Volume 4, Issue 8, August 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Wrapper Subset Evaluation Facilitates the Automated Detection of the Ground-Motion Intensity Measures and Derivation of the Seismic Fragility Curves

Abdulhameed Abdullah Yaseen* School of Civil Engineering and Surveying, Faculty of Technology, University of Portsmouth

David Begg School of Civil Engineering and Surveying, Faculty of Technology, University of Portsmouth

Nikos Nanos School of Civil Engineering and Surveying, Faculty of Technology, University of Portsmouth

Abstract—Fragility analysis and its graphical representation in terms of fragility curve is an effective tool for seismic risk assessment of structural systems. Fragility curve is a statistical tool representing the probability of exceeding a given damage state as a function of a ground-motion intensity measure (IM) that represents the ground motion. One of the most important issues in deriving fragility curves is the large number of IMs that have been proposed by researchers over the years. Any improvement in the proper selection of the IM would therefore represent a significant gain with respect to accurately predicting the structural seismic responses and mitigating the economic impacts of earthquake disasters besides reducing the loss of lives. In this study we apply automated machine learning to analyse IMs and assessing the seismic responses of two typical one- and two-storey unreinforced masonry (URM) buildings located in the Kurdistan region of Iraq. By applying wrapper feature selection method and using several classifier algorithms it was able to select the most relevant IMs to use as input to a fragility analysis tool. Furthermore, results suggest that the prediction of the failure pattern of buildings is also feasible using the wrapper method.

Keywords— Feature selection, Wrapper method, Weka, Fragility analysis, Unreinforced masonry buildings.

I. INTRODUCTION

One of the important recent tools in seismic vulnerability assessment of structures is fragility analysis tool [1]. Fragility analysis allows engineers to monitor the damage probability of a structure with respect to a seismic intensity measure (IM). Such an approach increases the possibility of failure classification and diagnosis before an earthquake happen. However, many factors may affect the outcome of the analysis. One of them is the large number of IMs proposed by researchers (e.g., [2], [3], [4]) to represent seismic action in a region of interest. Some of them may be irrelevant to fragility analysis; because only one ground-motion parameter is usually used by this method in order to derive fragility curves. Thus, selecting discriminatory IMs is critical to improving the accuracy and speed of prediction systems.

This paper presents wrapper feature selection method to select the best subsets of IMs. After IMs are selected by this method, its effectiveness is investigated by comparing error rate of seven traditional classification algorithms applied to only these selected IMs versus all IMs. The seven classification algorithms are Naive Bayes (NB), Nearest Neighbors (k-NN), logistic, multilayer perceptron neural network (MLP), decision table, and decision tree algorithms.

The organization of this paper is as follows. The two data sets used in this study are described in Section 2. The feature selection method and classifier algorithms are briefly described in Section 3. The results and discussions are reported in Section 4. Then conclusion and recommendations for future work are provided in Section 5.

II. FOUR SETS OF DATA

A total of 1168 time history analyses (using incremental dynamic analysis) are applied to the base of two typical one- and two-storey URM buildings located in the Kurdistan region (KR) of Iraq. These types of buildings constitute approximately 87% of the buildings in the region [5]. The KR is situated in the north and northeastern region of Iraq and is considered as the most seismic hazardous region in Iraq [6]. Based on the KR's seismological characteristics, two seismic hazard zones, referred to as 'B' and 'A' (representative of areas with 'high' to 'very high' levels of seismic hazards, respectively), can be defined for the region as shown in Fig. 1. Two sets of 35 and 38 time histories were selected to match the seismic characteristics of each seismic hazard zones A and B, respectively (Tables 1 and 2). These ground-motion time histories are furthermore scaled to eight levels of 0.02 g, 0.05 g, 0.1 g, 0.2 g, 0.4 g, 0.6 g, 0.8 g, and 1.0 g of peak ground acceleration (PGA). The scaled time histories are then applied to the base of two one- and two storey URM buildings located in two seismic hazard zones A and B, resulted in 1168 time history analyses and four sets of data. A total of 36 IMs were defined for each time history (Table 3). The dynamic response of the buildings was classified in terms of the damage state using the Milutinovic and Trendafiloski [7] criteria. However, the damage state

limits were reduced to two levels instead of five, as proposed in recent studies ([8], [9]), by merging the damage states from 'slight' to 'extensive' into 'yield' and the damage states of 'very heavy' and 'collapse' into 'collapse'. The collapse state is a level for which building repair is not practical or feasible, whereas in the yield state, the structure can still be used if suitable repairs are undertaken.

The numerical model was developed using the TREMURI code [10], which enables the representation of a complete three-dimensional (3D) model of the URM structures using an effective macro-element approach and demonstrates the nonlinear behavior of masonry panels and piers. The 3D view of the tested buildings is shown in Figure 1. The structures are 3.0 m high for the one-storey building and 6.0 m high for the two-storey buildings with plan dimensions of 15 m × 10 m. The mean values of the material properties for the masonry used in TREMURI are Young's modulus E=4,350 N/mm^2, shear modulus G=0.4 E, and specific weight=21 kN/m^3.

