2016 International Conference on Frontiers of Information Technology

Short Term Earthquake Prediction in Hindukush Region using Tree based Ensemble Learning

Khawaja Muhammad Asim[a], Adnan Idris[b], Francisco Martínez-Álvarez[c], Talat Iqbal[a]

[a]Centre for Earthquake Studies
National Centre for Physics
Islamabad, Pakistan
asim.khawaja@ncp.edu.pk

[b]Department of Computer Sciences & Information Technology
The University of Poonch
Rawalakot AJK, Pakistan
adnanidris@upr.edu.pk

[c]Department of Computer Science
Pablo de Olavide University
Seville, Spain
fmaralv@upo.es

***Abstract*— Earthquake prediction has been long considered as impossible phenomenon but recent research studies show some progress in this field by considering it as a data mining problem. There are numerous challenges in earthquake prediction, which includes highly non-linear behavior of seismic activity and non-availability of reliable seismic precursors. This work focuses on earthquake prediction in Hindukush region by employing mathematically computed seismic features and using these features to model earthquake occurrences through employing machine learning techniques. The study aims to consider earthquake prediction as a binary classification problem. The short term earthquake prediction is performed using tree based ensemble classifiers, where rotation forest has shown good prediction results, compared to random forest and rotboost.**

Keywords—earthqake prediction; seismic precursors; ensemble learning;

I. INTRODUCTION

Earthquake prediction has always been considered an impossible phenomenon but recent research studies have shown a new paradigm to explore this goal [1-3]. There are numerous difficulties in earthquake prediction including highly non-linear behavior of seismic activity and non-availability of reliable seismic precursors. This research study carries out earthquake prediction for Hindukush region by employing useful seismic features through machine learning techniques. Earthquake prediction is categorized into long term, midterm and short term prediction based on the horizon of prediction ranging from years to months and days, respectively. The use of seismic facts and data in combination with artificial intelligence has become increasingly popular in the field of seismic activity and earthquake prediction. The process of seismic activity prediction includes identifying suitable mathematically calculated parameters and co-relating them to the actual earthquake occurrences from the past [4-7] by means of intelligent data mining techniques. Such a methodology of earthquake prediction relies upon temporal sequence of past earthquakes, i.e. earthquake catalog.

The authors of [4, 5] carried out earthquake prediction research on monthly basis for southern California and San Francisco bay regions. In the research, eight mathematically computed seismic parameters have been proposed for earthquake prediction. These parameters highlight the foreshocks frequency, seismic energy release and frequency distribution of past earthquakes. These parameters are calculated using the temporal distribution of past earthquakes in the region. The authors have employed back propagation neural network and probabilistic neural networks, on calculated seismic parameters for earthquake prediction.

The authors of [6, 8] proposed the use of seven mathematically calculated seismic parameters for earthquake prediction in Chile and Iberian Peninsula. The parameters highlight the underlying Omori and Utsu law [9] and Gutenberg-Richter relationship [10]. In this work, artificial neural networks (ANN) have been used for earthquake prediction. The results are further enhanced by selecting best set of seismic parameters by measuring average information gain of every available feature in literature [7]. Finally, the best selected parameters yielded better results for the Chile and Iberian Peninsula compared to preceding research studies.

In another study [1], authors proposed an expert system based methodology for earthquake prediction. They consider the whole globe and divide it into four quadrants. Earthquake prediction model is based on association mining rules and predicate logic. The model takes earthquake occurrences around the globe as an input and earthquake prediction is made for one of the four quadrants. This association rules based expert system is capable

978-1-5090-5300-1/16 $31.00 © 2016 IEEE
DOI 10.1109/FIT.2016.70

of predicting earthquakes occurring in next 12 hours in every quadrant of globe.

Similarly, earthquake magnitude prediction in the Northern Red sea has been carried out using feed forward neural networks [11]. The features used in the research study are earthquake sequence number, location tile, earthquake depth and magnitude. The research study presented in [3], predict earthquakes in different regions of Israel on yearly basis. The parameters are based on foreshock frequency, seismic energy release and frequency distribution of earthquakes. The results obtained show that Multi-Objective Info-Fuzzy Network is producing better results for earthquake prediction.

