Use of machine learning algorithms for damage estimation of reinforced concrete buildings*

Swapnil Nayan and Pradeep Kumar Ramancharla**

Earthquake Engineering Research Centre, International Institute of Information Technology, Hyderabad, Telangana 500 032, India

Identifying the vulnerabilities in a building is a crucial step towards earthquake risk mitigation. Rapid visual screening is a quick and popular method for seismic vulnerability assessment. It helps identify buildings that require detailed investigation, which is done by modelling using seismic analysis software. This is a time-consuming and resource-intensive task. This article proposes the use of machine learning to bypass the seismic analysis of buildings. A case study using 1296 building models and maximum inter-storey drift ratio as the measure of damage has been presented. Random forest gives the best prediction accuracy in the study.

Keywords: Damage estimation, earthquakes, machine learning, rapid visual screening, reinforced concrete building.

ABOUT 60% of the land mass in India is part of the moderate to severe earthquake-prone areas, where over 80% of the population lives[1]. Combining this with poor construction and maintenance practices of concrete structures leads to loss of life and property, which was evident in the 2001 Bhuj earthquake that caused around 14,000 causalities[2]. Hence, identifying the vulnerabilities of a building is of utmost importance for earthquake preparedness.

For this, a three-tier approach is followed in most countries consisting of the following. Phase 1: Rapid visual screening (RVS). Phase 2: Preliminary assessment of buildings. Phase 3: Detailed seismic evaluation of the selected buildings.

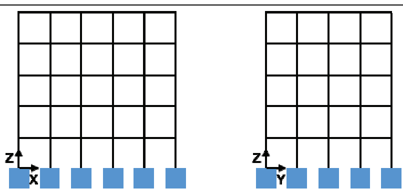
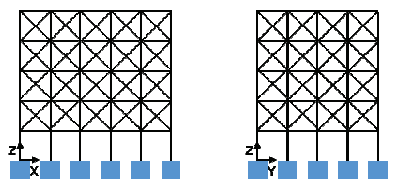
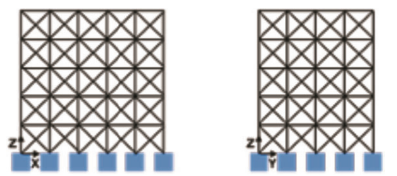
Phase 1 is a quick way of assessing the vulnerabilities of a building. It requires an experienced screener to visually inspect the buildings. Though it is a quick process, selection of RVS forms plays an important role in determining the vulnerability of buildings, as each area requires a different RVS form and each form can give different weightages to a particular vulnerability. Attempts to improve RVS forms like the one involving division and weighing of life-threatening factors and economic loss-inducing factors to arrive at a composite score have been made[3]. However, RVS forms are limited by the area for which they have been designed and hence a global vulnerability index is difficult to achieve. Preliminary and detailed evaluation is required, when there is a need for in-depth evaluation of buildings.

Detailed evaluation is (i) complex, as it requires collection of blueprints of building, identifying sizes of different members, load calculation and strength-related checks[2], (ii) resource intensive, requiring up-to-date computers and (iii) expensive, as it requires a lot of manpower to model and analyse all the buildings.

Machine learning (ML) has been used to solve many complex problems in multiple disciplines such as earthquake prediction, prediction of compressive strength of concrete[4,5], and damage estimation in buildings using image data[6–8].

Although attempts to tackle the problem of building damage assessment have been made[9,10], they are restricted in scope as they rely on data of a single earthquake and are tailored specifically to the buildings surveyed in the

Table 1. Types of structures

Structure type	Diagram
Bare frame	
Open ground storey	
Fully braced	

*The data and material, and code used in this study will be made available on request.

**For correspondence. (e-mail: swapnil.nayan@research.iiit.ac.in)

Table 2. Parameters used to make building models

Parameters	Minimum value	Maximum value	
Number of storeys	Two storeys	Ten storeys	
Number of bays in the X-direction	Two bays	Five bays	
Number of bays in the Y-direction	Two bays	Five bays	
Types of structures	Fully braced structures	Open ground-storey structures	Bare frame structures
Ratio of area moment of inertia of the column to that of beam (I_c/I_b)	0.75	1	1.25

Table 3. Building components dimension

No. of storeys	Column dimension (mm)	Beam dimension (mm)	I_c/I_b
2–5	230 × 230	245 × 245	0.75
6	245 × 245	265 × 265	0.75
7	270 × 270	290 × 290	0.75
8	290 × 290	310 × 310	0.75
9	325 × 325	350 × 350	0.75
10	370 × 370	400 × 400	0.75
2–5	230 × 230	230 × 230	1.00
6	240 × 240	240 × 240	1.00
7	270 × 270	270 × 270	1.00
8	290 × 290	290 × 290	1.00
9	325 × 325	325 × 325	1.00
10	355 × 355	355 × 355	1.00
2–5	230 × 230	220 × 220	1.25
6	235 × 235	220 × 220	1.25
7	265 × 265	250 × 250	1.25
8	290 × 290	275 × 275	1.25
9	310 × 310	295 × 295	1.25
10	345 × 345	325 × 325	1.25

Table 4. Damage state of buildings

Damage state	Maximum inter-storey drift ratio (%)
1	≤0.4
2	0.4–1
3	1–2
4	2–3
5	3–4
6	>0.4

affected area. This makes it difficult to use for any future earthquake events or events in other areas. The study by Chaurasia *et al.*[11], where neural network and Random Forest algorithms have been compared to obtain damages in buildings surveyed during the Gorkha earthquake in Nepal, is also plagued by the same problem.

