iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** VI **Month of publication:** June 2018

DOI: http://doi.org/10.22214/ijraset.2018.6056

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

Prediction of Seismic Activities in Coal Mines using Decision Tree based Ensemble Learning

Shrutika Ganesh[1], Karishma Kothari[2], Vaibhavi Nandode[3],Vedangi Godse[4] ,Nihar Ranjan[5]
[1, 2, 3, 4, 5]
Sinhgad Institute Of Technology And Science, Narhe, Pune.

Abstract: Seismic activities pose high risks to everyone who lives in an active coal mine region. Even though the hazard is well recognized, no one knows when it will strike or how severe it will be. In this paper, we introduce a method using Random Forest and Rotation Forest for building classifier ensembles based on decision trees using WEKA tool. While Random Forest dealt with the outliers of dataset efficiently, Rotation Forest proved to increase accuracy by using Principal Component Analysis (PCA) to each rotated subset of features. The system provides additional feature of notifying the engineer in charge prior with the danger. This paper mainly aims to advance knowledge and practice that could lead to prevention of seismic activities.
Keywords: PCA, WEKA, Random Forest, Rotation Forest

I. INTRODUCTION

After every damaging seismic activity around the world, a lot of analysis is done in newspapers and on TV regarding the issues of its safety. Numerous conferences are held and discussions are done by authorities all around the world and public feels reassured that the problem will be taken care of, until the next activity when they realize that nothing effective really got done since the last event.

Why do we face the above problem? In engineering, often it is more important and sometimes even more challenging to define the problem than the solution itself. Every stakeholder tends to think that his role is the most crucial in addressing an issue. Hence, differences of opinion are expected between scientists, engineers, administrators and social scientists on how to solve the problem. However, no one will disagree that the problem has to be addressed and can be prevented rather than cured or overcome with. There have been a lot of papers published, coming out with solutions to deal with the problem of seismic activities. There have been solutions put forth which use parallel feature extraction framework, Recurrent Neural Network, Naïve Bayes Classifier, Negation Handling, Histogram- Based feature engineering, most of them having the base as Machine Learning. Machine learning is a method of data analysis that automates analytical model building. Rising interest in machine learning is due to the factors such as growing volumes and varieties of available data, computational processing that is cheaper and more powerful and affordable data storage. Some aspects play a vital role to create good machine learning systems. They include Data preparation capabilities, algorithms, Automation and iterative processes and Ensemble modeling. The problems that these solutions are not able to cover include use of a lot of aggregated methods, lack of handling of dataset, output evaluated with dataset having outliers and noise, no technical feasibility and less accuracy and speed being the top of them all.

The proposed system however is an improvement upon the previous research and implementation where results were not satisfactory. This paper aims at solving the above issues using Random Forest and Rotation Forest to handle dataset efficiently and increase accuracy respectively. The system makes use of classifiers which aim at training the decision trees trained independently and create diversity in the same.

The paper is organized as follows. In Section II and III we present the related work and description of dataset respectively. In Section IV we provide detailed information about the Model, including insights of the ensemble classifiers being used. Next, in Section V, we describe the results and discussion on the model. Finally, in Section VI we summarize the work.

II. RELATED WORK

Seismic activities pose a great threat to miners and overall safety of the coal mining operation. Providing safety of miners working underground is the fundamental requirement of the coal mining industries. Coal mining companies are obligated by the law to introduce many safety measures to secure proper working conditions of their underground personnel. One of the techniques for addressing this problem is to use automatic feature engineering. This method does extraction from time series data that did not require any manual tuning. The paper argued that an ensemble of classifiers will produce a more robust and accurate system than a

single classifier. Another technique to predict the seismic events include using Recurrent Neural Network. This method uses Long Short-Term Memory cells. It requires almost no feature engineering, which makes it applicable to other domains with multivariate time series data. One of the many solutions currently available also includes using an efficient Naïve Bayes Classifier with Negation Handling for Seismic Hazard Prediction. This approach outperforms the traditional Naïve Bayes Classifier in terms of accuracy.

Most of the top current solutions rely heavily on feature engineering, either manual or automatic, such as: Automatic variable construction, window-based feature engineering, hand-crafted features or thousands of automatically generated features. Other techniques include prediction under distribution drift; Tree based Ensemble Learning, using transient features of seismic events etc. The main problems observed in predicting a seismic activity include accuracy of result and the time utilized to produce the results.

III. DATASET

Data in .arff format which is being given as input was downloaded from UCI Machine Learning Repository website which comes under Multivariate category. The data describes the problem of seismic bumps in a coal mine. Data has been collected two of long walls located in a Polish coal mine.