Earthquake prediction in Tokyo has also been studied in [12] using various seismic parameters introduced in literature through employing various machine learning techniques. This study extends the work proposed in [4, 6]. A parameterized approach has been used to extract the improved information, which is consequently used for earthquake prediction carried out for the year 2015, with prediction period of seven days. ANN is reported to have obtained better results. The authors in [13] investigate the prospective of Qesham earthquake in Southern Iran, occurred on September 10th, 2008. The study explores the applications of Radial Basis Function (RBF) and Adaptive Neural Fuzzy Inference System (ANFIS) to predict the occurrences large earthquakes using seismic features, based on seismic rate changes and earthquake frequency distributions.

Thus, numerous studies are found in contemporary literature, where seismic activity prediction is conducted through applying intelligent data mining techniques, on numerous mathematically computed seismic features. In our proposed work, we have employed a new combination of features, based on seismic concepts of Gutenberg-Richter relationship, seismic rate changes, foreshock frequency and tree based ensemble classification method to predict the seismic activity in Hindukush region. Our work is making a contribution in terms of newest ensemble classifiers being applied to predict the seismic activity in the Hindukush region. The obtained results show that rotation forest has strong ensemble classification capabilities by using the divergence of the data and thus attains improved prediction scores. Random forest attained second best performance in terms of AUC and precision as compared to RotBoost, and decision trees.

The rest of the manuscript is structured as, Section 2 explains the computation of seismic features. Ensemble based classification methodologies are presented in Section 3 and Section 4 contains the results and discussion

II. SEISMIC FEATURE CALCULATION

Feature calculation is the most important step in the research area of seismic activity prediction. In this research, a polygon shaped Hindukush region has been selected for short term seismic activity prediction, as proposed in [14]. Seismic features are calculated using the earthquake catalog of the selected region. A catalog is a complete list of earthquake magnitudes, location, time of occurrence and depth that have occurred in past. The study period for this work is selected from January, 1980 to January 2016. The earthquake catalog is obtained from US Geological Survey and is publically available [15]. The catalog must be evaluated for completeness magnitude or cut-off magnitude (M_c). M_c is defined as the minimum reported magnitude, above which the catalog is considered to be complete. There are different methods available in literate, which can be used for finding M_c [16]. The earthquakes listed in catalog below cut-off magnitude are removed before employing for feature calculation. M_c is calculated through visual examination of frequency magnitude distribution of events [17].

The Hindukush is one of the most seismically active regions in the world where small and medium magnitude earthquakes keep hitting on regular basis. The region came into existence when Indian plate collided with Eurasian plate during Eocene [18]. In this work, seismic features are calculated corresponding to every earthquake that has ever occurred in the region on the basis of previous 50 earthquakes. The feature vector corresponding to every earthquake is meant to represent the seismic state of the region when earthquake occurred. This earthquake prediction task is taken as a binary classification problem. The study is focused on predicting earthquakes as “Yes” or “No” rather than treating it as a regression problem, in which corresponding target magnitudes are converted to binary class through applying a feasible threshold on target magnitudes. After achieving a significant success of predicting earthquakes in binary form, the issue of forecasting exact magnitude may be considered in future. Thus in the current work, binary classification models are developed through inducing training from calculated features and consequently tested using 10 folds cross validation. These classification models are capable of predicting earthquakes of magnitude 5.0 and above for the horizon of 15 days.

There are total of 51 seismic features calculated for earthquake prediction in this study. Underlying concepts and governing rules of these features are seismic energy release, foreshock frequency distribution of earthquakes (Gutenberg-Richter’s law) and seismic rate changes. All the considered parameters are calculated using different possible procedures and techniques in order to retain maximum information about the seismic state of the region. The Detailed calculation of every seismic parameter is given below:

A. Time (T) of n events

It is the difference in days between of the time of occurrence of *1^{st}* event and *nth* event from the past [5]. The word event corresponds to earthquake occurrences. Eq. 1 shows the calculation of *T* and *n* in our case is selected to be 50. *T* signifies the foreshock frequency.

$$T = t_n - t_1 \tag{1}$$

B. Gutenberg-Richter law

Gutenberg-Richter (GR) law states that the number of earthquakes increase exponentially with decreasing magnitude. There exist an inverse relationship between number of earthquakes and their magnitude which is expressed in Eq. 2.

$$\log N_i = a - bM_i \qquad (2)$$

Where, N_i is the total number of events greater than and equal to magnitude M_i. For decreasing magnitude M of i^{th} earthquake, N increases exponentially. b is slope of curve and a is y-intercept. Both a and b are very important seismic parameters. a represents the productivity of the region whereas b shows the seismicity rate of the region. The authors of [5] suggest least square fitting method for the calculation of a and b value, while in [6], these parameters are calculated using maximum likelihood method. In this study, both methods have been separately applied for the calculation of a and b, therefore, it provides four seismic features.