Here, we identify vulnerable structures before an earthquake occurs using ML. For this study, maximum inter-storey drift ratio (MISDR), which is the ratio of the difference between the roof displacement of a storey and the storey below it to the height of the storey, has been divided into different damage states and is used as a measure of the amount of damage sustained by the structures. Logistic regression (log R), k-nearest neighbours (KNN), support vector machine (SVM), Naïve bayes (NB), decision tree classification (DTC) and Random Forest Classification (RFC) are the classification algorithms, and decision tree regression (DTR), linear regression (LR), polynomial regression (PolR), support vector regression (SVR) and random forest regression (RFR) are the regression algorithms that have been used to predict the damage state of buildings for Bhuj and Chamoli ground motions[12].

Dataset

One thousand two hundred and ninety-six different building models were designed and analysed using SAP2000 (ref. 13). Each building model belongs to one of the types mentioned in Table 1 and have been obtained using every permutation of the parameters listed in Table 2. Time-history (TH) analysis was performed on these models to obtain MISDR, which was then used for training and testing.

Modelling

All the buildings have been designed for gravity loads according to IS 456: 2000 (ref. 14). Table 3 lists the dimensions of beams and columns used in different types of models. Slab of thickness 150 mm is present at every storey level. The base of all the models is fixed. Diagonal struts have been designed according to IS 1893 (Part 1): 2016 (ref. 15). Beams and columns are made using M25-grade concrete and the grade of steel is HYSD 415. The length of the bay in each direction as well as the height of each storey is 3 m.

Methodology

Two approaches for damage estimation of reinforced concrete (RC) buildings during an earthquake have been discussed. Damage of building during an earthquake is defined by the damage state in which the building lies when the ground motion (GM) acts on it (Table 4).

FEMA 356, broadly divides the damage states of buildings into collapse prevention, life safety and immediate

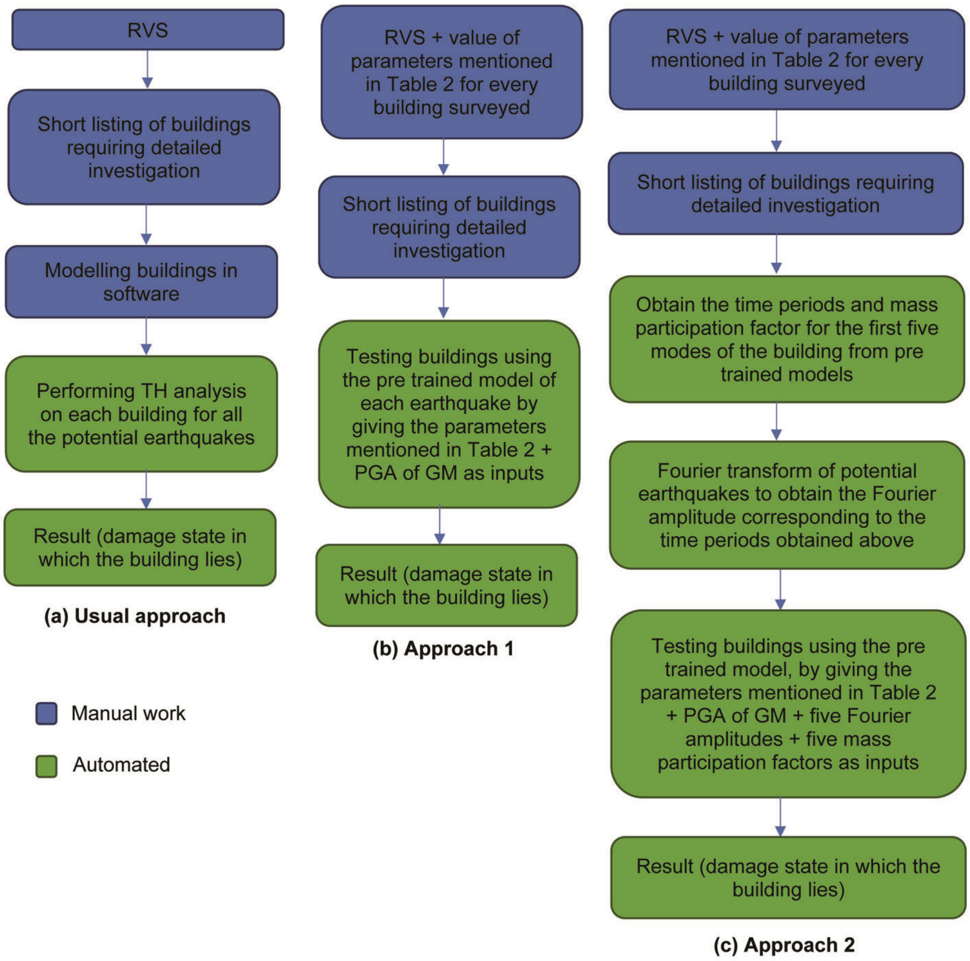


Figure 1. Approaches to estimate building damage.