The details of dataset used are as follows:

Table I Details Of Dataset

Data Set Characteristics:	Multivariate	**Number of Instances:**	2584
Characteristics of Attributes:	Real	**Number of Attributes:**	19
Task Associated With Dataset:	Classification	**Missing Values:**	N/A

The description of the 19 attributes in our dataset is described below

Table II Description Of Attributes

Features	Description
seismic	Result of shift seismic hazard assessment obtained by seismic method: a: lack of hazard, b: low hazard, c: high hazard, d: danger state.
seismoacoustic	Records collected in the form of sound waves which are further categorized as: a: lack of hazard, b: low hazard, c: high hazard, d: danger state.
shift	Type of shift observed: W- after coal is mined N- before coal is mined.
genergy	Seismic energy recorded within previous shift using geophone.
gpuls	Pulses of current recorded within previous shift.
gdenergy	Value of deviation in genergy recorded within previous shift.
gdpuls	Value of deviation in gpuls recorded within previous shift.
ghazard	Result of shift seismic hazard assessment in the mine working obtained from seismoacoustic method based on registration coming from GMax only.
nbumps	Number of seismic bumps recorded within previous shift.
nbumps2	Number of seismic bumps in energy range [10^2, 10^3] registered within previous shift.
nbumps3	Number of seismic bumps in energy range [10^3, 10^4] registered within previous shift.
nbumps4	Number of seismic bumps in energy range [10^4, 10^5] registered within previous shift.
nbumps5	Number of seismic bumps in energy range [10^5, 10^6] registered within previous shift.
nbumps6	Number of seismic bumps in energy range [10^6, 10^7] registered within previous shift.
nbumps7	Number of seismic bumps in energy range [10^7, 10^8] registered within previous shift.
nbumps89	Number of seismic bumps in energy range [10^8, 10^10] registered within previous shift.
energy	Total energy of seismic bumps registered within previous shift.
maxenergy	Maximum energy of seismic bumps registered within previous shift.
class	Denotes decision attribute: 1- high energy seismic bump occurred 0- low energy seismic bump occurred

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

Each row of the data describes the seismic activity in the rock mass within one shift (8 hours). A sample screenshot of dataset is as shown:

```
@attribute seismic {a, b, c, d}
@attribute seismoacoustic {a, b, c, d}
@attribute shift {W, N}
@attribute genergy real
@attribute gpuls real
@attribute gdenergy real
@attribute gdpuls real
@attribute ghazard {a, b, c, d}
@attribute nbumps real
@attribute nbumps2 real
@attribute nbumps3 real
@attribute nbumps4 real
@attribute nbumps5 real
@attribute nbumps6 real
@attribute nbumps7 real
@attribute nbumps89 real
@attribute energy real
@attribute maxenergy real
@attribute class {1, 0}

@data
a,a,N,15180,48,-72,-72,a,0,0,0,0,0,0,0,0,0,0,0
a,a,N,14720,33,-70,-79,a,1,0,1,0,0,0,0,0,2000,2000,0
a,a,N,8050,30,-81,-78,a,0,0,0,0,0,0,0,0,0,0,0
a,a,N,28820,171,-23,40,a,1,0,1,0,0,0,0,0,3000,3000,0
a,a,N,12640,57,-63,-52,a,0,0,0,0,0,0,0,0,0,0,0
a,a,W,63760,195,-73,-65,a,0,0,0,0,0,0,0,0,0,0,0
a,a,W,207930,614,-6,18,a,2,2,0,0,0,0,0,0,1000,700,0
a,a,N,48990,194,-27,-3,a,1,0,1,0,0,0,0,0,4000,4000,0
a,a,N,100190,303,54,52,a,0,0,0,0,0,0,0,0,0,0,0
a,a,W,247620,675,4,25,a,1,1,0,0,0,0,0,0,500,500,0
a,a,N,41950,135,-39,-36,a,1,0,1,0,0,0,0,0,5000,5000,0
a,a,N,53250,140,-19,-31,a,0,0,0,0,0,0,0,0,0,0,0
a,a,W,166180,448,-30,-19,a,1,1,0,0,0,0,0,0,400,400,0
a,a,N,64540,215,0,9,a,1,0,1,0,0,0,0,0,6000,6000,0
```

Fig 2: S enshot of dataset in .arff format

IV. MODEL

A. Preprocessing

Decision trees possess a great characteristic that the input data provided for processing does not require any specific preprocessing. Since the trees can choose equivalent splitting points, the results obtained are consistent regardless of any preprocessing technique applied on them. Hence, in our system, no explicit preprocessing is performed on the input dataset.