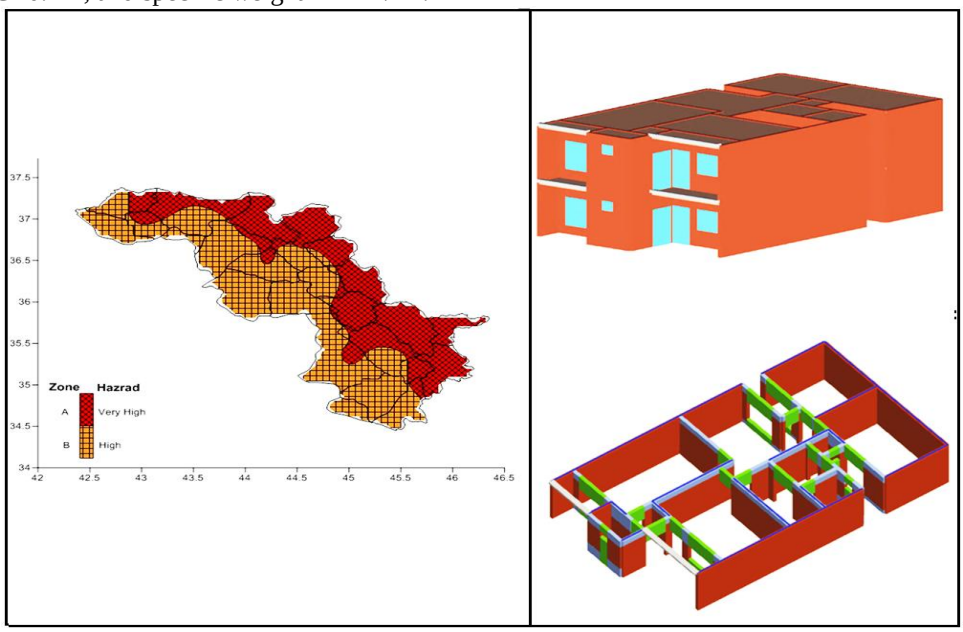

Fig. 1 Seismic hazard map of Kurdistan (proposed by authors) and 3D view of tested buildings using TREMURI software

TABLE I: The results of time history analyses applied to buildings in zone a using a set of 35 earthquake records

Earthquake name	Magnitude	Epicentral Distance (km)	PGA (g)	ASI (m/s)	VSI (cm)	One-storey		Two-storey	
						Disp. in X-direction	Damage state	Disp. in X-direction	Damage state
Gazli, USSR	6.8	12.8	0.60	4.71	230.18	0.94	Collapse	1.98	Collapse
Tabas, Iran	7.35	20.6	0.33	3.12	87.51	0.28	Yield	0.55	Yield
Tabas, Iran	7.35	55.2	0.84	8.02	339.33	1.81	Collapse	4.56	Collapse
Imperial Valley-06	6.53	28.7	0.27	2.76	151.34	0.17	Yield	0.76	Yield
Imperial Valley-06	6.53	27.8	0.52	4.14	178.16	0.44	Yield	1.23	Yield
Imperial Valley-06	6.53	28.1	0.60	3.67	196.44	0.45	Yield	1.17	Yield
San Salvador	5.8	7.9	0.88	5.96	240.75	1.5	Collapse	4.56	Collapse
Loma Prieta	6.93	18.5	0.97	5.51	430.27	2.1	Collapse	4.25	Collapse
Loma Prieta	6.93	27.2	0.51	2.98	191.26	0.42	Yield	1.36	Yield
Cape Mendocino	7.01	10.4	1.50	9.35	275.58	2.34	Collapse	4.26	Collapse
Cape Mendocino	7.01	4.5	0.59	3.57	197.92	0.72	Yield	2.41	Collapse
Landers	7.28	44.0	0.72	4.01	183.77	0.91	Collapse	4.12	Collapse
Northridge-01	6.69	40.7	0.57	5.14	212.36	0.85	Yield	2.15	Collapse
Northridge-01	6.69	13.0	0.57	4.91	244.04	0.98	Collapse	2.56	Collapse
Northridge-01	6.69	8.5	0.75	5.12	318.08	1.17	Collapse	4.86	Collapse