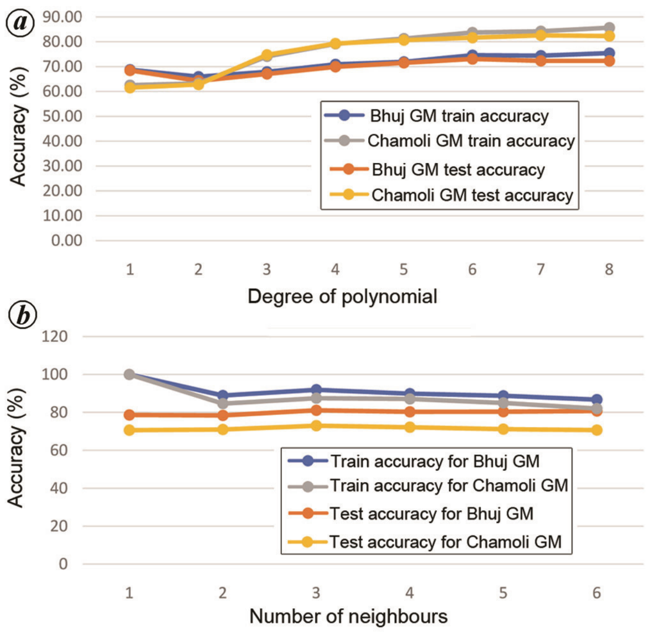


Figure 2. Parameter selection for (***a***) PolR: accuracy versus degree of polynomial and (***b***) KNN: accuracy versus no. of neighbours.

occupancy having drift limitations of 4%, 2% and 1% respectively[16]. Taking this into consideration, the damage states in this study are obtained by dividing the MISDR into six states as described in Table 4. This is done to get a more accurate prediction about the damage state in which the building lies.

Dataset is generated by carrying out linear TH analysis on 1296 building models using Bhuj and Chamoli GMs. TH analysis is carried out on each building model for both earthquakes by varying the peak ground acceleration (PGA) from 0.1 g to 1 g at an interval of 0.1 g. This creates a total of 12,960 data points corresponding to each earthquake. These data points are divided into training and testing datasets in 80 : 20 ratio.

In approach 1, the five variables mentioned in Table 2 along with PGA of GMs are the input to the model. In this approach, a separate model is required for every GM being considered.

For the second approach, modal analysis of each model is carried out. The time period and modal mass participation factors for the first five modes obtained from the modal analysis are fitted with the regression models. This

Table 5. Parameters used in machine learning (ML) models

Model	Parameters for Bhuj GM	Parameters for Chamoli GM	Parameters for combined data
LR	Loss function: Residual sum of squares	Loss function: Residual sum of squares	N/A
PolR	Loss function: Residual sum of squares Degree of polynomial: 6 Parameter selection: Backward elimination with significance level 5%	Loss function: Residual sum of squares Degree of polynomial: 7 Parameter selection: Backward elimination with significance level 5%	N/A
SVR (epsilon-SVR)	Kernel: Sigmoid Regularization parameter (*C*): 50 Regularization penalty: 12 Epsilon: 0.2 Kernel coefficient (gamma): 0.1 Independent term: –2 Parameter selection: Grid Search with mean squared error scoring	Kernel: Sigmoid C: 14 Regularization penalty: 12 Epsilon: 0.1 Gamma: 0.1 Independent term: –3 Parameter selection: Grid Search with mean squared error scoring	N/A
DTR	Loss function: Mean squared error Maximum depth: 195 Parameter selection: Grid Search with mean squared error scoring	Loss function: Mean squared error Maximum depth: 19 Parameter selection: Grid Search with mean squared error scoring	Loss function: Mean squared error Maximum depth: 70 Parameter selection: Grid Search with mean squared error scoring
RFR	Loss function: Mean squared error Number of estimators: 1000 Parameter selection: Grid Search with mean squared error scoring	Loss function: Mean squared error Number of estimators: 1000 Parameter selection: Grid Search with mean squared error scoring	Loss function: Mean squared error Number of estimators: 332 Parameter selection: Grid Search with mean squared error scoring
log*R*	Solver: Newton-cg Penalty: 12 C: 1 Iterations: 100	Solver: Newton-cg Penalty: 12 C: 1 Iterations: 100	N/A
KNN	Number of neighbours: 3 Distance metric used: Euclidean distance Parameter selection: Figure 2 *b*	Number of neighbours: 3 Distance metric used: Euclidean distance Parameter selection: Figure 2 *b*	N/A
NB	Distribution: Gaussian	Distribution: Gaussian	N/A
DTC	Criterion: Gini impurity Maximum depth: 18 Parameter selection: Grid Search with accuracy scoring	Criterion: Gini impurity Maximum depth: 19 Parameter selection: Grid Search with accuracy scoring	Criterion: Gini impurity Maximum depth: 25 Parameter selection: Grid Search with accuracy scoring
RFC	Criterion: Entropy Number of estimators: 162 Parameter selection: Grid Search with accuracy scoring	Criterion: Entropy Number of estimators: 177 Parameter selection: Grid Search with accuracy scoring	Criterion: Entropy Number of estimators: 280 Parameter selection: Grid Search with accuracy scoring
SVM	Kernel: rbf C: 100 Gamma: 0.026 Parameter selection: Grid Search with accuracy scoring	Kernel: rbf C: 101 Gamma: 0.026 Parameter selection: Grid Search with accuracy scoring	N/A

helps in the prediction of time periods and modal mass participation factor for a new building being tested. The first five modes are chosen as they give a combined mass participation factor of over 90%. Fourier transform is applied to each of the earthquake GMs to obtain the Fourier amplitudes corresponding to each of the time periods of the first five modes. The Fourier amplitudes along with mass participation factors are the additional variables in this model. This is done so that a single model is capable of predicting the damage state of building subjected to any GM. A total of 25,920 data points is available in this approach.

