to those obtained for regular models, 61.18% of the values being below 0.4 while 10.72% over 1.

4.2. *Machine learning analysis - results and discussion*

Drawing from the methodology depicted in Fig. 4, the initial phase in obtaining a good model coupled with a suitable set of features commences with the feature selection process. Our initial feature set consists of the following parameters which are detailed in Tables 3, 4 and 5: Fund_Freq1, Fund_Freq2, PGA_of_the_recording_scale_1, PGA_of_the_recording_scale_2, b_Height, dim_x, dim_y, b_st, h_st, b_gr, h_gr, E, MstY, MstX, Mgr, Lshape, bay2, no_span22, no_bay22, no_story22, T1, T2 and T3. It must be highlighted that these were cho-

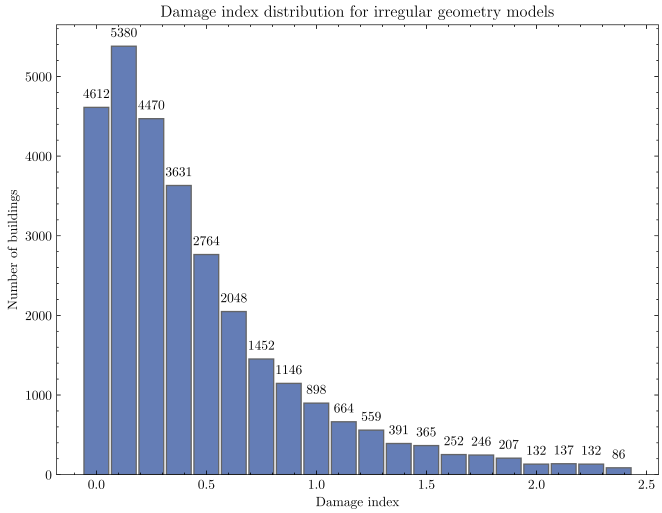


Fig. 13. Damage index distribution for irregular geometry models.

sen based on the specialized structural engineering knowledge of the authors. Additionally, the damage index is designated as the dependent variable. The data set is subsequently constituted of X data points, each characterized by the aforementioned features and a corresponding damage index value.

In pursuit of this, the feature selection process (referenced in Fig. 5) has been applied to the data set. As a result, each combination of a feature with a machine learning algorithm is evaluated, yielding an R2-score and an MAE value. The algorithms used at this step are a general linear regression model and the KNN algorithm with 5 neighbors. The goal is to ascertain the importance of each feature within the estimation process by employing these two indicators—R2 and MAE—for assessment on the test data set. Additionally, for this initial step, the data set is split in two components: a train data set representing 80% of the total data and a test the data set containing the rest of the data points. The results are depicted in Fig. 14.

By assessing the performances of the linear regression model, it seems that no particular feature can form a strong linear relationship with the damage index. Aside from the features associated with seismic input, all other features are associated with low R2 score values and high MAE scores, showcasing a significant error in estimation.

Conversely, when examining the performance metrics associated with the KNN algorithm, it is observed that a collection of nine features exerts a substantial influence on the estimation process. This influence is emphasised by both the R2 score and MAE values. In light of these findings, the following six feature sets are proposed for subsequent phases of the machine learning analysis:

- the "all features" set - which encompasses every proposed feature. This comprehensive set is a critical initial benchmark for consideration.
- the "9 best features" set - which is composed of the top nine features as determined by the feature selection process and evaluated based on the R2 score.
- The "8 best features" set, which mirrors the prior set but is restricted to the foremost eight features.
- The "7 best features" set, which, akin to its predecessors, is limited to the top seven features.
- The "6 best features" set, which is confined to the six most superior features.
- The "2 best features" set, which includes only the two best features.

Aside from finding the best model, these sets are chosen with a subsequent objective in mind, to obtain a minimal number of features that provide a good platform for a damage index estimation.

From the feature selection results illustrated in Fig. 14, it can be noticed that the primary factors influencing the extent of structural damage and its quantification through the R^2 score include the peak ground accelerations of the seismic motions ($PGA\_of\_the\_recording\_scale\_1$ and $PGA\_of\_the\_recording\_scale\_2$), building height ($b\_Height$), the dimensions and bending capacity of columns (b_st, h_st, Mst_y and Mst_X) as well as the main natural periods of the building ($T1$, $T2$, and $T3$). From a structural engineering perspective, this conclusion is expected, given that these parameters are either associated to the magnitude of the seismic input or the vibration of the building and its capacity to sustain deformation. Conversely, in estimating the Park-Ang structural damage index with the proposed method, we believe that most of these parameters would suffice to yield an accurate estimation. This aspect is advantageous, particularly since the analytical method requires all the features to be considered.

The subsequent phase of the analysis involves the training of models utilizing the previously obtained feature sets. Adhering to the methodology depicted in Fig. 4, three types of models are employed to estimate the damage index: a linear regression model, a KNN model with five neighbors, and an optimized version of the KNN model. This final optimized KNN model is determined through a systematic grid search that explores a combination of neighbor counts (3, 5, 11, 15) and methods of quantifying the distance among neighbors (either uniform or weighted). The uniform method implies that all neighbors contribute equally to the prediction, whereas the weighted approach points that are closer neighbors have a more significant influence on the prediction than those further away.

Consequently, for each pairing of feature set and machine-learning algorithm, a comparative analysis of actual versus estimated values is illustrated in Fig. 15. Within each graph, individual points symbolize the pairs corresponding to data points from the test data set; these pairs consist of the actual damage index and its estimated counterpart. To facilitate a visual reference of perfect estimation, a red dotted line is drawn from the plot's origin at an angle of 45 degrees, representing the line along which the real damage index is equal to the estimated damage index.

We can discuss the representations in Fig. 15 from two perspectives: the perspective of the model type and the perspective of the feature set size.

From the model types' perspective, it can be noticed that the best performances are given by the KNN algorithms, with the optimized version yielding a significant number of estimated data points that are equal to the real ones. Another outcome that can be noticed is related to the performance of the linear regression, which points that linear models cannot be used in this context for estimating the damage index. Graphically, this aspect is underscored by the spread of the data points relative to the 45 degrees reference axis. One could consider this outcome relatively expected, given the importance of each individual feature in the feature selection process using the linear regression model (Fig. 14).

Finally, a more in-depth analysis is presented by Fig. 16, where the same pairs consisting of a feature set and estimation algorithm are depicted in a 2-by-2 representation of three indicators: the R2-score, MAE and the feature usage ratio (FUR). The Feature Usage Ratio quantifies the proportion of utilized features in the estimation process relative to the total number of available features.

This graphical method aims to contextualize the performance of each scenario in relation to an optimal reference point, designated by the origin of each point. For example, in a plot between the mean absolute error and the feature usage ratio, the origin represents the hypothetical scenario of achieving perfect prediction without utilizing any features. This kind of representation is similar to the Pareto plot representation used in optimization ([23]), where a plot is also drawn to highlight the dominant points in terms of both performances.

As we can see, in the first plot showing the scenarios from the MAE and R2-score perspective, the best models and feature sets are shown

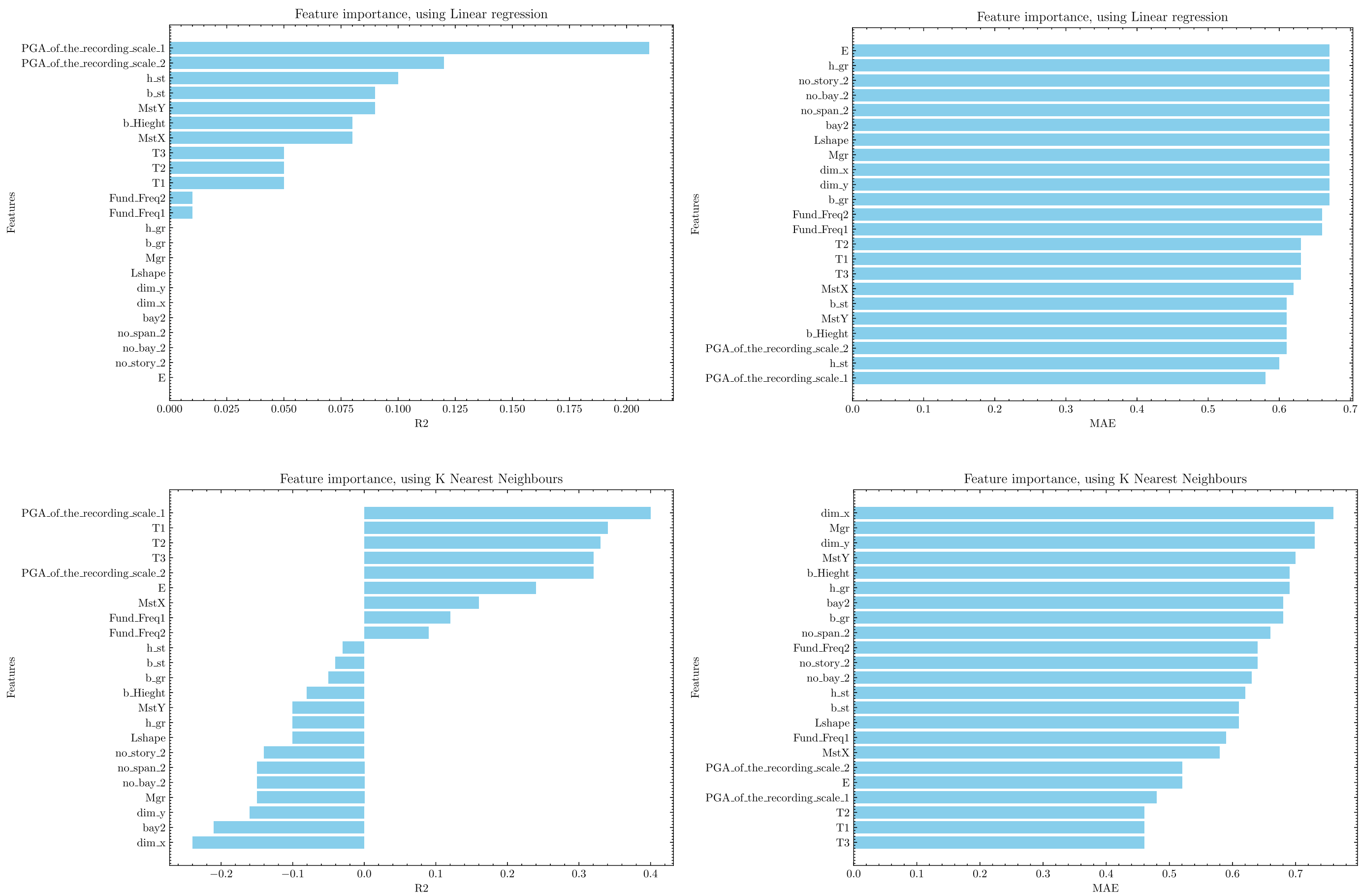


Fig. 14. Feature importance hierarchy, given the method proposed in Fig. 5. Results can be reproduced using the Jupyter Notebook available at [21].

closest to the origin of the plot. There, we see that the optimized KNN model applied to all 9 features provides the best R2 score and the smallest MAE. Subsequently, the optimized KNN version generally provides the best performances, closely followed in by the KNN model with 5 neighbors. This observation reinforces the suitability of the KNN model for damage index estimation and validates the effectiveness of the feature selection process in selecting the most impactful features from the entire collection.