Northridge-01	6.69	13.6	0.83	5.65	306.81	1.39	Collapse	3.24	Collapse
Northridge-01	6.69	16.8	0.60	4.10	259.38	0.87	Yield	2.44	Collapse
Kobe, Japan	6.9	8.7	0.51	5.53	164.25	0.94	Collapse	1.89	Collapse
Chi-Chi, Taiwan	7.62	32.7	0.65	6.15	312.53	1.22	Collapse	3.22	Collapse
Chi-Chi, Taiwan	7.62	31.7	0.97	6.17	533.03	2.48	Collapse	6.53	Collapse
Chi-Chi, Taiwan	7.62	88.8	0.31	2.29	63.19	0.16	Yield	0.53	Yield
Chi-Chi, Taiwan	7.62	26.7	0.81	3.78	355.87	1.51	Collapse	3.98	Collapse
Chi-Chi, Taiwan	7.62	28.7	0.50	3.37	304.86	0.46	Yield	4.11	Collapse
Chi-Chi, Taiwan	7.62	47.9	0.57	4.49	311.32	0.73	Yield	6.83	Collapse
Chi-Chi, Taiwan	7.62	15.4	0.57	5.71	238.28	0.91	Collapse	2.28	Collapse
Chi-Chi, Taiwan	7.62	21.4	0.49	4.34	246.10	1.07	Collapse	2.48	Collapse
Chi-Chi, Taiwan	7.62	19.1	0.60	4.15	366.62	1.49	Collapse	3.34	Collapse
Chi-Chi, Taiwan	7.62	16.0	0.30	3.39	132.01	0.32	Yield	0.76	Yield
Chi-Chi, Taiwan	7.62	5.0	0.44	4.91	172.01	0.91	Collapse	1.61	Collapse
Chi-Chi, Taiwan	7.62	7.6	0.74	5.67	258.29	1.51	Collapse	3.35	Collapse
Chi-Chi, Taiwan	7.62	8.9	1.16	6.87	717.04	3	Collapse	5.87	Collapse
Chi-Chi, Taiwan	7.62	95.7	0.38	3.78	123.40	0.32	Yield	1.51	Collapse
Chi-Chi, Taiwan	7.62	14.2	1.00	6.75	225.10	1.48	Collapse	2.48	Collapse
Chi-Chi, Taiwan	7.62	14.2	0.96	6.23	203.42	1.38	Collapse	2.19	Collapse
Manjil, Iran	7.37	40.4	0.51	4.93	155.77	0.5	Yield	1.25	Yield

TABLE II: The results of time history analyses applied to buildings in zone b using a set of 38 earthquake records

Earthquake name	**Magnitude**	**Epicentral Distance (km)**	**PGA (g)**	**ASI (m/s)**	**VSI (cm)**	**One-storey**		**Two-storey**	
						Disp. in X-direction	**Damage state**	**Disp. in X-direction**	**Damage state**
Gazli, USSR	6.8	12.8	0.60	4.71	230.2	0.94	Collapse	1.98	Collapse
Imperial Valley-06	6.53	33.7	0.24	2.35	115.6	0.12	Yield	0.52	Yield
Imperial Valley-06	6.53	29.4	0.36	4.01	147.1	0.34	Yield	1.05	Yield
Imperial Valley-06	6.53	27.1	0.36	2.40	148.6	0.11	Yield	0.64	Yield
Imperial Valley-06	6.53	27.8	0.52	4.14	178.2	0.44	Yield	1.23	Yield
Imperial Valley-06	6.53	27.5	0.41	2.69	199.6	0.26	Yield	1.10	Yield
Imperial Valley-06	6.53	27.6	0.46	3.09	242.7	0.59	Yield	3.13	Collapse
Imperial Valley-06	6.53	28.1	0.60	3.67	196.4	0.45	Yield	1.17	Yield
Imperial Valley-06	6.53	27.2	0.35	3.44	151.9	0.35	Yield	0.90	Yield
San Salvador	5.8	7.9	0.88	5.96	240.8	1.50	Collapse	4.56	Collapse
Superstition Hills-02	6.54	16.0	0.46	3.33	406.4	1.46	Collapse	3.62	Collapse
Superstition Hills-02	6.54	29.4	0.18	1.44	102.3	0.03	None	0.17	Yield
Loma Prieta	6.93	18.5	0.97	5.51	430.3	2.10	Collapse	4.25	Collapse
Loma Prieta	6.93	27.2	0.51	2.98	191.3	0.42	Yield	1.36	Yield
Loma Prieta	6.93	27.1	0.2	2.42	175.3	0.07	Yield	0.50	Yield

			5						
Erzican, Turkey	6.69	9.0	0.50	3.94	224.7	0.64	Yield	2.70	Collapse
Cape Mendocino	7.01	4.5	0.59	3.57	197.9	0.72	Yield	2.41	Collapse
Landers	7.28	44.0	0.72	4.01	183.8	0.91	Collapse	4.12	Collapse
Northridge-01	6.69	13.4	0.42	3.47	266.3	0.76	Yield	2.57	Collapse
Northridge-01	6.69	4.9	0.36	3.39	149.6	0.32	Yield	0.79	Yield
Northridge-01	6.69	40.7	0.57	5.14	212.4	0.85	Yield	2.15	Collapse
Northridge-01	6.69	13.0	0.57	4.91	244.0	0.98	Collapse	2.56	Collapse
Northridge-01	6.69	11.8	0.51	3.19	245.2	0.68	Yield	2.60	Collapse
Northridge-01	6.69	20.3	0.58	6.28	240.7	1.18	Collapse	2.78	Collapse
Northridge-01	6.69	3.4	0.37	3.55	150.4	0.18	Yield	0.89	Yield
Northridge-01	6.69	13.6	0.83	5.65	306.8	1.39	Collapse	3.24	Collapse
Northridge-01	6.69	16.8	0.60	4.10	259.4	0.87	Yield	2.44	Collapse
Kobe, Japan	6.9	38.6	0.69	5.02	317.6	1.17	Collapse	2.75	Collapse
Kocaeli, Turkey	7.51	53.7	0.22	1.32	42.1	0.04	None	0.30	Yield
Chi-Chi, Taiwan	7.62	32.0	0.35	2.37	154.9	0.17	Yield	0.82	Yield
Chi-Chi, Taiwan	7.62	77.5	0.47	4.25	116.2	0.44	Yield	1.99	Collapse
Chi-Chi, Taiwan	7.62	38.9	0.29	2.14	114.3	0.12	Yield	0.46	Yield
Chi-Chi, Taiwan	7.62	39.6	0.35	2.06	318.5	0.14	Yield	2.23	Collapse
Chi-Chi, Taiwan	7.62	47.9	0.57	4.49	311.3	0.73	Yield	6.83	Collapse
Chi-Chi, Taiwan	7.62	21.4	0.49	4.34	246.1	1.07	Collapse	2.48	Collapse
Chi-Chi, Taiwan	7.62	20.7	0.33	3.04	170.8	0.15	Yield	1.23	Yield
Duzce, Turkey	7.14	41.3	0.73	6.32	235.0	1.17	Collapse	2.74	Collapse
Duzce, Turkey	7.14	1.6	0.35	4.31	173.5	0.50	Yield	1.64	Collapse