Figure 1 *a* shows the workflow required to get the damage state of a building in case ML is not used. Figure 1 *b* displays the workflow when approach 1 is used, while Figure 1 *c* shows the workflow when approach 2 is used for the same task. Five regression and six classification algorithms mentioned above have been fitted to the data in approach 1, while only DTC, RFC, DTR and RFR are fitted to the dataset in approach 2.

Table 5 shows the parameter used for various ML models. Parameters for PolR and KNN models have been obtained by plotting accuracy versus degree of polynomial (Figure 2 *a*) and accuracy versus number of neighbours

Table 6. Confusion matrix of each damage state

	Confusion matrix	State 1		State 2		State 3		State 4		State 5		State 6	
Bhuj GM													
LR	Train data	8,471	685	9,345	66	9,291	129	9,543	159	9,686	211	2,267	1,987
		314	898	921	36	886	62	646	20	455	16	15	6,099
	Test data	2,075	181	2,336	18	2,312	22	2,397	44	2,417	58	607	493
		80	256	228	10	244	14	147	4	115	2	2	1,490
PolR	Train data	8,729	427	8,897	514	8,744	676	9,303	399	9,554	343	3,979	275
		480	732	638	319	536	412	449	217	337	134	194	5,920
	Test data	2,131	125	2,232	122	2,136	198	2,327	114	2,394	81	1,042	58
		140	196	171	67	142	116	111	40	85	32	49	1,443
SVR	Train data	8,665	486	9,095	326	8,899	513	9,104	605	9,484	407	3,890	366
		264	953	765	182	655	301	458	201	374	103	187	5,925
	Test data	2,129	132	2,264	80	2,207	135	2,287	147	2,383	98	1,011	87
		68	263	198	50	170	80	111	47	82	29	50	1,444
DTR	Train data	9,151	0	9,421	0	9,421	0	9,709	0	9,891	0	4,256	0
		0	1,217	0	947	0	956	0	659	0	477	0	6,112
	Test data	2,232	29	2,260	84	2,258	84	2,350	84	2,430	51	1,042	56
		38	293	65	183	96	154	76	82	72	39	41	1,453
RFR	Train data	8,632	13	8,856	44	8,940	40	9,298	44	9,429	53	6,443	48
		32	1,691	32	1,436	32	1,356	55	971	72	814	19	3,858
	Test data	2,100	8	2,217	28	2,229	14	2,312	34	2,349	35	1,607	27
		20	464	16	331	25	324	27	219	44	164	14	944
log*R*	Train data	8,955	201	8,988	423	8,856	564	9,589	113	9,897	0	3,383	871
		168	1,044	326	631	474	474	629	37	471	0	104	6,010
	Test data	2,199	57	2,238	116	2,183	151	2,415	26	2,475	0	905	195
		53	283	93	145	113	145	148	3	117	0	21	1,471
KNN	Train data	9,096	60	9,279	132	9,061	359	9,549	153	9,870	27	4,150	104
		28	1,184	90	867	145	803	249	417	289	182	34	6,080
	Test data	2,208	48	2,253	101	2,207	127	2,339	102	2,439	36	1,025	75
		50	286	73	165	96	162	109	42	103	14	58	1,434
NB	Train data	6,960	2,196	9,411	0	9,420	0	9,702	0	9,897	0	3,286	968
		0	1,212	957	0	948	0	666	0	471	0	122	5,992
	Test data	1,680	576	2,354	0	2,334	0	2,441	0	2,475	0	883	217
		0	336	238	0	258	0	151	0	117	0	29	1,463
DTC	Train data	9,156	0	9,411	0	9,420	0	9,702	0	9,897	0	4,254	0
		0	1,212	0	957	0	948	0	666	0	471	0	6,114
	Test data	2,224	32	2,276	78	2,259	75	2,347	94	2,396	79	1,050	50
		40	296	68	170	90	168	75	76	76	41	59	1,433
RFC	Train data	9,156	0	9,411	0	9,420	0	9,702	0	9,897	0	4,254	0
		0	1,212	0	957	0	948	0	666	0	471	0	6,114
	Test data	2,238	18	2,304	50	2,265	69	2,353	88	2,430	45	1,044	56
		26	310	43	195	65	193	71	80	84	33	37	1,455
SVM	Train data	9,076	80	9,229	182	9,196	224	9,498	204	9,806	91	4,111	143
		79	1,133	146	811	176	772	197	469	234	237	92	6,022
	Test data	2,232	24	2,286	68	2,256	78	2,355	86	2,416	59	1,051	49
		33	303	52	186	72	186	74	77	85	32	48	1,444
Chamoli GM													
LR	Train data	8,176	465	8,694	203	8,522	464	8,657	672	8,427	1,083	5,465	1,012
		583	1,144	1,039	432	900	482	609	430	651	207	117	3,774
	Test data	1,986	126	2,197	51	2,128	109	2,164	195	2,094	262	1,387	261
		157	323	246	98	250	105	133	100	188	48	30	914
PolR	Train data	8,337	304	8,456	441	8,717	269	9,162	167	9,329	181	6,350	127
		298	1,429	501	970	231	1,151	171	868	212	646	76	3,815
	Test data	2,029	83	2,111	137	2,134	103	2,315	44	2,303	53	1,610	38
		97	383	154	190	66	289	57	176	60	176	24	920
SVR	Train data	8,450	195	8,575	325	8,669	311	9,068	274	9,129	353	6,245	246
		228	1,495	384	1,084	229	1,159	325	701	383	503	155	3,722
	Test data	2,072	36	2,145	100	2,169	74	2,289	57	2,304	80	1,565	69
		66	418	70	277	60	289	87	159	100	108	33	925
DTR	Train data	8,645	0	8,900	0	8,980	0	9,342	0	9,482	0	6,491	0
		0	1,723	0	1,468	0	1,388	0	1,026	0	886	0	3,877