Subsequent plots offer additional insights into performance while also incorporating the Feature Usage Ratio. In this case, the best solutions are connected by a plot that emphasises the dominance relative to the other solutions. Moreover, these plots provide a framework for stakeholders to determine the most suitable feature set and the best model for their specific needs. For instance, if data acquisition is not constrained and all nine features are obtainable, then the optimized KNN model would be recommended for estimation as it ensures very good performance in terms of MAE and R2 score - denoted by the blue square marker positioned on the left of the plot. Alternatively, if not many features are available, the best feature set choice might be placed to the right side of the plot, towards the solution encompassing only two features and relatively low estimation performance.

Overall, by using the proposed feature selection process and using the optimized KNN version, the damage index may be accurately estimated based on the features of the building and of the associated seismic input. Even though linear models such as the linear regression do not represent a suitable estimation solution, the KNN model is able to obtain very good performances both in terms of R2-score and MAE. For example, Table 1 highlights the performance of the optimized K-Nearest Neighbors (KNN) algorithm in estimating the damage index, given the best features identified during the feature selection process and evaluated using the R^2 metric (Fig. 14). It is noteworthy that an adequate estimation ($R^2 \sim 0.75$) is achieved using the peak ground accelerations of the seismic inputs and the natural periods of the building, whereas only the two best features yield an R^2 score of just 0.57. This result is a significant indicator of the potential to estimate the damage index with a minimal number of features efficiently.

Table 1
Damage index estimation performances using the best features identified from the feature selection process, based on the KNN algorithm and the R^2 metric (Fig. 14).

Id	Feature set	Algorithm	Metric	Value
1	best 6 features	Optimised KNN	MAE	0.1012
2	best 6 features	Optimised KNN	R2	0.749
3	best 2 features	Optimised KNN	MAE	0.1668
4	best 2 features	Optimised KNN	R2	0.5789

Finally, given the artificially generated, extensive set of buildings that serve as the training platform for the machine learning algorithms, various test-train ratios have been investigated. The purpose of this analysis is to ascertain the performance of the algorithms relative to the size of the test set. Consequently, in Table 2, we analyze the performances in terms of Mean Absolute Error (MAE) and R2 for two feature sets: one containing all features, and one comprising only the best features identified through the feature selection process. Additionally, we consider the K-nearest neighbors (KNN) algorithm with optimized hyperparameters, which reflects the algorithm that showed the best performance in Fig. 16. The results demonstrate how performance improves as the test dataset increases in size. For instance, with a test size constituting 20% of the initial dataset, the R2 index is approximately 0.83, a significant improvement from scenarios where the test size was

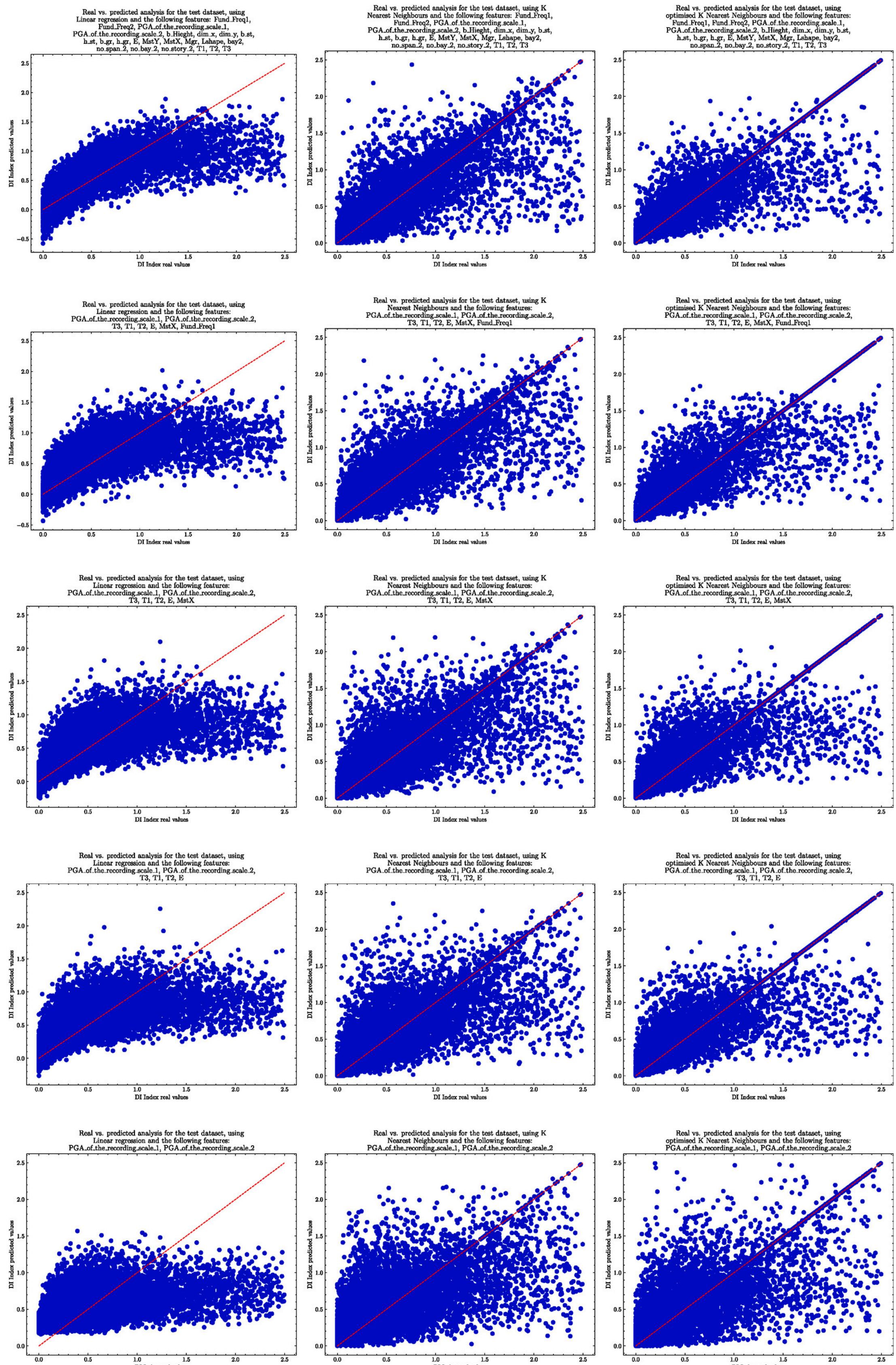


Fig. 15. Comparison between the estimated damage index and the real damage index for each building-seismic activity data point in the test data-set, for each pair of machine learning model (Linear Regression, KNN and optimized KNN) and feature set. Results can be reproduced using the Jupyter Notebook available at [22].

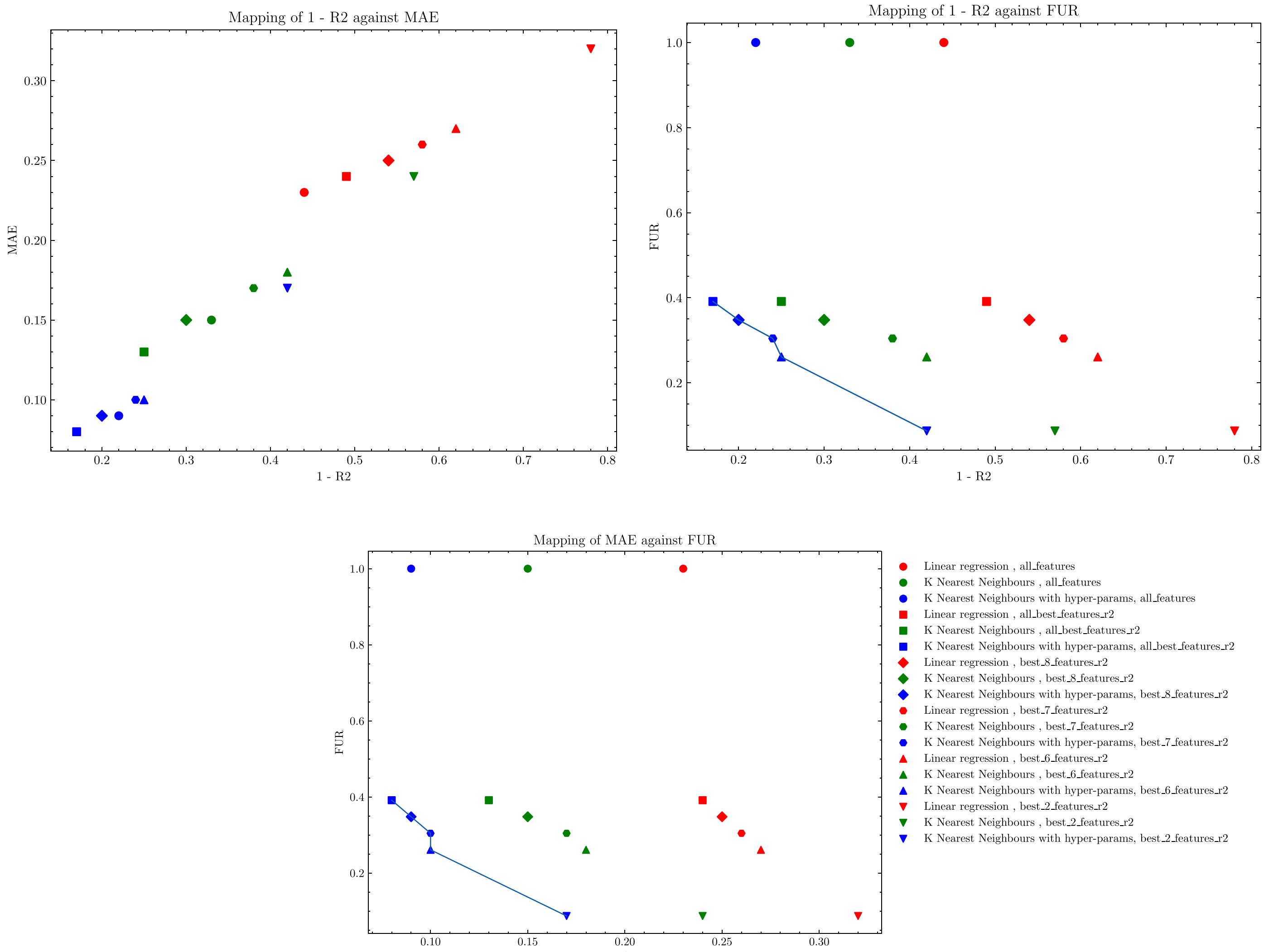


Fig. 16. Results representation in 2x2 objective space, with the ideal point in the origin of the plot. In the last two plots, the blue line connects the dominant solutions, relative to the indicators used for evaluation. Plots can be reproduced using the Jupyter Notebook available at [22].