TABLE III: Ground-motion IMs considered in the current study.

IMs	Name
Acceleration-based	Peak ground acceleration (PGA), root mean square of acceleration (ARMS), Arias intensity (IA), characteristic intensity (IC), cumulative absolute velocity (CAV), acceleration spectrum intensity (ASI), sustained maximum acceleration (SMA), effective design acceleration (EDA), A95 parameter, and spectral acceleration at different periods Sa(nT1), where n=1, 2, 3, 4, 8, 16, 32 and T1 is the fundamental period of the structure
Velocity-based	Peak ground velocity (PGV), root mean square of velocity (VRMS), specific energy density (SED), velocity spectrum intensity (VSI), sustained maximum velocity (SMV), and Housner intensity (IH)
Displacement-based	Peak ground displacement (PGD), root mean square of displacement (DRMS), spectral displacement at different periods Sd(nT1), where n=1, 2, 3, 4, 8, 16, 32 and T1 is the fundamental period of the structure
Hybrid	PGV/PGA, PGA/PGV, and PGV^2/PGA

IMs	Name
Duration	Predominant period (Tp) and mean period (Tm)

Note: refer to Kramer [11] for the explicit explanation of the IMs examined.

III. FEATURE SELECTION METHOD AND CLASSIFICATION ALGORITHMS

A. Feature Selection Method

Feature selection is considered as an effective method for selecting a good subset of features by reducing dimensionality [12], removing irrelevant data, increasing learning accuracy, and improving the interpretability of results [13]. It can also display the relationship between features and identify the most important features in the prediction of results [14]. Four basic steps involved within a typical feature selection process are: subset generation, subset evaluation, stopping criterion, and result validation [15]. Through a search procedure, several candidate feature subsets are produced and based on a specific search strategy each candidate subset is evaluated and compared with the previous best one using a certain evaluation criterion. This process is repeated until a given stopping criterion is satisfied.

Feature selection methods can be classified into three main categories: wrapper methods, filter methods, and embedded methods [16]. The wrapper method requires one predefined learning algorithm and uses its performance as the evaluation criterion, whereas the filter technique only relies on general characteristics of the data without using mining algorithm. However, the wrapper model is more computationally expensive than the filter model ([13], [17]). The embedded model attempts to take advantage of the two aforementioned models and it is typically less complex than wrapper methods and more complex than filter methods [18]. Kohavi and John [13] recommended the use of the wrapper approach as it fits with the definition that the optimal features depend on the specific learning algorithm and the training set at hand. Since the fragility analysis is usually undertaken for a specific class of structures located in a specific region using a specific dataset of earthquake records, it is therefore decided to use the wrapper method in the present study.

B. Classification Algorithms

A number of automated classifier algorithms or learning models are available for wrapper feature selection method using the Weka (Waikato Environment for Knowledge Analysis) toolbox [19]. The considered classifiers in the study are briefly discussed below and include the Naive Bayes (NB) [20], Nearest Neighbors (k-NN) [21], logistic, Multilayer Perceptron neural network (MLP), decision table, and decision tree algorithms. Weka is a powerful open source Java-based machine learning software package developed at the University of Waikato in New Zealand. It is publicly available online at http://www.cs.waikato.ac.nz/ml/weka.

NB as a probabilistic learning algorithm that uses Bayes's rule is among the most practical approaches to certain types of learning problems [22].

In k-NN algorithm and based on Euclidean distance from nearest example in the training set, the unknown pattern of a new testing sample is assigned with the same class of that example. Usually, the most common class in the k closest neighbors is used, where k is a parameter set defined by the user. In this work, we use the algorithm with k=1 (known as 1-NN) and k=3 (known as 3-NN).

Le Cessie and van Houwelingen [23] algorithm with selected modifications is used by Weka to build a multinomial logistic regression model using a ridge estimator. The probability of an event occurrence is predicted by logistic regression and it is commonly used to model binary data.

The multilayer perceptron is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes (i.e. input, hidden, output), with each layer fully connected to the next one. A supervised learning technique called backpropagation is utilized by MLP for training the network. The MLP can discriminate data that are not linearly separable.

In decision table algorithm, the dataset is divided into cells and each cell contains identical examples. The unknown class for a new testing example is then assigned by the most frequent class in the cell [24].