(*Contd*)

Table 6. (*Contd*)

	Confusion matrix	State 1		State 2		State 3		State 4		State 5		State 6	
	Test data	2,098	10	2,217	28	2,226	17	2,319	27	2,344	40	1,613	21
		15	469	16	331	24	325	31	215	36	172	21	937
RFR	Train data	8,635	10	8,859	41	8,936	44	9,295	47	9,430	52	6,443	48
		28	1,695	33	1,435	34	1,354	55	971	74	812	18	3,859
	Test data	2,101	7	2,215	30	2,228	15	2,313	33	2,347	37	1,604	30
		22	462	15	332	25	324	30	216	46	162	14	944
log*R*	Train data	8,418	223	8,197	700	8,249	737	8,855	474	9,400	110	5,808	669
		141	1,586	541	930	702	680	646	393	768	90	115	3,776
	Test data	2,063	49	2,039	209	2,042	195	2,236	123	2,332	24	1,481	167
		38	442	129	215	204	151	157	76	208	28	31	913
KNN	Train data	8,589	52	8,606	291	8,523	463	8,966	363	9,446	64	6,418	59
		31	1,696	123	1,348	291	1,091	350	689	449	409	48	3,843
	Test data	2,088	24	2,097	151	2,027	210	2,202	157	2,287	69	1,559	89
		39	441	104	240	156	199	146	87	194	42	61	883
NB	Train data	6,953	1,688	8,398	499	8,282	704	8,965	364	9,477	33	5,867	610
		7	1,720	1,278	184	874	508	757	282	835	23	138	3,753
	Test data	1,677	435	2,125	123	2,053	184	2,264	95	2,351	5	1,487	161
		3	477	309	35	251	104	174	59	230	6	36	908
DTC	Train data	8,641	0	8,897	0	8,986	0	9,329	0	9,510	0	6,477	0
		0	1,727	0	1,471	0	1,382	0	1,039	0	858	0	3,891
	Test data	2,104	8	2,217	31	2,199	38	2,324	35	2,310	46	1,629	19
		12	468	25	319	31	324	45	188	41	195	23	921
RFC	Train data	8,641	0	8,897	0	8,986	0	9,329	0	9,510	0	6,477	0
		0	1,727	0	1,471	0	1,382	0	1,039	0	858	0	3,891
	Test data	2,106	6	2,201	47	2,198	39	2,299	60	2,311	45	1,589	59
		27	453	24	320	37	318	47	186	99	137	22	922
SVM	Train data	8,560	81	8,756	141	8,909	77	9,145	184	9,373	137	6,317	160
		91	1,636	121	1,350	127	1,255	118	921	240	618	83	3,808
	Test data	2,092	20	2,178	70	2,183	54	2,273	86	2,298	58	1,595	53
		40	440	49	295	65	290	61	172	95	141	31	913
Combined data													
DTR	Train data	17,698	0	18,320	0	18,415	0	19,075	0	19,388	0	10,784	0
		0	3,038	0	2,416	0	2,321	0	1,661	0	1,348	0	9,952
	Test data	4,432	35	4,511	79	4,461	101	4,651	105	4,756	94	2,626	69
		39	678	74	520	91	531	109	319	111	223	50	2,430
DTC	Train data	17,732	0	18,299	0	18,385	0	19,051	0	19,405	0	10,808	0
		0	3,004	0	2,437	0	2,351	0	1,685	0	1,331	0	9,928
	Test data	4,400	33	4,528	83	4,521	71	4,647	133	4,724	109	2,598	73
		39	712	65	508	107	485	102	302	123	228	66	2,447
RFR	Train data	17,672	26	18,211	109	18,299	116	18,932	143	19,238	150	10,658	126
		30	3,008	49	2,367	53	2,268	59	1,602	129	1,219	54	9,898
	Test data	4,455	12	4,512	78	4,482	80	4,652	104	4,776	74	2,621	74
		32	685	16	578	22	600	18	410	17	317	14	2,475
RFC	Train data	17,698	0	18,320	0	18,415	0	19,075	0	19,388	0	10,784	0
		0	3,038	0	2,416	0	2,321	0	1,661	0	1,348	0	9,952
	Test data	4,432	35	4,506	84	4,458	104	4,647	109	4,768	82	2,591	104
		34	683	75	519	104	518	122	306	147	187	36	2,453

(Figure 2 *b*) respectively. The value of degree of polynomial in PolR and number of neighbours in KNN after which the test accuracy decreases even when the train accuracy increases, i.e. indicating overfitting, has been chosen as the best fitting parameter in the respective models.