Table 2
Results analysis for different test size proportions. Results can be reproduced using the script *compare_dataset_ratio.py* available at [22].

Id	Feature set	Test set size ratio of total set size	Metric	Value
1	best features	0.2	MAE	0.0782
2	best features	0.2	R2	0.8313
3	all features	0.2	MAE	0.0882
4	all features	0.2	R2	0.7831
5	best features	0.4	MAE	0.0913
6	best features	0.4	R2	0.8106
7	all features	0.4	MAE	0.1001
8	all features	0.4	R2	0.7659
9	best features	0.6	MAE	0.1092
10	best features	0.6	R2	0.7718
11	all features	0.6	MAE	0.1176
12	all features	0.6	R2	0.733
13	best features	0.8	MAE	0.1424
14	best features	0.8	R2	0.7081
15	all features	0.8	MAE	0.1527
16	all features	0.8	R2	0.6576

80% of the initial dataset and the R2 score was only 0.65. This underscores the importance of dataset size in achieving accurate predictions and highlights the capability of the KNN algorithm to provide adequate estimations—a 0.65 R2 score—even with a smaller dataset. However, usually real data of such dimensions are very difficult to acquire from actual buildings, so future research should also aim to improve the estimation accuracy with even smaller data sets.

5. Conclusions

The aim of this paper is to introduce an innovative methodology for the assessment of seismic structural damage of moment-resisting frame structures. This entails the application of machine learning algorithms which are trained and tested on large set of data generated through a complete simulation procedure. The machine learning algorithms employ different regression approaches in order to estimate the Park-Ang structural damage index.

The learning database used in this paper is composed of 59569 unique structural models and their associated nonlinear seismic response synthetically expressed through the Park - Ang damage index. Compared to the studies presented in [7] and [8], our research surpasses them in both data set size and complexity. The structural models considered are randomly generated according to specific plausibility criteria and have both regular and irregular geometry layouts providing a comprehensive representation of the diverse building structures encountered in current practice.

Table 3
Data set describing the regular structural models.

Number of storeys (no story1) - no_story										
Number of story	4	5	6	7	8	9	10	11	12	13
No. of building	2795	3225	3415	3155	3164	2887	2865	2983	2676	2832
Storey height - Height										
Range (m)	3.0-3.1	3.1-3.2	3.2-3.3	3.3-3.4	3.4-3.5	3.5-3.6	3.6-3.7	3.7-3.8	3.8-3.9	3.9-4.0
No. of building	2502	2772	2822	3033	2833	3148	2967	2910	2602	4408
Building height - b_Height										
Range (m)	12-16	16-20	20-24	24-28	28-32	32-36	36-40	40-44	44-48	48-52
No. of building	3240	3658	3708	3803	3255	3026	3554	2821	1862	1070
Column section width - b_st										
Range (m)	0.3-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	1.0-1.1	1.1-1.2	1.2-1.3	1.3-1.4
No. of building	4289	3363	3452	3560	3234	2772	2294	3294	2021	1718
Column section height - h_st										
Range (m)	0.3-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	1.0-1.1	1.1-1.2	1.2-1.3	1.3-1.4
No. of building	4373	3337	3316	3774	3269	2594	2336	3112	2206	1680
Uniform distributed load - py										
Range (kN/m)	6.0-6.8	6.8-7.6	7.6-8.4	8.4-9.2	9.2-10.0	10.0-10.8	10.8-11.6	11.6-12.4	12.4-13.2	13.2-14.0
No. of building	2946	3138	3012	3046	2849	2903	3694	2083	3014	3312
First natural period of vibration - T1										
Range (s)	0.4-0.6	0.6-0.8	0.8-1.0	1.0-1.2	1.2-1.4	1.4-1.5	1.5-1.7	1.7-1.9	1.9-2.1	2.1-2.3
No. of building	2444	9220	7766	4909	2907	1566	767	304	77	37
Second natural period of vibration - T2										
Range (s)	0.3-0.5	0.5-0.7	0.7-0.9	0.9-1.1	1.1-1.3	1.3-1.5	1.5-1.7	1.7-1.9	1.9-2.1	2.1-2.3
No. of building	1703	10155	8106	5046	2734	1397	498	254	87	17
Third natural period of vibration - T3										
Range (s)	0.3-0.4	0.4-0.6	0.6-0.7	0.7-0.9	0.9-1.0	1.0-1.2	1.2-1.3	1.3-1.5	1.5-1.6	1.6-1.8
No. of building	957	7239	9378	6218	3485	1867	685	131	24	13
The maximum number of formed plastic hinges divided by the total number of plastic hinges defined - no_ph_max/nph										
Range	0.0-0.1	0.1-0.1	0.1-0.2	0.2-0.3	0.3-0.3	0.3-0.4	0.4-0.4	0.4-0.5	0.5-0.6	0.6-0.6
No. of building	3294	2704	5086	5610	5471	4207	2336	999	253	37
Hysteretic energy divided by the input seismic energy - HE(i)/IE(i)										
Range	-0.0-0.1	0.1-0.2	0.2-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.5	0.5-0.6	0.6-0.7	0.7-0.8
No. of building	3625	1617	1888	2141	2733	3155	3823	4774	4789	1452
Concrete Young Modulus - E										
Range*1e+07 (kN/m^2)	1.3-1.5	1.5-1.8	1.8-2.0	2.0-2.3	2.3-2.5	2.5-2.7	2.7-2.9	2.9-3.2	3.2-3.5	3.5-3.7
No. of building	27	31	364	4190	4343	4244	4253	4496	3987	4062
Maximum base shear on X direction - FbXmax										
Range*1e+04 (kN)	0.04-0.3e	0.3-0.6	0.6-0.9	0.9-1.2	1.2-1.5	1.5-1.8	1.8-2.1	2.1-2.3	2.3-2.6	2.6-2.9
No. of building	4749	9862	7441	4163	2204	973	385	126	67	27
Maximum base shear on Y direction - FbYmax										
Range*1e+03 (kN)	0.4-3.5	3.5-6.6	6.6-9.7	9.7-13	13-16	16-19	19-22	22-25	25-28	28-31
No. of building	6255	10634	7124	3356	1571	625	278	118	29	7
The maximum number of formed plastic hinges - nr_ph_max										
Range*1e+02	0-2.1	2.1-4.1	4.1-6.2	6.2-8.2	8.2-10	10-12	12-14	14-16	16-18	18-21

Table 3 (*continued*)

Beams bending capacity - Mgr										
Range*1e+02 (kNm)	0.9-2.2	2.2-3.5	3.5-4.8	4.8-6.1	6.1-7.4	7.4-8.7	8.7-10	10-11	11-13	13-14
No. of building	3022	5908	6259	5207	4429	2242	1679	653	493	105

Column bending capacity on X direction - MstX										
Range*1e+03(kNm)	0.5-0.8	0.8-1.1	1.1-1.3	1.3-1.6	1.6-1.9	1.9-2.1	2.1-2.4	2.4-2.6	2.6-2.9	2.9-3.2
No. of building	2986	4475	4499	4802	4050	3745	3031	1667	631	111

Column bending capacity on Y direction - MstY										
Range*1e+03 (kNm)	0.6-0.8	0.8-1.1	1.1-1.3	1.3-1.5	1.5-1.7	1.7-1.9	1.9-2.1	2.1-2.4	2.4-2.6	2.6-2.8
No. of building	3822	3283	4012	3993	3454	3247	3151	2782	1708	545

The number of spans (no span1) - no_span					**The number of bays (no bay1) - no_bay**			
Range	2	3	4	5	2	3	4	5
No. of building	7506	7377	7466	7648	7763	7509	7413	7312

Beam section height - h_gr							
Range (m)	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
No. of building	201	1924	4277	6434	8527	7763	871

Beam section width - b_gr				
Range (m)	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6
No. of building	9176	14975	5249	597

Damage Index - DI										
Range	0.0-0.2	0.2-0.5	0.5-0.7	0.7-1.0	1.0-1.2	1.2-1.5	1.5-1.7	1.7-2.0	2.0-2.2	2.2-2.5
No. of building	9251	7419	4908	3031	1958	1278	924	574	421	233

Table 4
Data set describing the irregular structural models.

Number of storeys (no story21) - no_story										
Number of story	4	5	6	7	8	9	10	11	12	13
No. of building	2605	3285	3436	3213	2908	2749	2895	2813	2886	2782

Number of storeys (no story22) - no_story2										
Number of story Range	2	3	4	5	6	7	8	9	10	11
No. of building	8712	5864	4434	2967	2684	1611	1376	884	717	323

Storey height - Height										
Range (m)	3.0-3.1	3.1-3.2	3.2-3.3	3.3-3.4	3.4-3.5	3.5-3.6	3.6-3.7	3.7-3.8	3.8-3.9	3.9-4.0
No. of building	2543	2572	2543	2757	2918	2920	2968	2865	3153	4333

Building height - b_Height										
Range (m)	12-16	16-20	20-24	24-28	28-32	32-36	36-40	40-44	44-48	48-52
No. of building	3160	3682	3613	3524	3223	3079	3196	3020	1939	1136

Column section width - b_st										
Range (m)	0.3-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	1.0-1.1	1.1-1.2	1.2-1.3	1.3-1.4
No. of building	4196	3246	3508	3568	3119	2387	2447	3255	2163	1683

Column section height - h_st										
Range (m)	0.3-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	1.0-1.1	1.1-1.2	1.2-1.3	1.3-1.4
No. of building	4230	3341	3319	3792	2856	2415	2669	3157	2059	1734

Uniform distributed load - py										
Range (kN)	6.0-6.8	6.8-7.6	7.6-8.4	8.4-9.2	9.2-10.0	10.0-10.8	10.8-11.6	11.6-12.4	12.4-13.2	13.2-14.0
No. of building	2892	2701	2793	2971	2778	3084	3877	2128	2775	3573

First natural period of vibration - T1										
Range (s)	0.3-0.5	0.5-0.8	0.8-1.0	1.0-1.2	1.2-1.4	1.4-1.7	1.7-1.9	1.9-2.1	2.1-2.3	2.3-2.5
No. of building	2641	10472	7586	4970	2409	1081	310	92	10	1

(*continued on next page*)

Table 4 (*continued*)