C. Standard Deviation of b value

Standard deviation σb of b value is calculated using Eq. 3 [17]. Where, n denotes the total number of events used to calculate b value and $\bar{M}$ shows the mean magnitude of all the earthquake magnitudes involved in calculation of b value. Since, in this study b values are calculated by two different methods as mentioned above, so both the values are separately used to calculate σb , therefore adding two more features to the dataset.

$$\sigma b = 2.3b^2 \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{M}_i - \bar{M})^2}{n(n-1)}} \qquad (3)$$

D. Seismic Rate Changes

Seismicity rate changes are another important seismic parameters. There are two different ways proposed to calculate seismic rate changes. Habermann (1988) [19] proposed z value for measuring seismic rate change between two intervals as given in Eq. 4.

$$z = \frac{R_1 - R_2}{\sqrt{\frac{S_1}{n_1} + \frac{S_2}{n_2}}} \qquad (4)$$

$R_{1,2}$ denote the mean seismicity rates, $S_{1,2}$ represent the standard deviation of the rates and $n_{1,2}$ are the number to events during period 1 and 2. Period 1 correspond to the whole interval except the interval of interest while period 2 is the interval of interest about which seismic rate change is calculated.

The other parameter proposed to evaluate seismicity rate change is known as β value [20]. The β value is defined as follows in Eq. 5.

$$\beta = \frac{M(t, \delta) - n\delta}{\sqrt{n\delta(1-\delta)}} \qquad (5)$$

Where, n is the total number events in the dataset, t is total time of duration and δ is the normalized duration of interest. $M(t, \delta)$ corresponds to the number of the events observed defined using end time t and interval of interest δ. Both z and β value are independent from each other and possess opposite signs to each other.

E. Probabilistic Recurrence Time

Probabilistic Recurrence Time (T_r) is calculated on the basis of a and b values calculated for earthquakes during interval T (Eq. 1) for the selected magnitude of equal and greater than M'. T_r is calculated using Eq. 6

$$T_r = \frac{T}{10^{a-bM'}} \qquad (6)$$

T_r is also known as total recurrence time. It is basically the estimate of time between two earthquakes of magnitude greater than and equal to M'. In order to retain maximum information for earthquake prediction dataset, T_r is calculated for every earthquake magnitude ranging from 4.0 to 6.0. As we already have a and b values calculated using two different methods, so both the values are used separately for the calculation of T_r, Therefore enabling T_r to contribute 42 more features to the earthquake prediction dataset.

There are total of 51 meaningful seismic parameters calculated for our dataset based on well-known seismic facts as explained above. The combination of these 51 seismic features, in which every parameter is calculated using multiple available methods and possibilities, on the principle of retaining maximum information for earthquake occurrence. Previously proposed earthquake prediction datasets in [4-6, 8, 13] use only single value for calculation of every seismic parameter. Therefore, the approach of calculating every parameter in multiple ways makes this dataset unique. The dataset comprises of 4351 features instances. Earthquake prediction problem is treated as a binary classification problem, which aims to predict earthquakes of magnitude equal to and greater than 5.0. Various classification methodologies are explored to obtain the predictions on calculated seismic feature set. WEKA freely available data mining tool and MATLAB implementations are used for experiments.

III. Earthquake Prediction Models

A. Decision Tree J48

Decision tree algorithm is initially applied on the computed seismic dataset, for obtaining the binary predictions. Decision tree operates through selecting a root note from the original dataset ***S***. The information gain for every unused attribute is computed in each iteration. The attributes having smallest information gain are selected and in this way the set ***S*** is split to produce the subsets of the original dataset. The algorithm traverses each subset and considers only those attributes which are not selected in previous iterations. Thus, a learning model based on decision tree is developed, where values of the input variables are depicted by the path from the root the leaf and each leaf represents the value of the target variable. Moreover, decision trees are also considered as a classifier, which learns the variations present in the data effectively.