The decision tree-simple CART [25] model is used in Weka as a classifier to build multivariate decision trees and remove the redundant attributes. The decision trees can clearly recognize the most important features for the prediction system and can classify unknown patterns rapidly [26]. However, as indicated by Perner [26] the construction of a decision tree can be affected by correlated and irrelevant attributes and as a result its performance degrades.

IV. RESULTS AND DISCUSSIONS

To reduce the necessary computing resources and speeding up the construction of the classification model in addition to avoiding bias in the results introduced by availability of a large number of irrelevant features, the VARCLUS procedure is used in the first step to merge the IMs within two clusters. Tanagra software [27] is used for that purpose and based on a succession of principal component analyses involved in the VARCLUS procedure the clusters are generated. The cluster that more closely correlates with the structural response is chosen for use in the wrapper subset evaluation procedure. The correlation coefficients between the top displacement and the two clusters are represented in Table 4 for each dataset. This first analysis demonstrates that the relevant ground-parameters should be selected from the first cluster, whereas cluster 2 is poorly correlated with the structural response in all cases. As shown, the spectral accelerations and displacements at different periods are poorly correlated with the damage measure even though this correlation should be superior to the other IMs, especially near the natural period of the buildings. This behavior can be explained by the increase in the natural period of the tested buildings due to a loss of rigidity and the progressive degradation.

For the four datasets, after carrying out the wrapper feature selection, the optimum feature subsets identified are summarized in Tables 5-8. To limit overfitting and reduce variability, 10 rounds of cross-validation are performed in the wrapper method for each classifier, and the validation results are averaged over the rounds. For most algorithms, high accuracy is achieved in the data classification. However, the smallest error rate in the prediction of the response of buildings located in zone B is achieved by the 1-NN, while for buildings in zone A, the 3-NN and NB were the best classifiers. The higher precision in the nearest neighbors indicates that this algorithm returned substantially more relevant results than the other classifiers. Furthermore, the wrapper method's effectiveness is clearly seen by comparing error rate of seven classification algorithms applied to the wrapper-based selected IMs versus all IMs in cluster1 for each data set (see Tables 5-8). Different feature sets produced by different classifier algorithms applied to the four different datasets and their success in the prediction of the results, support the findings of Kohavi and John [13], by showing that the attribute set chosen should be considered as part of the classifier algorithm.

To select only one IM which is required by fragility analysis and furthermore, to find which of the involved IMs has the greatest effect on the response of the structure, Receiver Operating Characteristics (ROC) analysis is applied to a subset with the minimum rate of error per each data set. Based on the ROC curves obtained from the ROC analysis, the satisfactory IMs can be determined if their curves are as close as possible to the (0, 1) corner. Alternatively, the IMs can be analysed by measuring the area under the ROC curve, which is commonly known as the AUC indicator. An AUC close to unity indicates an efficient model. The ROC curves plot the true positive rates (sensitivity) versus false positive rates (1-specificity) for a given classification procedure.

The ROC analysis (Figs 2 and 3) indicates that the ASI and VSI produced better results than the other IMs in classifying data based on the damage state of the one- and two- storey buildings, respectively. These parameters consider the spectral acceleration for ASI and the spectral velocity for VSI over a wide range of periods (e.g., 0.1-0.5 s for ASI and 0.1-2.5 s for VSI). Hence, this range can explain the increase in the natural period of the buildings due to a loss of rigidity and the progressive degradation produced in the buildings considered in this study.

TABLE IV: Clusters generated from the accelerogram dataset and the corresponding ground-motion parameters, the correlation coefficients between them, and the drift obtained from the simulations

Seismic hazard zone	Building	Cluster	Ground-motion parameter	Top displacement
A	One-storey	1	PGA, PGV, PGD, PGV/PGA, PGA/PGV, PGV^2/PGA, ARMS, VRMS, DRMS, IA, IC, SED, CAV, ASI, VSI, IH, SMA, SMV, EDA, A95, Tm	0.8842
		2	Sa T1, Sa 2T1, Sa 3T1, Sa 4T1, Sa 8T1, Sa 16T1, Sa 32T1, Sd T1, Sd 2T1, Sd 3T1, Sd 4T1, Sd 8T1, Sd 16T1, Sd 32T1, Tp	-0.0333
	Two-storey	1	PGA, PGV, PGD, PGV/PGA, PGA/PGV, PGV^2/PGA, ARMS, VRMS, DRMS, IA, IC, SED, CAV, ASI, VSI, IH, SMA, SMV, EDA, A95, Tm	0.9041
		2	Sa T1, Sa 2T1, Sa 3T1, Sa 4T1, Sa 8T1, Sa 16T1, Sa 32T1, Sd T1, Sd 2T1, Sd 3T1, Sd 4T1, Sd 8T1, Sd 16T1, Sd 32T1, Tp	0.0699
B	One-storey	1	PGA, PGV, PGD, PGV/PGA, PGA/PGV, PGV^2/PGA, ARMS , VRMS, DRMS, IA, IC, SED, CAV, ASI, VSI, IH, SMA, SMV, EDA, A95, Tp, Tm	0.8932
		2	Sa T1, Sa 2T1, Sa 3T1, Sa 4T1, Sa 8T1, Sa 16T1, Sa 32T1, Sd T1, Sd 2T1, Sd 3T1, Sd 4T1, Sd 8T1, Sd 16T1, Sd 32T1	0.1452
	Two-storey	1	PGA, PGV, PGD, PGV/PGA, PGA/PGV, PGV^2/PGA, VRMS, DRMS, IA, IC, SED, CAV, VSI, IH, SMA, SMV, EDA, A95, Tp, Tm	0.9284
		2	Sa T1, Sa 2T1, Sa 3T1, Sa 4T1, Sa 8T1, Sa 16T1, Sa 32T1, Sd T1, Sd 2T1, Sd 3T1, Sd 4T1, Sd 8T1, Sd 16T1, Sd 32T1, ARMS, ASI	0.1921