For fitting the regression algorithms to the data, the actual value of MISDR obtained from TH analysis is used for fitting and predictions. Then these values are converted into the damage states to obtain prediction accuracies. For the classification algorithms MISDR of the entire dataset is converted into damage states. This is used instead of the actual MISDR for fitting and prediction.

The accuracies of prediction for each model can be calculated using eq. (1) below.

$$\text{Accuracy} = \frac{\text{Correctly classified data points}}{\text{Total number of data points}}. \tag{1}$$

The accuracies for models have been obtained by ten-fold cross-validation. Tables 6 *a–c* shows the confusion matrices for Bhuj GM, Chamoli GM and combined data. Each element of the confusion matrix contains four numbers. The top left, top right, bottom left, bottom right numbers represent the number of true negative, false positive, false negative and true positive respectively.

Accuracy can also be obtained by adding all the true positive numbers for a particular model and dividing it by the total number of data points, which can be obtained by adding the four numbers of a cell.

For example, in case of state 1 train data of LR in Bhuj GM, true negative = 8471, false positive = 685, false negative = 314, true positive = 898 and total data points = 10,368. Similarly, true positive values for states 2, 3, 4, 5 and 6 are 36, 62, 20, 16 and 6099 respectively.

Hence, accuracy = (898 + 36 + 62 + 20 + 16 + 6099)/ 10,368 = 68.78%.

Time-history analysis

TH analysis provides dynamic structural response under loading which varies with time. TH analysis has been carried out in SAP2000, and Figures 3 and 4 show the results. The 1296 building models consist of 432 buildings in each type of structure, i.e. open ground storey, bare frame and fully braced structures. These 432 models in each type consist of 48 structures each of two storeys, three storeys up to ten storeys. It is evident from Figure 3 that when all the structures are considered, about 59% and 37% of the buildings are in damage state 6 when Bhuj and Chamoli GMs act respectively.

Fully braced buildings perform significantly better than the other types of structures. Only a minuscule percentage of buildings go into damage states above 3 for both GMs. While no building was present in damage state 1 or 2 when Bhuj GM acted on open ground-storey buildings, 63% of the buildings were in damage states 1 and 2 when the structure was fully braced.

Bare frame structures perform worse than fully braced buildings, but are better than open ground-storey structures. Only 44% of buildings go into damage state 6 as opposed to 68% in the case of buildings with an open ground storey when Chamoli GM acts on them. Majority of the buildings up to six storeys stay in damage state 1 and only a small percentage of structures go into damage state 4 or above. However, in the case of open ground-storey structures, majority of the buildings are present in damage state 6.

Bhuj earthquake appears to be more detrimental for buildings compared to Chamoli GM. From Figure 3, it can be observed that in fully braced buildings, a pattern similar to that observed in Chamoli GM is followed, with the exception of some buildings that do go into damage states 5 and 6. Also, 80% of open ground-storey buildings in each case other than the two-storey structures go into damage state 6. Similar observations can be made for bare frame buildings.

Results and observations

The usual approach of tackling the problem requires manual work in modelling buildings using software, which requires significant amount of time and computing power.

We tackled these problems using two approaches which give instant results. Only a few additional parameters mentioned in Figure 1 *a* and *b* are required to obtain the results.

From TH analysis, it can be observed that fully braced structures perform significantly better than bare frame structures, which marginally outperform structures with an open ground-storey. This can be attributed to the fact that open ground-storey buildings suffer a drop in stiffness and strength at the ground-storey level, which dissipates most of the energy and causes damage, while fully braced structures have similar strengths and stiffness at each storey level.

Figure 5 indicates the relative importance of different parameters in influencing the damage state prediction of structures. For individual GM data, the type of structure, PGA and number of storeys are more important to obtain the results. Whereas, in a combined dataset, Fourier amplitude, PGA and type of structure are more useful to determine the results.

Figure 6 *a* and *b* shows the train and test accuracies obtained by ten-fold cross-validation for models using approach 1. In both cases, DTC, RFC, DTR and RFR perform better than the other algorithms, with the train and test accuracies being close to each other in case of RFR.

Hence for approach 2, only DTC, RFC, DTR and RFR were used. Figure 6 *c* presents the results for approach 2. While approach 2 is more beneficial for the problem, RFR yields good results in both approaches.

The accuracy calculated from the confusion matrix may vary slightly from that mentioned in Figure 6 as it is created by taking a random distribution of data and not from ten-fold cross validation.

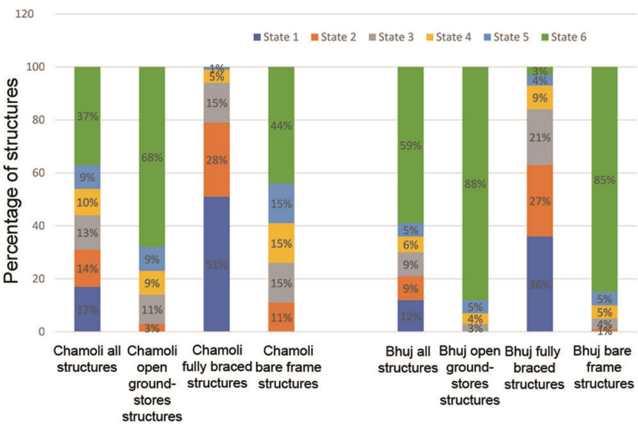


Figure 3. Damage state distribution of structures for Bhuj and Chamoli ground motions.