B. Random Forest

Random Forest [21] is an ensemble classifier that is developed on the basis of majority voting of decision trees. Various number of decision trees are generated over bootstrap samples of the training dataset. The final decision is made by aggregating the predictions obtained by all the decision trees. For, multiple decision trees are involved in random forest, the results obtained are improved compared to the use of single decision tree. Random forest minimizes the overall error rate and focuses to improve the prediction accuracy. Moreover, random forest is also chosen to be used in this study for its capabilities to obtain improved results in the case of balanced class distribution in the dataset, which nearly exists in our case of seismic activity prediction. Random Forest is developed using 20 number of decision trees, while having each depth of 10 nodes in each tree.

C. Rotation Forest

Rotation Forest [22] is another ensemble classifier that instantaneously attains diversity and accuracy. *K* feature subsets are extracted from the original feature space and then Principal component analysis is applied on each feature subset. Consequently, a rotation matrix is constructed with the help of principal components extracted on each subset. The rotation matrix extends *K* axis rotation which rotates the input resulting in higher diversity. Decision trees are used as base classifier, which exploits the diversity attained through rotation forest. Moreover, Rotation forest also uses all the principal components extracted from *K* subsets, which results in achieving better accuracy. We implemented the Rotation Forest by selection 8 feature subsets for the construction of rotation matrix.

D. RotBoost

RotBoost is a recently introduced ensemble classification algorithm that combines rotation forest and adaboost [23]. In RotBoost, rotation matrix construction is inherited from Rotation Forest, whereas iterative weights updating process for hard instances is taken from Adaboost. Generally, a good performance of an ensemble classifier is also attributed to the use of base classifier. Thus decision tree is also considered as base classifier in the case of RotBoost. The RotBoost's weight updating was attained with 12 numbers of iterations.

IV. Results and Discussion

Results of the applied machine learning techniques on binary classification problem of earthquake prediction have been evaluated using 10 fold cross validation. We have considered various performance measures, which include:

True Positive (TP): Actual earthquake occurrences, which are also predicted by the classification model are known as true positives.

False Positive (FP): When prediction model predicts earthquake but there is not actual event, is known as false positive.

True Negative (TN): When there is no earthquake and the model also predicts the same is known as true negative.

False Negative (TN): This corresponds to the situation in which prediction model fails to predict an earthquake.

There are other important performance evaluation criteria, derived from the above mentioned parameters which are known to be balanced performance measures such as sensitivity, specificity, precision, F measure and area under curve (AUC). Sensitivity (S_n) is capability of the classification model to predict actual positives while Specificity (S_p) refers to the ability of the model to predict actual negatives among all the negatives. S_n and S_p are calculated using Eq. 7 and Eq. 8, respectively. S_p is also known as recall.

$$S_n = \frac{TP}{TP+FN} \tag{7}$$

$$S_p = \frac{TN}{TN+FP} \tag{8}$$

Precision is the ratio of true positive predictions and all positive predictions of algorithm. It is also known is positive predictive value. It is the measure of false of alarms, indirectly because if a classification model possesses higher value of precision, it means there are less false alarms. It is calculated using Eq. 9.

$$P_1 = \frac{TP}{TP+FP} \tag{9}$$

F measure is another evaluation criterion, also known as F score. It is the harmonic mean of precision and recall, which can be calculated using Eq. 10.

$$F = \frac{2TP}{2TP+FP+FN} \tag{10}$$

The performance of a binary classification model is also evaluated by drawing a Receiver Operating Curve (ROC). It is drawn between True Positive Rate (TPR) and False Positive Rate (FPR) for different decision thresholds. AUC of 1 corresponds to the maximum classification performance.