TABLE V: Results of the testing classifiers using the optimum feature subsets chosen for each classifier and the one-storey building dataset (38 records-zone b) using the wrapper method

Classifiers	Feature set	Precision	ROC Area	Error rate	Feature set	Precision	ROC Area	Error rate
1-NN	ARMS, VSI, IH, ASI, A95	0.959	0.95	4.1	Cluster 1	0.902	0.897	9.76
Trees (Simple CART)	VRMS, EDA, A95	0.957	0.912	4.3	Cluster 1	0.91	0.928	9.15
Logistic	PGA, PGA/PGV, ASI	0.945	0.979	5.5	Cluster 1	0.91	0.943	9.15

MLP	PGA, VSI, SMA, EDA	0.939	0.966	6.1	Cluster 1	0.921	0.966	7.93
3-NN	VRMS, IA, EDA	0.951	0.943	4.9	Cluster 1	0.939	0.942	6.1
NB	PGA/PGV, ASI, VSI, SMA, A95	0.936	0.98	6.7	Cluster 1	0.887	0.942	12.19
Decision table	PGA, VRMS, IC, EDA	0.948	0.949	5.5	Cluster 1	0.866	0.929	13.41

TABLE VI: Results of the testing classifiers using the optimum feature subsets chosen for each classifier and the two-storey building dataset (38 records-zone b) using the wrapper method

Classifiers	**Feature set**	**Precision**	**ROC Area**	**Error rate**	**Feature set**	**Precision**	**ROC Area**	**Error rate**
1-NN	IC, CAV, VSI	0.974	0.976	2.7	Cluster 1	0.914	0.9	8.6
Trees (Simple CART)	VSI	0.947	0.922	5.4	Cluster 1	0.935	0.928	6.45
Logistic	VSI, EDA	0.963	0.99	3.8	Cluster 1	0.919	0.937	8.06
MLP	PGA, VSI	0.0957	0.993	4.3	Cluster 1	0.919	0.981	8.06
3-NN	SED, VSI, EDA, A95	0.957	0.969	4.3	Cluster 1	0.93	0.965	6.98
NB	IA, CAV, VSI, IH, EDA	0.965	0.992	3.8	Cluster 1	0.919	0.977	9.1
Decision table	DRMS, IA, VSI, SMA	0.946	0.987	5.4	Cluster 1	0.93	0.98	6.98

TABLE VII: Results of the testing classifiers using the optimum feature subsets chosen for each classifier and the one-storey building dataset (35 records-zone a) using the wrapper method

Classifiers	**Feature set**	**Precision**	**ROC Area**	**Error rate**	**Feature set**	**Precision**	**ROC Area**	**Error rate**
1-NN	VRMS, EDA	0.947	0.937	5.55	Cluster 1	0.841	0.838	15.97
Trees (Simple CART)	EDA	0.897	0.879	10.42	Cluster 1	0.896	0.896	10.41
Logistic	VRMS, ASI, IH, SMA, SMV, EDA	0.944	0.976	5.55	Cluster 1	0.868	0.934	13.19
MLP	IA, SED, ASI, VSI, IH, SMA	0.937	0.966	6.25	Cluster 1	0.917	0.953	8.33
3-NN	PGA, IA, IC, ASI	0.952	0.938	4.86	Cluster 1	0.889	0.943	11.11
NB	PGA, PGA/PGV, ASI, VSI	0.938	0.977	6.25	Cluster 1	0.899	0.949	10.41
Decision table	IC, ASI, EDA	0.903	0.946	9.72	Cluster 1	0.871	0.957	13.19

TABLE VIII: Results of the testing classifiers using the optimum feature subsets chosen for each classifier and the two-storey building dataset (35 records-zone a) using the wrapper method

Classifiers	**Feature set**	**Precision**	**ROC Area**	**Error rate**	**Feature set**	**Precision**	**ROC Area**	**Error rate**
1-NN	PGA, IC, SED, EDA	0.945	0.941	5.55	Cluster 1	0.92	0.915	8.02
Trees (Simple CART)	A95	0.927	0.883	7.4	Cluster 1	0.892	0.915	11.11
Logistic	PGA, IC, IH, A95	0.951	0.984	4.95	Cluster 1	0.867	0.955	13.6
MLP	PGA, IH, A95	0.945	0.975	5.55	Cluster 1	0.918	0.963	8.64
3-NN	PGA, PGV, PGD, IH	0.946	0.961	5.55	Cluster 1	0.914	0.953	8.64