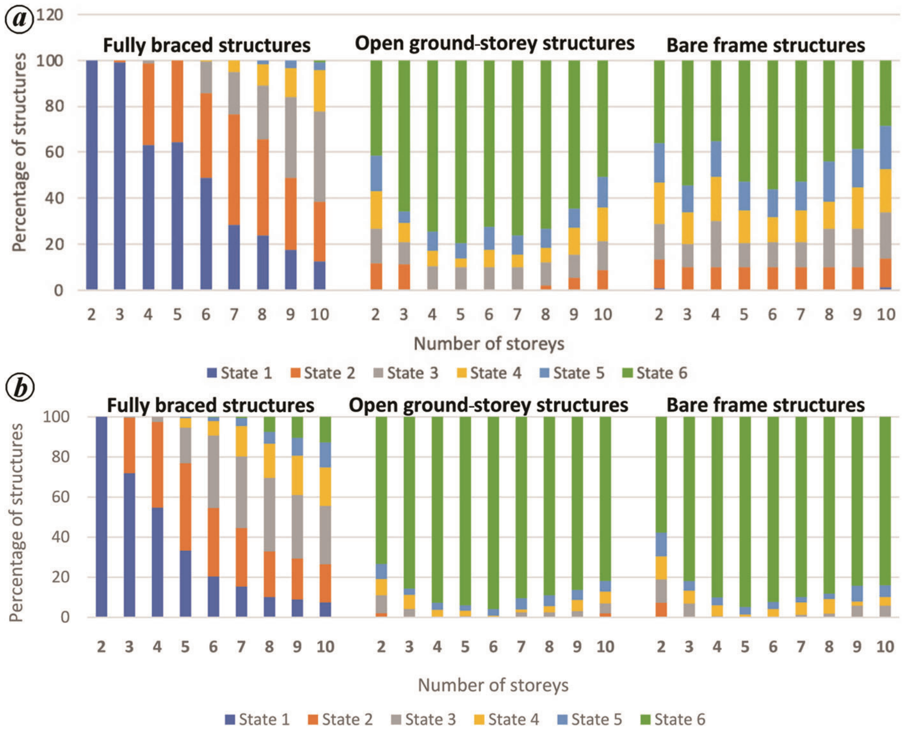


Figure 4. Storey-wise damage state distribution of buildings. ***a***, Chamoli GM; ***b***, Bhuj GM.

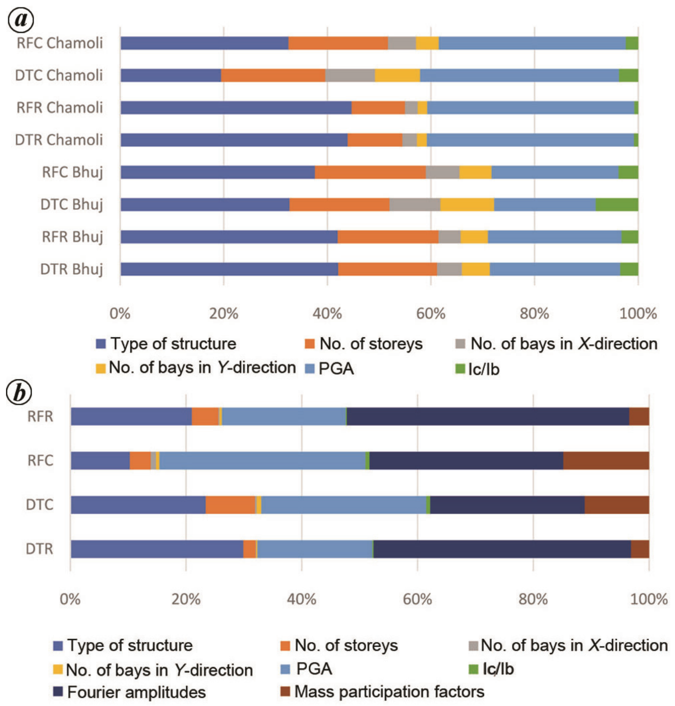


Figure 5. Relative importance of parameters used in various models. ***a***, Bhuj and Chamoli GM datasets; ***b***, Combined dataset.