Second natural period of vibration - T2										
Range (s)	0.3-0.5	0.5-0.7	0.7-0.8	0.8-1.0	1.0-1.2	1.2-1.4	1.4-1.6	1.6-1.7	1.7-1.9	1.9-2.1
No. of building	2679	9504	7433	4968	2601	1481	596	245	58	7
Third natural period of vibration - T3										
Range (s)	0.2-0.4	0.4-0.5	0.5-0.6	0.6-0.8	0.8-0.9	0.9-1.1	1.1-1.2	1.2-1.3	1.3-1.5	1.5-1.6
No. of building	1765	8159	8291	5700	3158	1569	599	237	87	7
The maximum number of formed plastic hinges divided by the total number of plastic hinges defined - nr_ph_max/nph										
Range	0.0-0.1	0.1-0.1	0.1-0.2	0.2-0.3	0.3-0.3	0.3-0.4	0.4-0.5	0.5-0.5	0.5-0.6	0.6-0.6
No. of building	3812	3688	4735	5085	4841	3772	2306	1078	229	26
Hysteretic energy divided by the input seismic energy - HE(i)/IE(i)										
Range	-0.0-0.1	0.1-0.2	0.2-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.6	0.6-0.7	0.7-0.8
No. of building	3408	2043	2070	2481	2928	3479	3991	4317	3790	1065
Concrete Young Modulus - E										
Range*1e+07 (kN/m^2)	1.3-1.6	1.6-1.8	1.8-2.0	2.0-2.3	2.3-2.5	2.5-2.8	2.8-3.0	3.0-3.2	3.2-3.5	3.5-3.7
No. of building	25	26	758	4388	4206	3963	4244	4102	4049	3811
Maximum base shear on X direction - FbXmax										
Range*1e+03 (kN)	0.4-4.9	4.9-9.5	9.5-14	14-18	18-23	23-27	27-32	32-36	36-41	41-45
No. of building	6652	11269	6807	3068	1210	377	126	37	19	7
Maximum base shear on Y direction - FbYmax										
Range*1e+03 (kN)	0.6-5.3	5.3-9.9	9.9-15	15-19	19-24	24-28	28-33	33-38	38-42	42-47
No. of building	6460	11342	6806	3087	1225	448	128	33	27	16
The maximum number of formed plastic hinges - nr_ph_max										
Range*1e+02	0-2.7	2.7-5.3	5.3-8.0	8.0-11	11-13	13-16	16-19	19-21	21-24	24-27
No. of building	13611	8613	4120	1928	823	274	126	61	9	7
Beams bending capacity - Mgr										
Range*1e+02 (kNm)	0.9-2.2	2.2-3.5	3.5-4.8	4.8-6.1	6.1-7.4	7.4-8.7	8.7-10	10-11	11-13	13-14
No. of building	2893	6083	6619	5223	3914	2141	1597	557	439	106
Column bending capacity on X direction - MstX										
Range*1e+03 (kNm)	0.5-0.8	0.8-1.1	1.1-1.3	1.3-1.6	1.6-1.9	1.9-2.1	2.1-2.4	2.4-2.6	2.6-2.9	2.9-3.2
No. of building	3023	4206	4650	4580	3946	3745	2911	1718	691	102
Column bending capacity on Y direction - MstY										
Range*1e+03 (kNm)	0.6-0.8	0.8-1.1	1.1-1.3	1.3-1.5	1.5-1.7	1.7-1.9	1.9-2.1	2.1-2.4	2.4-2.6	2.6-2.8
No. of building	3810	3036	4110	3851	3391	2806	3208	3029	1750	581
The bay dimension - bay2										
Range (m)	5.0-5.1	5.1-5.2	5.2-5.3	5.3-5.4	5.4-5.5	5.5-5.6	5.6-5.7	5.7-5.8	5.8-5.9	5.9-6.0
No. of building	2588	2776	2768	2593	2825	2891	2764	2819	2593	4955
Beam section height - h_gr										
Range (m)	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0			
No. of building	124	1903	4514	6672	8188	7457	714			
The number of span (no span21) - no_span					**The number of spans (no span22) - no_span2**					
Range	2	3	4	5	2	3	4	5		
No. of building	7671	7454	7291	7156	7341	7626	7490	7115		
No. of building	3812	3688	4735	5085	4841	3772	2306	1078	229	26

Table 4 (*continued*)

The number of bays (no bay21) - no_bay					**The number of bays (no bay22) - no_bay2**					
Range	1	2	3	4	1	2	3	4		
No. of building	15555	7943	4362	1712	6009	6063	6113	11387		
Beam section width - b_gr										
Range (m)	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6						
No. of building	9419	14853	4764	536						
Damage Index - DI										
Range	0.0-0.2	0.2-0.5	0.5-0.7	0.7-1.0	1.0-1.2	1.2-1.5	1.5-1.7	1.7-2.0	2.0-2.2	2.2-2.5
No. of building	9992	8101	4812	2598	1562	950	617	453	269	218

Table 5
Data set describing main parameters of the seismic inputs.

PGA of the recording on X direction										
Range (m/s^2)	0.7-1.7	1.7-2.7	2.7-3.8	3.8-4.8	4.8-5.8	5.8-6.8	6.8-7.8	7.8-8.8	8.8-9.8	9.8-10.8
No. of building	9868	16189	15326	7985	4953	2750	1418	812	253	15
PGA of the recording on Y direction										
Range (m/s^2)	0.7-1.8	1.8-2.9	2.9-4.1	4.1-5.2	5.2-6.3	6.3-7.4	7.4-8.5	8.5-9.6	9.6-10.8	10.8-11.9
No. of building	11720	17104	13399	6780	4873	2546	1604	1179	336	28
First natural frequency of the recording Fund Freq1 & Fund Freq2										
Range (Hz)	0.4-0.7	0.7-1.1	1.1-1.4	1.4-1.8	1.8-2.1	2.1-2.4	2.4-2.8	2.8-3.1	3.1-3.4	3.4-3.8
No. of building	12	2	1	0	0	0	0	2	2	1

Furthermore, a machine-learning framework is proposed to determine both the most suitable features and an appropriate estimation algorithm capable of accurately predicting the damage index based on the seismic input and building characteristics. The procedure involves selecting a minimal set of features with the highest importance and choosing the best algorithm for estimating the damage index. Results show that a trained K-Nearest Neighbors algorithm with optimized hyper-parameters can accurately estimate the damage index on a separate test data set, achieving an R2 score of 1 and a mean absolute error of approximately 0.1. These results show very promising performances, considering the extensive data set used, the minimal number of features and the simplicity of the proposed machine-learning algorithm. This aspect holds considerable importance for structural practitioners, as the traditional analytical method for calculating the damage index [1] is notably complex and requires all building parameters, making it time-consuming. With this approach, engineers can estimate the damage index of a building more rapidly and achieve an adequate estimation even with a reduced set of parameters.

It must be mentioned that, the utilization of such a substantial and diverse data set mitigates the risk of overfitting in our machine learning algorithm, contributing to the robustness of the prediction models proposed in this paper.

The authors' main contribution is the development of a advanced numerical tool which incorporates two main features: a finite element and force analogy method based computational routine that is able to conduct fast nonlinear dynamic simulations and a machine learning analysis framework that can be efficiently used to determine the best features and the best estimation algorithm for predicting the damage index. The proposed method can be adjusted for other structure types, given that the associated damage index can be analytically determined for each structure and the structure generation method can be adapted with the required structural engineering knowledge.

Compared to other relevant works, our study represents a step forward in the application of machine learning techniques to structural engineering. Although the proposed method may fundamentally seem similar to the applicability principles of existing machine learning algorithms, our approach is distinguished by the utilization of 3D simulation modes compared to those presented by [7] and by the usage of nonlinear dynamic analysis compared to [8] which uses 3D structural models but linear analysis. At the same time, the approach presented in this paper seems to provide satisfactory results similar to those obtained by [9] and [10] using ANNs, but employing fewer features that describe the seismic motions. Moreover, compared to those presented by [11], the current study discusses RC structures instead of masonry buildings and uses a different damage estimation procedure, based on Park and Ang damage index rather than the vulnerability index formulation. Furthermore, our paper distinguishes by the utilization of a much larger structural response training data set compared to all of the papers cited before. Consequently, our method introduces a novel technique for generating structures, in order to create a comprehensive dataset as the foundational platform for the machine learning algorithm. With these innovations, our study aims to propose a new approach to addressing the data requirements limitations highlighted in various other studies [4,12,13].

Furthermore, this article showcases a tool that could prove to be a valuable resource for researchers and practitioners aiming to establish a simplified, fast and accurate approach for estimating structural damage. The proposed numerical instrument, is readily accessible at the following web address [24], providing a practical and applicable resource for the scientific and engineering communities.

Future research should consider the development of fast nonlinear simulation procedures for a wider range of structural systems and the use of more advanced algorithms such as Random Forrest for better estimating the damage index. Furthermore, future research will also address interoperability challenges in order to provide a better framework for structural engineers in estimating relevant building parameters, even for other building types.

CRediT authorship contribution statement

Vasile Calofir: Software, Formal analysis, Conceptualization. **Ruben-Iacob Munteanu:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Mircea-Stefan Simoiu:** Writing – review & editing, Writing – original draft, Software, Investigation, Conceptualization. **Karol-Cristian Lemnaru:** Software, Formal analysis, Data curation.

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

We have shared the link to the code and data in the references section.

Acknowledgement

We want to express our gratitude to the "Academy of Romanian Scientists, Ilfov 3,050044 Bucharest, Romania" for funding this research work (Contract No. 37/11.04.2023). We would also like express our gratitude to the National University of Science and Technology POLITEHNICA Bucharest (UNSTPB), Bucharest, Romania for funding this work.

References

[1] Ruben Munteanu, Vasile Calofir, Florin Mota, George Nica, Large scale damage assessment framework for buildings in urban areas; the effect of active control implementation, 2023.

[2] George Bogdan Nica, Munteanu Ruben Iacob, Vasile Calofir, Mihail Iancovici, Modelling nonlinear behavior of 3d frames using the force analogy method, Structures 35 (2022) 1162–1174.

[3] Theodora Dumitrescu, Radu Dobrescu, Improved algorithm for seismic extreme events prediction, UPB Sci. Bull. 83 (1) (2021).

[4] Ehsan Harirchian, Seyed Ehsan Aghakouchaki Hosseini, Viviana Novelli, Tom Lahmer, Shahla Rasulzade, Utilizing advanced machine learning approaches to assess the seismic fragility of non-engineered masonry structures, Results Eng. 21 (March 2024) 101750.

[5] Haytham F. Isleem, Naga Dheeraj Kumar Reddy Chukka, Alireza Bahrami, Solomon Oyebisi, Rakesh Kumar, Tang Qiong, Nonlinear finite element and analytical modelling of reinforced concrete filled steel tube columns under axial compression loading, Results Eng. 19 (September 2023) 101341.