A. Comparative Analysis

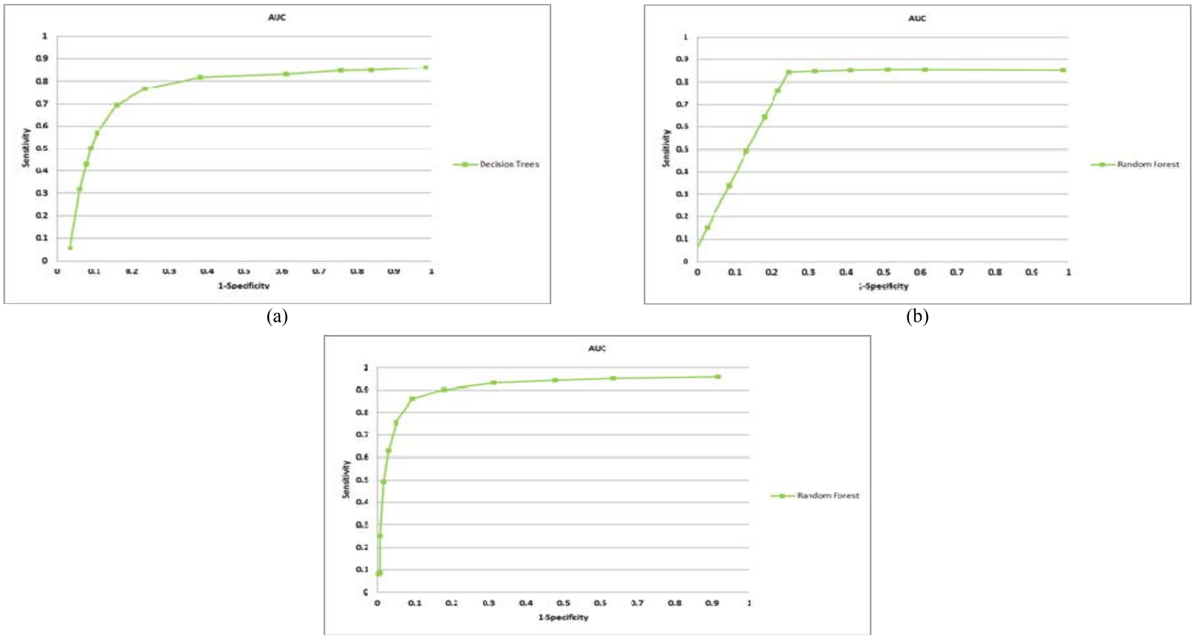
Four aforementioned tree based classifiers have been trained and tested on the calculated earthquake prediction dataset using 10 fold validation strategy. The performance of every model is evaluated on basis of above discussed evaluation criteria. All four classification techniques applied in this research showed impressive performance and performed exceptionally well for earthquake prediction thus representing the strength and the effectiveness of new seismic parameters computation strategy. Table 1 summarizes the performance attained for all four classification models in numeric form.

Rotation forest takes lead among all the models through showing better performance in all the evaluation measures with AUC and F measure of 95.9% and 92.8%, respectively. The other classifiers are not much behind, with random forest showing second best performance in avoiding false alarms with precision of 79%. ROC corves are potted for decision tree J48, random forest and rotation forest to depict their performance as shown in Figure 1.

Table 1: Results of classifiers obtained through 10 fold cross validation strategy

Classification Model	J48	Random Forest	Rotation Forest	RotBoost
S_n	0.778	0.803	0.952	0.948
S_p	0.960	0.950	0.951	0.806
Precision	0.785	0.791	0.905	0.724
F Measure	0.861	0.797	0.928	0.821
AUC	0.861	0.854	0.959	0.636

Since every region has its own properties and is different from other regions, like, Hindukush region is formed due to subduction of Indian plate under Eurasian plate. Therefore, for every region, separate training will be required to train a prediction model based upon seismic features of that particular region.



(a)

(b)

(c)

Figure 1: ROC Curves for classifiers showing the performance and AUC. (a): Decision tree J48 (b): Random forest (c): Rotation forest

CONCLUSION

Results obtained in this work, depict that machine learning approaches can play considerable role in prediction of seismic activity. Simulation results represent that Rotation forest is able to attain higher prediction performance with regards of seismic activity in Hindukush region, compared to other classification methodologies such as decision trees, random forest and rotboost. Computation of catalog based seismic features through a different approach has shown encouraging performance in combination with tree based ensemble classifiers.

ACKNOWLEDGMENTS

Authors would like to show gratitude towards friends and colleagues for their continuous support and advice. Furthermore, authors would like to thank US Geological Survey which has made the earthquake catalog publically available for use through website, since the study relies on catalog, therefore without catalog, this study would not have been possible.

References

[1] A. Ikram and U. Qamar, "Developing an expert system based on association rules and predicate logic for earthquake prediction," *Knowledge-Based Systems,* vol. 75, pp. 87-103, 2015.