NB	PGA, ARMS, VSI, IH	0.951	0.986	4.93	Cluster 1	0.913	0.972	9.26
Decision table	VRMS, SMV, A95	0.94	0.968	6.17	Cluster 1	0.89	0.96	11.11

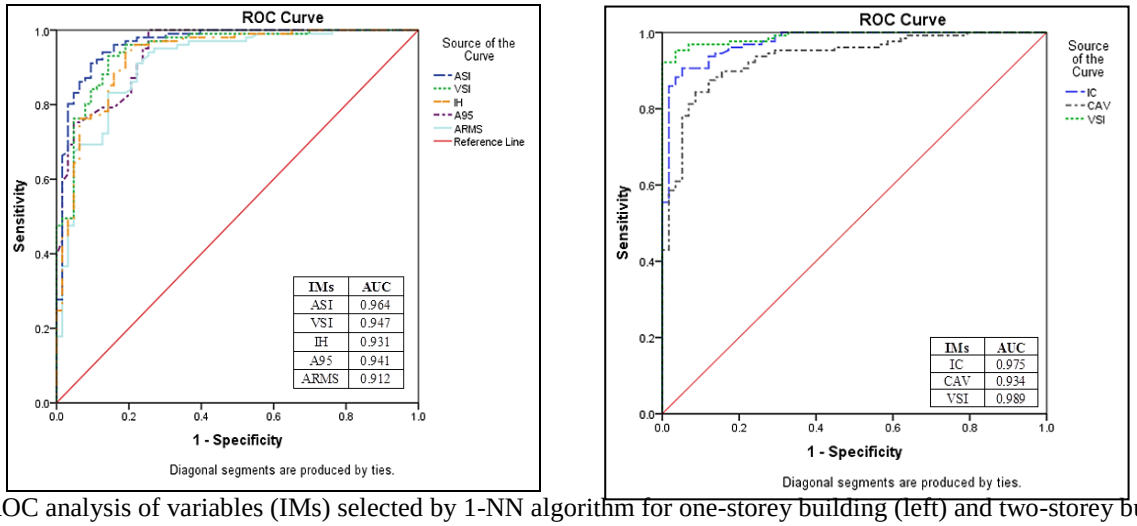

Fig. 2 ROC analysis of variables (IMs) selected by 1-NN algorithm for one-storey building (left) and two-storey building (right) in zone B (Collapse damage state)

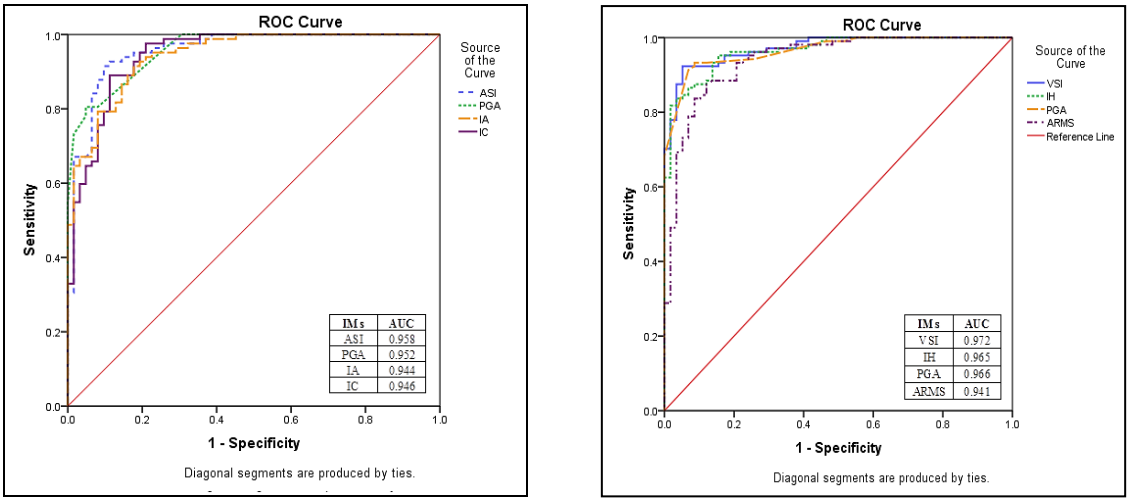

Fig. 3 ROC analysis of variables (IMs) selected by 3-NN and NB algorithms, respectively for one-storey building (left) and two-storey building (right) in zone A (Collapse damage state)

V. CONCLUSIONS

The feasibility of detecting optimal ground-motion intensity measure (IM) from a large number of IMs has been demonstrated in the study. By using wrapper feature selection method, the 36 IMs reduced to subsets of one to six IMs providing more accurate prediction of the seismic responses of the two tested unreinforced masonry buildings. The most successful classifier algorithms tested were the Nearest Neighbours, followed by Navies Bias. Minimum error rate of 2.7 was achieved by 1-NN classification algorithm using two-storey building dataset of seismic hazard zone B. On the other hand, 10.4% of error (as the maximum error rate) was produced by decision tree (simple CART) algorithm in the predicted results of the one-storey building located in the seismic hazard zone A. Overall, without using wrapper model, the minimum and maximum error rate were 6 and 16, respectively. The relative speed with which the analysis presented here for detection of the correct IM is quite promising despite the small datasets available. Furthermore, it has also been shown that the feature set chosen should be considered as part of the classifier algorithm.