[6] P. Thisovithan, Harinda Aththanayake, D.P.P. Meddage, I.U. Ekanayake, Upaka Rathnayake, A novel explainable AI-based approach to estimate the natural period of vibration of masonry infill reinforced concrete frame structures using different machine learning techniques, Results Eng. 19 (September 2023) 101388.

[7] Oliver Richard de Lautour, Piotr Omenzetter, Prediction of seismic-induced structural damage using artificial neural networks, Eng. Struct. 31 (2) (2009) 600–606.

[8] Musa Arslan, M. Ceylan, T. Koyuncu, An ann approaches on estimating earthquake performances of existing rc buildings, Neural Netw. World 22 (2012) 443.

[9] Konstantinos Morfidis, Konstantinos Kostinakis, Approaches to the rapid seismic damage prediction of r/c buildings using artificial neural networks, Eng. Struct. 165 (2018) 120–141.

[10] Konstantinos Morfidis, Konstantinos Kostinakis, Seismic parameters' combinations for the optimum prediction of the damage state of r/c buildings using neural networks, Adv. Eng. Softw. 106 (2017) 1–16.

[11] Tiago Ferreira, João Estêvão, Rui Maio, Romeu Vicente, The use of artificial neural networks to estimate seismic damage and derive vulnerability functions for traditional masonry, Frontiers of Structural and Civil Engineering 14 (2020).

[12] Shashank Reddy Vadyala, Sai Nethra Betgeri, John C. Matthews, Elizabeth Matthews, A review of physics-based machine learning in civil engineering, Results Eng. 13 (March 2022) 100316.

[13] Rosette Niyirora, Wei Ji, Elyse Masengesho, Jean Munyaneza, Ferdinand Niyonyungu, Ritha Nyirandayisabye, Intelligent damage diagnosis in bridges using vibration-based monitoring approaches and machine learning: a systematic review, Results Eng. 16 (December 2022) 100761.

[14] Kevin K.F. Wong, Rong Yang, Inelastic dynamic response of structures using force analogy method, J. Eng. Mech. 125 (10) (1999) 1190–1199.

[15] Li Gang, K.F. Wong Kevin, Theory of Nonlinear Structural Analysis: The Force Analogy Method for Earthquake Engineering, 2014, pp. 1–352.

[16] G.C. Hart, K. Wong, Structural Dynamics for Structural Engineers, Wiley, 1999.

[17] Munteanu Ruben Iacob, Florin Moţa, Vasile Calofir, Cătălin Baciu, New approach to nonlinear dynamic analysis of reinforced concrete 3d frames; an accurate and computational efficient mathematical model, Appl. Sci. 12 (3) (2022).

[18] George-Bogdan Nica, Florin Pavel, Gabriel Hojda, A fast nonlinear dynamic analysis automated approach to produce fragility curves for 3d rc frames, Eng. Struct. 281 (2023) 115695.

[19] Zhongheng Zhang, Introduction to machine learning: k-nearest neighbors, Ann. Transl. Med. 4 (11) (June 2016) 218.

[20] Anabella Cotovanu, Radu Vacareanu, Local site conditions modeling in stochastic simulation of ground motions generated by vrancea (Romania) intermediate-depth seismic source, J. Seismol. 24 (2020) 229–241.

[21] Vasile Calofir, Mircea Stefan Simoiu, Ruben Munteanu, DIGITERRA - feature selection JupyterNotebook, https://mybinder.org/v2/zenodo/10.5281/zenodo.10200266/?labpath=feature_selection.ipynb.

[22] Vasile Calofir, Mircea Stefan Simoiu, Ruben Munteanu, DIGITERRA - estimation models comparison JupyterNotebook, https://mybinder.org/v2/zenodo/10.5281/zenodo.10200266/?labpath=compare_models.ipynb.

[23] Kenneth V. Price, Rainer M. Storn, Jouni A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization. Natural Computing Series, Springer, Berlin, New York, 2005.

[24] Vasile Calofir, Mircea Stefan Simoiu, Ruben Munteanu, DIGITERRA - digital platform for estimating the building degradation index with machine learning techniques, https://zenodo.org/doi/10.5281/zenodo.10200266, November 2023.

Universitatea Nationala de Stiinta si Tehnologie POLITEHNICA Bucuresti

Splaiul Independenţei 313, Sector 6, cod 060042 , Bucureşti, ROMÂNIA

Articolul:

"SYSTEM ANALYSIS AND DESIGN OF A SEISMIC DAMAGE ESTIMATION WEB PLATFORM USING MODEL-BASED ENGINEERING"

Autori:

Vasile Calofir, Mircea Stefan Simoiu, Ruben Munteanu, Sergiu Stelian Iliescu

a fost recenzat şi aprobat şi este în curs de publicare în revista "**Scientific Bulletin**", Series C, Electrical Engineering and Computer Science, ISSN 2286 – 3540.

23.09.2024

SYSTEM ANALYSIS AND DESIGN OF A SEISMIC DAMAGE ESTIMATION WEB PLATFORM USING MODEL-BASED ENGINEERING

Vasile Calofir[1], Mircea Stefan Simoiu[2], Ruben Munteanu[3], Sergiu Stelian Iliescu[4]

Predicting seismic damage is crucial in seismic engineering for constructing resilient buildings and enabling effective emergency responses. Traditional dynamic models for damage index estimation are complex and time-consuming, typically requiring specialized users. However, this process can be optimized and simplified using machine learning models, making it accessible to both engineers and non-specialized users. This paper proposes a formal system analysis procedure to determine the functional requirements and optimal architecture for a web platform that estimates building damage using machine learning, based on building parameters and seismic motion. The analysis identifies user profiles and adapts requirements to satisfy a wide range of users.

Keywords: system analysis, model-based system enrgineering, SysML, web application, damage index

1. **Introduction**

Accurate estimation of seismic damage in structures is a crucial component of earthquake engineering, serving as a foundation for designing resilient buildings and infrastructure. The potential consequences of earthquakes—ranging from structural failure to economic losses and loss of life—necessitate a thorough understanding of how various structures respond to seismic forces. Seismic damage estimation provides essential insights into the behavior of structures during earthquakes, enabling engineers and researchers

Associate Professor PhD Eng, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania, e-mail: vasile.calofir@upb.ro

Lecturer PhD Eng, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania, e-mail: mircea_stefan.simoiu@upb.ro

PhD Eng, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania, e-mail: ruben.munteanu@upb.ro

Professor PhD Eng, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania, e-mail: stelian.iliescu@upb.ro

to predict potential failure modes and assess the effectiveness of design strategies. By identifying vulnerable elements within a structure, this process informs the development of targeted retrofitting and reinforcement measures, thereby enhancing structural resilience.

Furthermore, seismic damage estimation is integral to the formulation of building codes and standards. It supports the creation of regulations that ensure new constructions are capable of withstanding seismic events, while also guiding the assessment and upgrading of existing structures. This proactive approach not only mitigates the risk of catastrophic failure but also contributes to public safety and economic stability. In addition to its role in engineering and design, seismic damage estimation is critical for disaster management and policy-making. It enables the quantification of potential economic losses, informing the allocation of resources for emergency preparedness and post-disaster recovery. The data generated from these estimations also play a significant role in the insurance industry, aiding in the assessment of seismic risk and the pricing of policies.

Damage assessment is a topic of significant interest among researchers, and the scientific literature is rich with important papers in this field. In one important recent paper [1] the authors proposed a regional-scale time history analysis method for evaluating seismic damage risk, addressing gaps in existing assessment techniques. The method included automated modeling, response calculation, and 3-D visualization, alongside a calculation method for building damage degrees (DD). The method's accuracy was validated using real earthquake damage data from 192 buildings affected by the Ludian Earthquake, providing a valuable tool for accurate seismic damage assessment and risk mitigation. In another recent paper [2], the authors assessed the seismic vulnerability of structures in high seismic zones using fragility curves derived from the HAZUS method. They applied a performance-based design to evaluate building capacity against seismic events in Padang. By calculating damage probability through a lognormal distribution, the study reveals a significant vulnerability to collapse, highlighting the importance of accurate seismic assessments to prevent structural failures during earthquakes. In [3] the authors proposed a rapid seismic-damage assessment method using deep learning and spectrum-compatible data augmentation. They addressed the challenge of limited strong ground motion data by using continuous wavelet transform (CWT) to generate augmented strong-motion data. This data was then applied to deep-learning algorithms for predicting building damage on a regional scale. The method was tested through case studies and compared to traditional data-augmentation approaches. Results showed that the proposed method reduces dispersion in seismic responses and improves prediction accuracy, achieving 87.4% accuracy with a processing time under 1 second. In another important paper [4], the authors developed a numerical model database for typical regular reinforced concrete (RC) frame structures to improve seismic damage simulation in large cities. Unlike existing models, which are based solely on structural theories and design codes, the models in their were calibrated using refined numerical models to enhance accuracy. The study includes the impact of reinforcement corrosion and construction year on structural performance. The accuracy of the models was validated by comparing them with refined models and real RC frame buildings. Furthermore in [5] the authors tackled the challenge of correlating changes in a structure's dynamic properties with specific damage levels for effective structural health monitoring (SHM). They applied a methodology using numerical analyses to assess damage based on SHM data. The study focused on 3D models of reinforced concrete buildings designed to non-seismic standards in the Mediterranean region. By conducting non-linear dynamic analyses and modal analyses, they captured variations in dynamic properties due to seismic events. The study assessed the probability of frequency changes at different damage levels, offering insights into damage detection and assessment.

The web platform developed in this paper estimates structural damage induced by earthquake excitation using the well-established Park-Ang Damage Index [6], which has been refined in subsequent research [7]. This index offers a robust and widely accepted method for quantifying damage to structural elements, providing a reliable assessment of seismic impacts on structural integrity. The platform leverages data from a large number of nonlinear dynamic analyses, utilizing a numerically efficient methodology previously developed by the authors [8, 9, 10]. Building on our previous work [11], which introduced non-linear machine learning algorithms for damage index estimation, this paper extends the analysis to propose these algorithms for broader public use via a web platform. We formally define the requirements for this platform, addressing the needs of both experienced and non-experienced users, and conclude with an architecture that guides implementation, mitigating the risks of developing an inaccessible or ineffective tool.

2. Relevant Research Initiatives in the Field

In recent years, researchers have explored various approaches to simplify the dynamic methods used to estimate damage indices. These approaches frequently utilize machine learning algorithms, which have emerged as powerful tools for providing accurate predictions in an efficient manner. Consequently, researchers and governing institutions have made significant contributions both in terms of datasets and software tools.