[2] G. Asencio-Cortés, F. Martínez-Álvarez, A. Morales-Esteban, and J. Reyes, "A sensitivity study of seismicity indicators in supervised learning to improve earthquake prediction," *Knowledge-Based Systems,* vol. 101, pp. 15-30, 2016.

[3] M. Last, N. Rabinowitz, and G. Leonard, "Predicting the Maximum Earthquake Magnitude from Seismic Data in Israel and Its Neighboring Countries," *PloS one,* vol. 11, p. e0146101, 2016.

[4] H. Adeli and A. Panakkat, "A probabilistic neural network for earthquake magnitude prediction," *Neural networks,* vol. 22, pp. 1018-1024, 2009.

[5] A. Panakkat and H. Adeli, "Neural network models For earthquake magnitude prediction using multiple seismicity indicators," *International Journal of Neural Systems,* vol. 17, pp. 13-33, 2007.

[6] J. Reyes, A. Morales-Esteban, and F. Martínez-Álvarez, "Neural networks to predict earthquakes in Chile," *Applied Soft Computing,* vol. 13, pp. 1314-1328, February 2013.

[7] F. Martínez-Álvarez, J. Reyes, A. Morales-Esteban, and C. Rubio-Escudero, "Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula," *Knowledge-Based Systems,* vol. 50, pp. 198-210, 2013.

[8] A. Morales-Esteban, F. Martínez-Álvarez, and J. Reyes, "Earthquake prediction in seismogenic areas of the Iberian Peninsula based on computational intelligence," *Tectonophysics,* vol. 593, pp. 121-134, 2013.

[9] T. Utsu and Y. Ogata, "The centenary of the Omori formula for a decay law of aftershock activity," *Journal of Physics of the Earth,* vol. 43, pp. 1-33, 1995.

[10] B. Gutenberg and C. Richter, *Seismicity of the earth and associated phenomena*: Princeton University Press, 1954.

[11] A. S. N. Alarifi, N. S. N. Alarifi, and S. Al-Humidan, "Earthquakes magnitude predication using artificial neural network in northern Red Sea area," *Journal of King Saud University - Science,* vol. 24, pp. 301-313, 2012.

[12] G. Asencio-Cortés, F. Martínez-Álvarez, A. Troncoso, and A. Morales-Esteban, "Medium–large earthquake magnitude prediction in Tokyo with artificial neural networks," *Neural Computing and Applications,* pp. 1-13.

[13] A. Zamani, M. Sorbi, and A. Safavi, "Application of neural network and ANFIS model for earthquake occurrence in Iran," *Earth Science Informatics,* vol. 6, pp. 71-85, 2013/06/01 2013.

[14] K. Asim, F. Martínez-Álvarez, A. Basit, and T. Iqbal, "Earthquake magnitude prediction in Hindukush region using machine learning techniques," *Natural Hazards,* pp. 1-16.

[15] U. S. G. Survey. Quaternary fault and fold database for the United States [Online]. Available: http//earthquake.usgs.gov/hazards/qfaults/

[16] S. Wiemer and M. Wyss, "Minimum magnitude of completeness in earthquake catalogs: examples from Alaska, the western United States, and Japan," *Bulletin of the Seismological Society of America,* vol. 90, pp. 859-869, 2000.

[17] S. Wiemer and M. Wyss, "Mapping spatial variability of the frequency-magnitude distribution of earthquakes," *Advances in geophysics,* vol. 45, pp. 259-302, 2002.

[18] A. Farah, G. Abbas, K. A. De Jong, and R. D. Lawrence, "Evolution of the lithosphere in Pakistan," *Tectonophysics,* vol. 105, pp. 207-227, 1984.

[19] R. Habermann, "Precursory seismic quiescence: past, present, and future," *Pure and Applied Geophysics,* vol. 126, pp. 279-318, 1988.

[20] M. V. Matthews and P. A. Reasenberg, "Statistical methods for investigating quiescence and other temporal seismicity patterns," *Pure and Applied Geophysics,* vol. 126, pp. 357-372, 1988.

[21] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[22] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE transactions on pattern analysis and machine intelligence,* vol. 28, pp. 1619-1630, 2006.

[23] C.-X. Zhang and J.-S. Zhang, "RotBoost: A technique for combining Rotation Forest and AdaBoost," *Pattern Recognition Letters,* vol. 29, pp. 1524-1536, 2008.