Conclusion

Damage estimation of buildings due to earthquakes is important to identify the vulnerabilities and take appropriate retrofitting measures to mitigate them. ML has the ability to identify these risks more accurately and efficiently than most RVS methodologies which rely solely on regression analysis. Previous studies for damage estimation

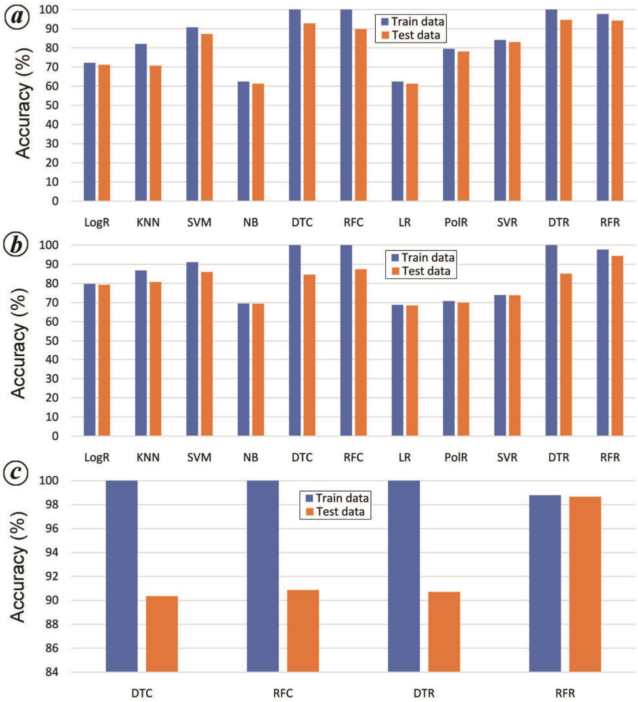


Figure 6. Accuracy of forecasting the damage state by various machine learning algorithms. ***a***, Approach 1, Chamoli GM; ***b***, Approach 1, Bhuj GM; ***c***, Approach 2, combined data.

using ML were limited in scope due to the use of data from earthquake events that had spatial and temporal variations. This meant that the results from one event were not comparable to another event. To overcome this problem, a dataset of 1296 buildings was developed and tested to find best-solution approaches.

RFR with the second approach was found to be an effective way to estimate damage to buildings during earthquakes. Fourier amplitude, PGA and type of structure played a significant role in determining the damage state of buildings. Fully braced structures performed much better than open ground-storey or bare frame structures. This study can be extended by further increasing the complexity of the dataset and testing other ML algorithms.

Conflict of interest: The authors declare that there is no conflict of interest.

1. Pradeep, K. R. and Murty, C. V. R., Earthquake safety of houses in India: understanding the bottlenecks in implementation. *Indian Concr. J.*, 2014, **88**(9), 51–63.
2. Srikanth, T. *et al.*, Earthquake vulnerability assessment of existing buildings in Gandhidham and Adipur cities, Kachchh, Gujarat (India). *Eur. J. Sci. Res.*, 2010, **41**(3), 336–353.
3. Murty, C. V. R. *et al.*, A methodology for documenting housing typologies in the moderate–severe seismic zones. In 15th World Conference on Earthquake Engineering, Lisbon, Portugal, September 2012.
4. Chou, J.-S. *et al.*, Machine learning in concrete strength simulations: multi-nation data analytics. *Construct. Build. Mater.*, 2014, **73**, 771–780; doi:https://doi.org/10.1016/j.conbuildmat.2014.09.054.
5. Chopra, P. *et al.*, Comparison of machine learning techniques for the prediction of compressive strength of concrete. *Adv. Civil Eng.*, 2018, **3**, 1–9; doi:10.1155/2018/5481705.
6. Nia, K. R. and Mori, G., Building damage assessment using deep learning and ground level image data. In 14th Conference on Computer and Robot Vision (CRV), Edmonton, AB, Canada, 2017, pp. 95–102; doi:10.1109/CRV.2017.54.
7. Nex, F. *et al.*, Structural building damage detection with deep learning: assessment of a state-of-the-art CNN in operational conditions. *Remote Sensing*, 2019, **11**(23), 2765; doi:10.3390/rs11232765.
8. Naito, S. *et al.*, Building-damage detection method based on machine learning utilizing aerial photographs of the Kumamoto earthquake. *Earthq. Spectra*, 2020, **36**(3), 1166–1187; https://doi.org/10.1177/8755293019901309.
9. Sujith, M. *et al.*, Classifying earthquake damage to buildings using machine learning. *Earthq. Spectra*, 2020, **36**(3), 1166–1187; doi:10.1177/8755293019878137.
10. Sreerama, A. K., Chenna, R., Mishra, S., Ramancharla, P. and Karanth, A., Rapid visual screening of different housing typologies in Himachal Pradesh, India. *Nat. Hazard.*, 2016, **85**, 1851–1875; doi:10.1007/s11069-016-2668-3.
11. Chaurasia, K. *et al.*, Predicting damage to buildings caused by earthquakes using machine learning techniques. In IEEE 9th International Conference on Advanced Computing, December 2019, pp. 81–86; doi:10.11109/1ACC48062.2019.8971453.
12. Bhuj and Chamoli ground motion record, Cosmos Strong Motion Virtual Data Center (VDC); https://www.strongmotioncenter.org/
13. CSI, *SAP2000 Integrated Software for Structural Analysis and Design*, Computers and Structures Inc., Berkeley, California, USA.
14. IS 456: 2000, Plain and reinforced concrete – code of practice, Fourth revision.
15. IS 1893 (Part 1), Criteria for earthquake resistant design of structures. Part 1 General Provisions and Buildings, Fifth revision, 2016.
16. Structural Engineering Institute, American Society of Civil Engineers, Prestandard and commentary for the seismic rehabilitation of buildings. Federal Emergency Management Agency, Washington DC, USA, 2000.

Received 14 October 2020; revised accepted 7 January 2022

doi: 10.18520/cs/v122/i4/439-447