At the international level, the Global Earthquake Model Foundation has been developing tools and datasets for seismic hazard and risk assessment, with a key focus on the OpenQuake platform. OpenQuake is an open-source software capable of integrating machine learning models to predict the seismic performance of buildings and infrastructure. By standardizing seismic risk assessment methodologies globally, GEM initiatives facilitate the development of resilient infrastructure in earthquake-prone regions. Another notable platform is the Prompt Assessment of Global Earthquakes for Response (PAGER) system developed by the USGS, which estimates the impact of earthquakes globally. PAGER provides information on the impact of significant earthquakes worldwide, informing emergency response authorities, governments, aid agencies, and the media about the potential disaster's extent. It rapidly assesses earthquake impact by comparing the population exposed at each intensity level using economic and human loss models based on previous earthquakes in various countries or regions.

Another example is the SimCenter platform, under development at the University of California, Berkeley, which offers state-of-the-art computational modeling and simulation tools, user support, and educational materials for the natural hazards engineering research community. The goal is to enhance the national capability to simulate the impact of natural hazards on structures, utility networks, and communities. Additionally, the center enables leaders to make better-informed decisions regarding the necessity and effectiveness of potential mitigation strategies. It employs machine learning to improve modeling and simulation using data from experimental tests, field investigations, and previous simulations.

The Next Generation Attenuation (NGA) project, an ongoing research initiative at the Pacific Earthquake Engineering Research Center, contributes to the development of predictive models for seismic motion attenuation. This project integrates precise models that can estimate structural responses in various seismic scenarios.

Although the aforementioned research initiatives significantly advance hazard impact estimation, they often operate with large-scale data without providing specific analyses for certain types of buildings. In this context, investigating solutions that focus on estimating the damage index considering specific building-level characteristics would be valuable. Moreover, given the complexity of the parameters involved in both building and earthquake software modeling, exploring an approach for developing a platform suitable for both specialists and non-specialists in the construction field would be beneficial. Such a platform would engage a broader audience, fostering a deeper connection between people and the residential building construction domain.

A web platform would also represent a useful tool from a software perspective [12, 13, 14]. Web-based tools have become increasingly popular for interfacing various complex applications [15, 16], dedicated to both specialized and non-specialized users. However, in the research domain of structural analysis, to the best of our knowledge, the availability of such tools is either limited or confined to private institutions, making them inaccessible to the public.

A useful approach for desging such a web platform, given its novel character, is system analysis through the model-based system engineering (MBSE)

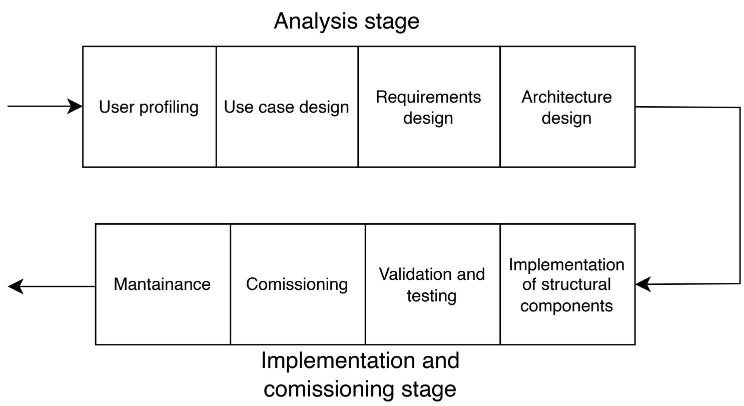


FIGURE 1. System design proposed method

formalism. This formal analysis involves developing diagram models that enable a robust, maintainable formulation of the system requirements and architecture. The foundation for this formalism is defined in reference works such as [17] and [18]. The diagrams used in this formalism, designed under the *SysML* standard, facilitate a more in-depth analysis of the connection between the system user and the informational system itself, regardless of the complexity [19].

3. **Proposed method and formal analysis**

The system analysis was based on a structured method that ensures a clear and detailed understanding of the necessary steps for developing a web platform integrating machine learning models. This method involves identifying user profiles, defining requirements, and designing the system architecture, followed by implementation, testing, and maintenance [17]. The diagram below illustrates these essential stages of the life cycle of such a system:

Consequently, the stages presented in Fig. 1 refer to:

- **Determining user profiles** - This first stage of the life cycle involves accurately identifying the actors who interact with the information system to be implemented. The user profile should outline the type of interaction, whether these users are the primary beneficiaries of the system or require a certain level of training to be able to interact with the system.
- **Identifying use cases** - This stage focuses on outlining the interactions between users and the system. Use case diagrams can be used for actual modeling of the use cases, or sequence diagrams can be employed to add a temporal dimension to the interaction representation [20].
- **Determining functional requirements** - This critical stage for system analysis highlights the main functionalities of the system, related to the needs of the beneficiaries. Functional requirements should express what the system does to facilitate user-system interactions defined previously in the use cases [21].

- **Designing an architecture** - This stage involves defining a system architecture that outlines how the system implements the proposed functional requirements. The architecture highlights potential technologies used and how they interact with each other.
- **Actual implementation** - This stage immediately follows system analysis and aims at the actual implementation of the system based on the proposed architecture.
- **Testing and validation** - At this stage, the functionalities of the system are tested and validated in various scenarios. The goal is to ensure that the system meets all the requirements set out in the design phase [21].
- **Commissioning and maintenance** - This stage includes both the commissioning of the system and its continuous adaptation according to new user needs or legislative or business changes that may require system adaptation.

3.1. Profiling users that interact with the damage estimation platform

In the initial stage of system analysis, we need to outline the user profiles of the system. This step can be based on knowledge of the process, past experiences in implementing similar processes, or even interviews with the people directly involved in operating the respective process.

Users must be categorized into two main groups [18]:

- End users - interact with the system without requiring specialized training; can be considered beneficiaries of the system's capabilities.
- System users - have extended rights over system usage, but also need specialized training to interact with it.

For the web platform estimating the degradation index, we can consider four types of users as follows:

- ***Non-specialist* end user** - This type of user includes individuals who, although having a low level of knowledge in structural engineering, are interested in using the platform.
  This user should be able to enter basic data about buildings into the platform, such as: number of floors, building height, and plan dimensions. As a result of interacting with the web platform, this type of user will obtain adequate estimates to provide an overview of the potential structural degradations of the analyzed building, considering the limitations imposed by the types of data entered.
  Regarding the configuration of a degradation index estimation experiment, this user will use a default estimation model and will have the option to choose from a series of pre-defined accelerograms in the platform.
  This category may include, for example, real estate agents or insurance agents.

- ***Intermediate level*** **end user** - This category includes users who have a relatively advanced level of training in civil engineering. These users can provide the application with additional parameters, such as the dimensions of structural elements (columns, beams), modulus of elasticity, or the type of building (regular or irregular).
  In configuring simulation experiments, these users can select both an estimation model and the desired accelerogram.
  As a result of interacting with the system, these users would typically obtain a more accurate degradation index estimate.
- ***Specialist*** **end user** - This category includes users with specialized training in civil engineering. These users need to provide the platform with all the specific parameters (number of storeys, storey hieght, building height, column section width, column section height, uniform distributed load, first natural period of vibration, second natural period of vibration, third natural period of vibration, the maximum number of formed plastic hinges, divided by the total number of plastic hinges defined, hysteretic energy divided by the input seismic energy, concrete Young Modulus, maximum base shear on X direction, maximum base shear on Y direction, the maximum number of formed plastic hinges, beams bending capacity, column bending capacity on X direction, column bending capacity on Y direction, the bay dimension, beam section height, the number of span, the number of bays, beam section width) of a building required for estimating the index with machine lerning algorithms.
  They also have the ability to define new accelerograms in the system, which they can use later in various simulation experiments. Similar to intermediate users, these users can choose an estimation model and the desired accelerogram.
  Regarding the estimation of the degradation index, these users will obtain the highest level of precision in conducting simulation experiments.
- ***Administrator*** **system user** - Users in this category have extended capabilities of using the application, including rights to manage and configure the platform.
  In configuring a simulation experiment, these users, in addition to the capabilities of specialist users, can introduce new estimation algorithms and calibrate algorithms already defined in the platform.
  Concerning the administration function, these users can create, modify, or delete users and can extract specific reports with all simulation experiments executed in the platform.

3.2. Defining Use Cases

With the user categories already defined, the next step is to define the use cases. As mentioned in the previous section, a use case defines an interaction of a user type with the system, forming the basis for defining functional requirements.