As shown in the study, feature selection is very important for classifying IMs and prediction of the failure pattern of buildings. For the future work, we will widen our scope to consider more feature selection and classification algorithms such as boosting, genetic algorithm, evolutionary algorithm, and support vector machine. So that we can find an optimal approach to determining discriminatory features. To find common features from different feature selection methods is another interesting problem. We may also consider the different datasets obtained from different locations.

ACKNOWLEDGMENT

The authors are grateful to the TREMURI staff for providing the academic and commercial versions of the TREMURI software, which were used in the example analyses.

REFERENCES

[1] J. Park, P. Towashiraporn, J. I. Craig, and B. J. Goodno, "Seismic fragility analysis of low-rise unreinforced masonry structures," *Eng Struct* , vol. 3, pp. 125–137, 2009.

[2] G. W. Housner and P. C. Jennings, "Generation of artificial earthquakes*," ASCE J. Eng. Mech. Div.*, vol. 90, pp. 113–150, 1964.

[3] A. Arias, *A measure of earthquake intensity, in Seismic design for nuclear power plants*, R. J. Hansen ed., Cambridge: Massachusetts Institute of Technology Press, 1970, pp. 438–483.

[4] N. Shome, C. A. Cornell, P. Bazzurro, J. E. Carballo, "Earthquakes, records, and nonlinear responses," *Earthq Spectra* , vol. 14, pp. 469–500, 1998.

[5] Central Statistical Organization. Buildings, *Dwelling and Establishment Census and households listing. Enumeration and listing report* (Report No. 1, Buildings, Dwelling and Households- National level), Baghdad: CSO, 2011.

[6] S. A. Ameer, M. L. Sharma, H. R. Wason, and S. A. Alsinawi, "Probabilistic Seismic Hazard Assessment for Iraq Using Complete Earthquake Catalogue Files," *Pure and Applied Geophysics [Pure Appl. Geophys.]*, vol. 162, pp. 951-966, 2005.

[7] Z. V. Milutinovic and G. S. Trendafiloski, "WP4: Vulnerability of current buildings," RISK-UE project of the EC: an advanced approach to earthquake risk scenarios with applications to different European towns , 2003.

[8] H. Crowley, M. Colombi, V. Silva, N. Ahmad, M. Fardis, G. Tsionis, A. Papailia et al., "Fragility functions for common RC building types in Europe," vol. 3, Tech. Rep, 2011.

[9] P. Gehl, D. M. Seyedi, and J. Douglas, "Vector-valued fragility functions for seismic risk evaluation," *Bulletin of Earthquake Engineering*, vol. 11, pp. 365-384, 2013.

[10] S. Lagomarsino, A. Penna, and A. Galasco, "TREMURI program: Seismic analysis program for 3D masonry buildings," University of Genoa, Italy, 2006.

[11] S. L. Kramer, *Geotechnical earthquake engineering*, Upper Saddle River, New Jersey: Prentice Hall, 1996.

[12] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," *ICML*, vol. 3, pp 856–863, 2003.

[13] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell*., vol. 97, pp. 273–324, 1997.

[14] A. Padmapriya and K. S. C. Maragatham, "Algorithms for computer aided diagnosis–an overview," *Int J Comput Trends Technol* , vol. 4, pp. 472–478, 2013.

[15] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic, 1998.

[16] I. Kojadinovic and T. Wottka, "Comparison between a filter and a wrapper approach to variable subset selection in regression problems," in *Proceedings of the European Symposium on Intelligent Techniques (ESIT)*, 2000, pp. 311–321.

[17] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature Selection for Clustering- a Filter Solution," in *Proc. Second Int'l Conf. Data Mining*, 2002, pp. 115-122.

[18] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics* , vol. 23, pp. 2507–2517, 2007.

[19] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques with Java implementations*, San Francisco: Morgan Kaufmann Publishing House, 1999.

[20] Mr. Bayes, and Mr Price, "An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS," *Philosophical Transactions (1683-1775)*, pp. 370-418, 1763.

[21] E. Eix and J. L. Hodges Jr, *Discriminatory analysis-nonparametric discrimination: consistency properties*. University of California, Berkeley, Technical Report Project 21-49-004, Report no. 4, 1951.

[22] H. Liu, Li. Jinyan, and W. Limsoon, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informatics*, vol. 13, pp. 51-60, 2002.

[23] S. Le Cessie and J. C. van Houwelingen, "Ridge estimators in logistic regression," *Appl Stat* , vol. 41, pp. 191–201, 1992.

[24] R. Kohavi, "The power of decision tables," in *Machine Learning: ECML-95*, 1995, pp. 174-189.

[25] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[26] P. Perner, "Improving the accuracy of decision tree induction by feature preselection," *Appl Artif Intell.*, vol. 15, pp. 747–760, 2001.

[27] Tanagra – free data mining software for teaching and research [Online]. Available: http://eric.univ-lyon2.fr/_ricco/tanagra/en/tanagra.