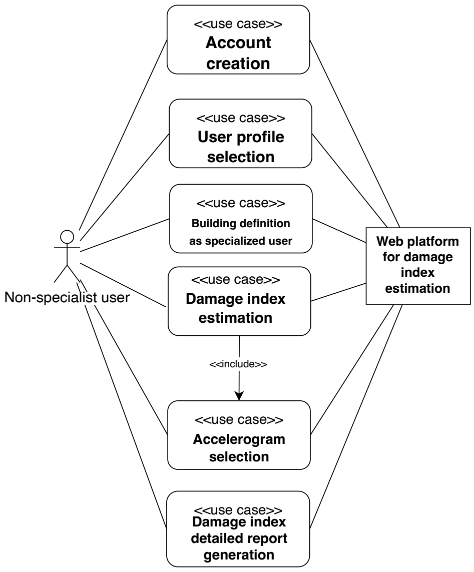


FIGURE 2. *Non-specialist* end user

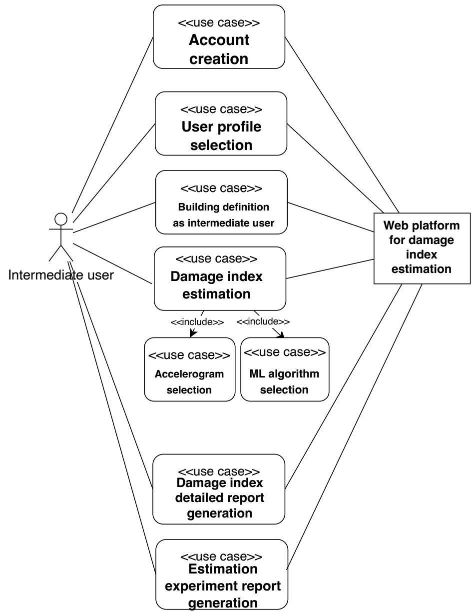


FIGURE 3. *Intermediate* end user

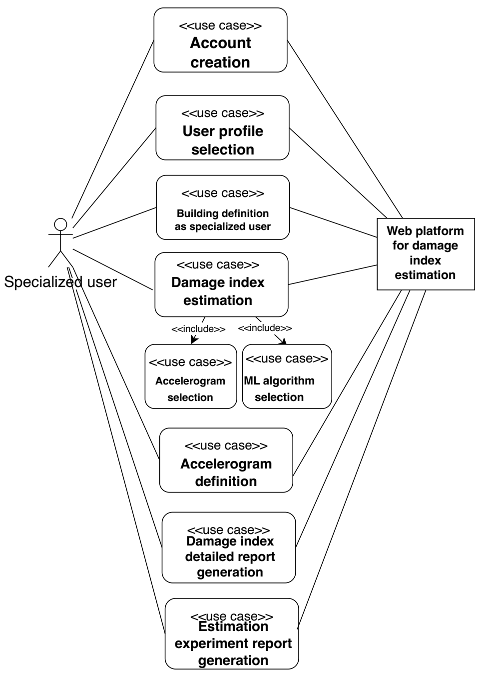


FIGURE 4. *Specialist* end user

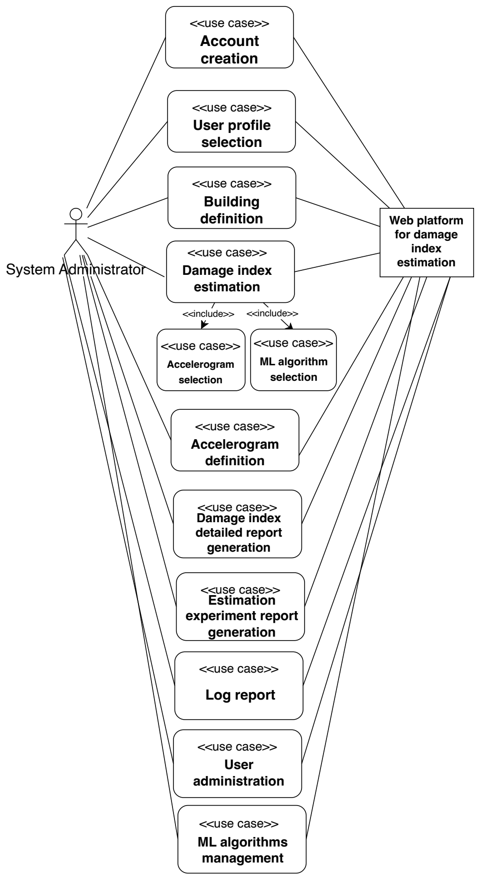


FIGURE 5. *Administrator* end user

FIGURE 6. Use case diagrams for the end users and system users of the platform

In Fig. 6, we can analyze the use case diagrams for the end users and system users of the platform. These diagrams reflect a relatively simple pattern of user interaction with the system. This pattern is highlighted by the fact that in each interaction, there is only one human actor involved. This aspect is particularly useful as it indicates, from an early stage of system analysis, that the system can be implemented on a relatively simplistic architecture.

Moreover, we observe that there are three main stages using the platform for estimating the degradation index. These stages are:

- Choosing the user type
- Configuring a simulation experiment in a manner appropriate to the involved user
- Generating the results of a simulation experiment in a form suitable to the involved user

This procedural sequence implies analyzing user-system interaction from a temporal perspective. To further highlight this temporal dimension, sequence diagrams have been developed.

Given the complexity of each interaction and the subset of similar capabilities that can be identified among user categories, we can analyze the sequence diagrams for the experienced user (Fig. 7). Both sequence diagrams and use case diagrams were developed based on the *SysML* standard [19].

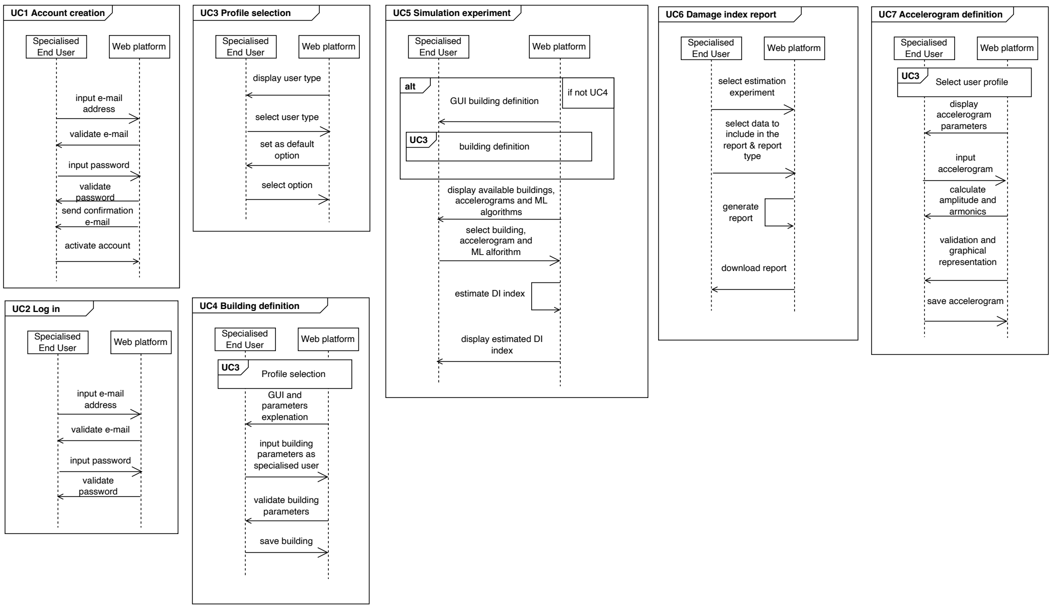


FIGURE 7. Sequence diagram for the *specialised* end user

Before analyzing the sequence diagrams, it is important to mention that each sequence diagram corresponds to a use case defined by a unique identifier. For example, the Create Account use case is identified by the synthetic key *UC1*.

Conceptually, sequence diagrams describe the interactions of users with the system as a whole, without going into specific details that involve certain additional functional or non-functional constraints. Thus, relative to the main stages defined earlier, we can highlight the following classification:

- Use cases UC1, UC2, and especially UC3 are associated with choosing the user type
- Use cases UC4, UC5, UC7 are associated with configuring and executing the simulation experiment
- Use case UC6 is associated with the reporting stage

The temporal dimension attributed to the sequence diagrams provides an important detail regarding the existing conditions in the application. Specifically, sub-processes that are conditioned by the execution of other sub-processes can be identified. For example, entering building data is conditioned by selecting a user type, as the functionality of entering data will be influenced by the restrictions associated with each user category. Another relevant example is the use case associated with executing a simulation experiment, as this execution involves a configuration stage where the user needs to enter building data.

3.3. Developing Functional Requirements

Based on the use cases, functional and non-functional requirements have been developed to concretely express what the platform must do in relation to user needs.

Since the requirements are a consequence of the use cases, they are presented in a structured manner with respect to each use case. Fig. 8 presents the functional requirements diagram, designed using the SysML standard. Fig. 8 depicts all functional requirements developed, depending on the user profile. Given that there are many common requirements among end-users, one can notice the predominance of the green color associated with the typical end-user profile. Additionally, it is important to note that most of the requirements are linked through two relationship types:

- **Derivation relationship** - signifying that one child requirement is derived from another parent requirement.
- **Composition relationship** - enforcing that a parent requirement is satisfied only when all the child requirements are satisfied.

By analyzing the requirements, we can observe, for example regarding the building generation requirements, that there are certain validation mechanisms that can be implemented through a software component of the platform, as well as certain storage capabilities that the platform must have. For instance, the platform must allow a user to enter data associated with multiple buildings, which is associated with the need for the platform to store the data entered by the user.

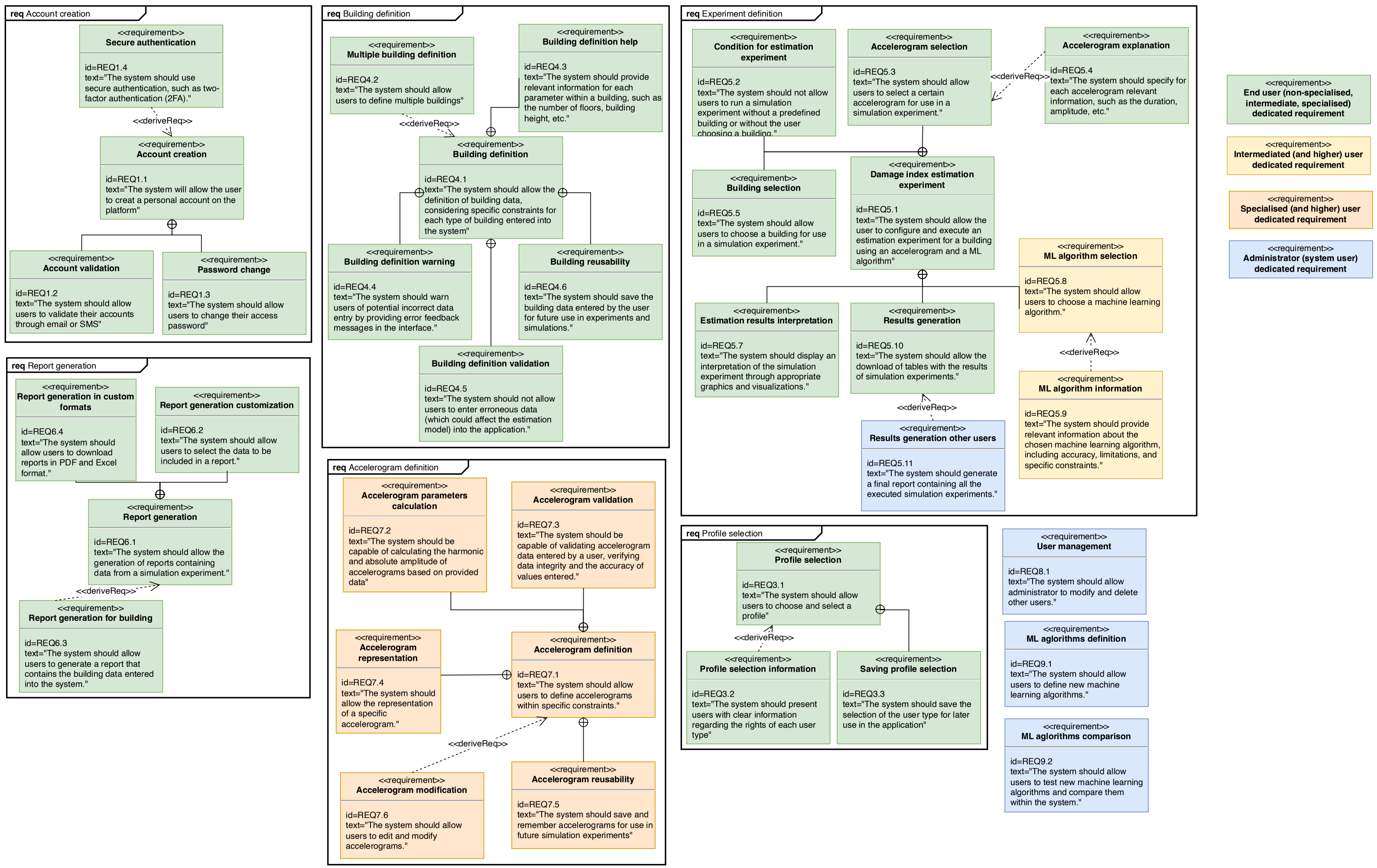


FIGURE 8. Requirements diagram and associated legend, depending on the user profile

In another case, for example the requirements related to the estimation experiment, the requirements defined here describe how the system must represent information in interaction with the user so that the user can configure a simulation experiment in an accessible way. These requirements aim to provide relevant information to the user when choosing a building, proposing an accelerogram, or a machine learning algorithm. Also, the system must be able to provide detailed explanations regarding the index estimation, adapted to the type of user involved in the interaction.

The administrator represents a system user that has complete flexibility when using the web platform. He can also use the platform for damage index estimation, but also has a user-management component. He also has different reporting tools available, while also being the only user profile who can manage machine learning models inside the system, given the sensible nature of these algorithms.

Overall, the proposed functional requirements focus primarily on outlining a service deeply oriented towards the human user and their experience in the process of estimating the degradation index. This aspect comes as a consequence of the fact that civil engineering concepts can be relatively complex for non-specialized users, which is why the platform must offer a friendly and useful environment to the target audience, regardless of their level of training.

3.4. **Platform Architecture for Degradation Index Estimation**

Finally, based on the functional requirements, the architecture described in Fig. 9 was proposed.

From a hardware perspective, the platform for estimating the degradation index operates on a single server. This server allows storing estimation experiments and user data in a *Sqlite* database, considering the estimated volume of experiments and the number of users involved.

At the software level, the interaction between the database and the end user will be realized through an application developed in Python. This application enables both the development of an intuitive interface based on the *Streamlit* framework and the integration with the already implemented module containing pre-trained machine-learning algorithms.

The central point of this architecture is the user interface developed, as mentioned earlier, in the *Streamlit* framework. This technology was specifically chosen to offer the user an intuitive environment through which they can use the machine learning algorithms. The framework can provide alerts, notifications, allow user authentication under certain conditions, and include graphical or tabular representations for reporting. Being an open-source framework and able to cover numerous functional and non-functional requirements defined earlier, this development framework is more than suitable for implementing the user interface.

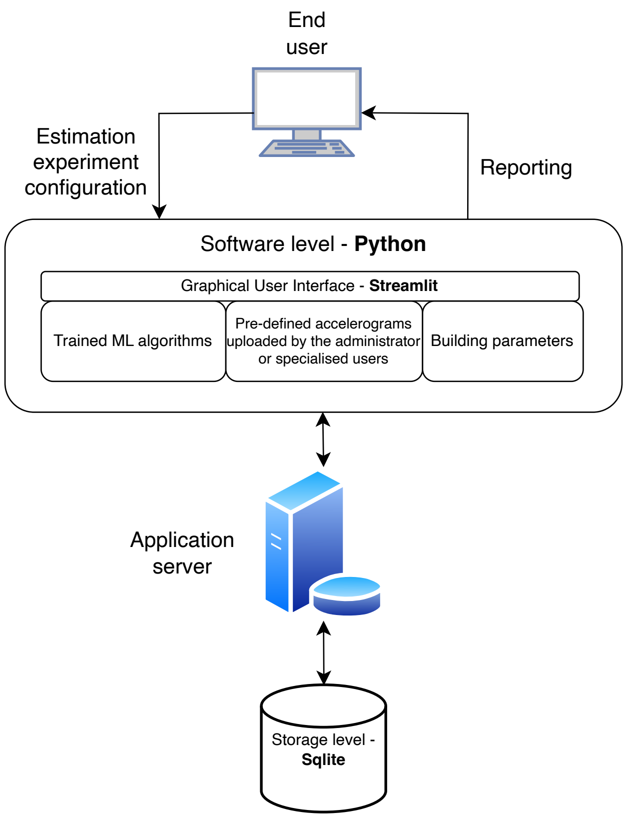


FIGURE 9. Proposed architecture for the degradation index estimation platform

4. Conclusions

In conclusion, the paper presents a system analysis conducted for developing a web platform for estimating the seismic degradation index of buildings. We went through the fundamental stages of the software product life cycle, including determining user profiles, identifying use cases, and developing functional and non-functional requirements. Through this analysis, we were able to outline the needs and expectations of end users and design an architecture that efficiently meets these requirements. The system analysis was essential to define the platform's functional requirements and establish the optimal architecture to guide subsequent implementation.

An important contribution of the paper is related to identifying and classifying users into four main categories: non-specialist end users, intermediate end users, specialist end users, and system administrator users and adapting the system analysis principles to design adequate requirements in the context depicted by these user profiles. Each user category has specific needs and interactions with the system, leading to the definition of detailed requirements for each use case.

Additionally, use case and sequence diagrams were developed to illustrate the interactions between users and the system in a clear and structured manner. The functional and non-functional requirements were presented using the SysML standard, highlighting the requirements associated with different use cases. These requirements ensure that the platform will meet all necessary functionalities to provide users with an optimal experience and facilitate the process of estimating the seismic degradation index.

Finally, the proposed architecture represents another important contribution. By using the system analysis approach, a robust architecture was determined, mitigating the risk of implementing a system that is inadequate for its targeted end-users.

Future work will involve the actual implementation of the proposed architecture, followed by testing and validating the system to ensure that all functional and non-functional requirements are met. Successful implementation of this platform will significantly contribute to improving the seismic resilience of buildings, offering a valuable tool for engineers and companies in designing safer and more resilient structures.

Acknowledgement

We want to express our gratitude to the "Academy of Romanian Scientists, Ilfov 3,050044 Bucharest, Romania" for funding this research work. We would also like express our gratitude to the National University of Science and Technology POLITEHNICA Bucharest (UNSTPB), Bucharest, Romania for funding this work.

REFERENCES

[1] Xuchuan Lin, Xueyan Liu, Jiang Hui, and Wenchen Shan. Assessment on detailed regional seismic damage risk of buildings based on time-history dynamic analyses. *Bulletin of Earthquake Engineering*, 22:1–21, 03 2024.

[2] M. Zuher, Ade Nasution, Zairah Sidiq, Masrilayanti Masrilayanti, and Jafril Tanjung. Fragility assesment of mid-rise rc building using hazus method in high seismic zone. *Jurnal Bangunan: Konstruksi & Desain*, 1:79–89, 08 2023.

[3] Qingle Cheng, Aiqun Li, Haotian Ren, Chea Por, Wenjie Liao, and Linlin Xie. Rapid seismic-damage assessment method for buildings on a regional scale based on spectrum-compatible data augmentation and deep learning. *Soil Dynamics and Earthquake Engineering*, 01 2024.

[4] Xiaoyan Song, Xiaowei Cheng, Yi Li, Guo Ruijie, Haoyou Zhang, Zihan Liang, and Senna Wang. A numerical model database for rapid seismic damage assessment of typical regular reinforced concrete frame structures in urban building clusters. *Journal of Building Engineering*, 90:109392, 08 2024.

[5] Alessandro Lubrano Lobianco, Marta Del Zoppo, and M. Di Ludovico. *Seismic Damage Assessment for RC Buildings from SHM Data*, pages 1366–1373. 06 2023.

[6] Young-Ji Park and Alfredo H.-S. Ang. Mechanistic seismic damage model for reinforced concrete. *Journal of Structural Engineering*, 111, 04 1985.

[7] S. K. Kunnath, a. M. Reinhorn, and R. F. Lobo. *IDARC Version 3.0: A Program for the Inelastic Damage Analysis of Reinforced Concrete Structures. Technical Report NCEER-92-0022*. U.S. National Center for Earthquake Engineering Research, 1992.

[8] George Bogdan Nica, Munteanu Ruben Iacob, Vasile Calofir, and Mihail Iancovici. Modelling nonlinear behavior of 3d frames using the force analogy method. *Structures*, 35:1162–1174, 2022.

[9] Munteanu Ruben Iacob, Florin Moţa, Vasile Calofir, and Cătălin Baciu. New approach to nonlinear dynamic analysis of reinforced concrete 3d frames; an accurate and computational efficient mathematical model. *Applied Sciences*, 12(3), 2022.

[10] Munteanu Ruben Iacob, Enache Ruxandra, Baciu Cătălin, and Calofir Vasile. A new perspective into torsional inelastic response of actively controlled irregular multistorey buildings. *Alexandria Engineering Journal*, 71:691–706, 2023.

[11] Vasile Calofir, Ruben-Iacob Munteanu, Mircea-Stefan Simoiu, and Karol-Cristian Lemnaru. Innovative approach to estimate structural damage using linear regression and K-nearest neighbors machine learning algorithms. *Results in Engineering*, 22:102250, June 2024.

[12] Adriana Olteanu, Anca Daniela Ioniţă, and Traian Ionescu. LEVERAGING OPEN SOURCE E-LEARNING SYSTEMS WITH WEB 2.0 AND KNOWLEDGE STRUCTURES. *U.P.B. Sci. Bull., Series C, Vol. 73, Iss. 2, 2011.*

[13] Mohammed Tali Almalchy, Sarmad Monadel Algayar, and Nirvana Popescu. HCare WEB APPLICATION FOR EHEALTH MONITORING SYSTEM. *U.P.B. Sci. Bull., Series C, Vol. 73, Iss. 2, 2011.*

[14] O B Alaba, Ioana Făgărăşan, and Radu Dobrescu. SYSTEM ANALYSIS FOR E-LEARNING GRIDS. *U.P.B. Sci. Bull., Series C, Vol. 71, Iss. 3, 2009.*

[15] Mircea Stefan Simoiu, Vasile Calofir, Ioana Fagarasan, and Cristina Nichiforov. eLearning Remote Simulator for Implementing Control Systems - A Case Study on a DC Motor. In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, pages 1062–1067, Metz, France, September 2019. IEEE.

[16] Calofir Vasile, Simoiu Mircea-Stefan, Fagarasan Ioana, and Sergiu Stelian Iliescu. Simulation Environment of a 3-Tank Control System for Educational Purposes. In *2020 European Control Conference (ECC)*, pages 1023–1028, Saint Petersburg, Russia, May 2020. IEEE.

[17] Benjamin S. Blanchard, editor. *Systems engineering and analysis*. Pearson, Essex, 5th ed edition, 2014.

[18] Charles S Wasson. System Engineering Analysis, Design, and Development.

[19] Sanford Friedenthal, Alan Moore, and Rick Steiner. *A practical guide to SysML: the systems modeling language*. Morgan Kaufmann, Waltham, MA, 2nd ed edition, 2012. OCLC: ocn754518532.

[20] Jon Holt and Simon Perry. *SysML for Systems Engineering*. Institution of Engineering and Technology, January 2008.

[21] Bruce Powel Douglass. *Agile model-based systems engineering cookbook: improve system development by applying proven recipes for effective agile systems engineering*. Packt, Birmingham, 2021. OCLC: 1261874525.