B. Architecture

After studying the research already done in this field, we have prepared a generic model which we will use to accomplish the purpose of prediction of seismic activities. In order to improve the efficiency of results obtained, the random forest and rotation forest classifiers will be trained before actual classification occurs. This training will use the dataset downloaded from UCI Machine Learning Repository website which comes under Multivariate category. We propose a system that uses an ensemble of classifiers to classify the state. The components of this system will be written in Java programming language. The approach is depicted in figure below:

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

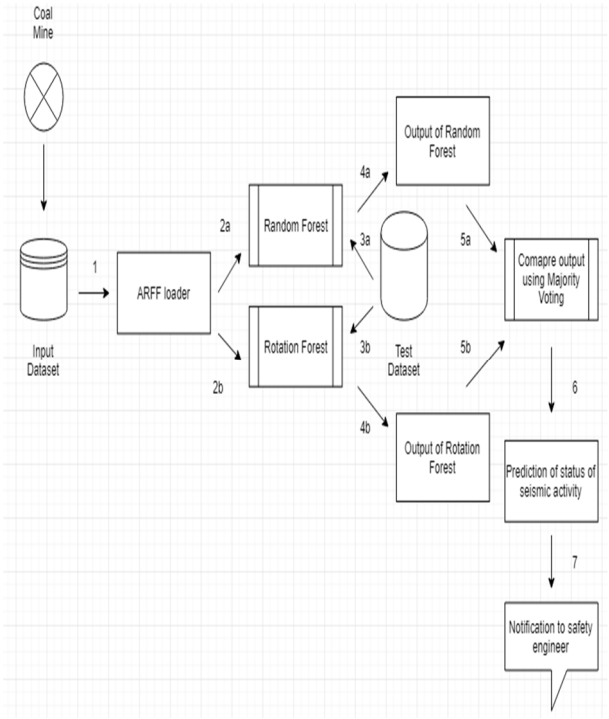


Fig 3: System Architecture

The architecture works as follows:

1) Training dataset is given to the system in .arff extension file.
2) Training dataset is loaded to train Random Forest Classifier.
3) Training dataset is loaded to train Rotation Forest Classifier.
4) After Random Forest gets trained, testing dataset is passed for prediction.
5) Output of Random Forest Classifier that is, the value of predicted class is stored in a List named 'randomoutput'.

C. Model Selection

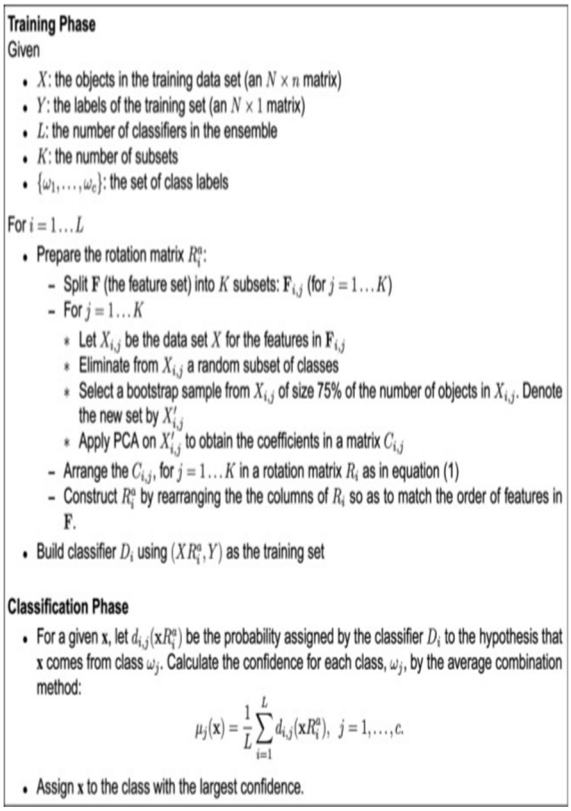
Model selection was a significant challenge. Being a critical application, it was necessary that we made a thorough research on the working, limitations, advantages and other aspects of feasibility study of these classifiers we chose. In the dataset used for our estimation, there are instances where the reading of tremors with energy >10^4j is recorded, decision attribute, which generally describes the event as hazard or non-hazard by 1 or 0, is either kept null or irrelevant. To deal with such issues, we are using Random Forest which is proven as one on the best classifiers to deal with such datasets. Also Rotation Forest is used for better accuracy in prediction. Here is the detailed explanation of the classifiers under consideration:

1) *Random Forest*Random Forest 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 : improved compared to the use of single decision tree. Random forest minimizes the overall error rate and focuses to improve the prediction accuracy. Random Forest is developed using 20 number of decision trees, while having each depth of 10 nodes in each tree. The pseudo code for randomized tree can be given as: Randomly select "k" features from total "m" features, where k << m. mong the "k" features, calculate the node "d" using the best split point.of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).

2) *Rotation Forest*

Rotation Forest is another ensemble classifier that instantaneously attains diversity an accuracy. K feature subsets are extracted from the original feature space and then Principal Component Analysis (PCA) 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. Rotation Forest also uses all the principal components extracted from K subsets, which results in achieving better accuracy. The pseudo code for Rotation Forest is as follows:



Training Phase

Given

- X: the objects in the training data set (an $N \times n$ matrix)
- Y: the labels of the training set (an $N \times 1$ matrix)
- L: the number of classifiers in the ensemble
- K: the number of subsets
- $\{\omega_1, \ldots, \omega_c\}$: the set of class labels

For $i = 1 \ldots L$

- Prepare the rotation matrix R_i^a:
  - Split $\mathbf{F}$ (the feature set) into K subsets: $\mathbf{F}_{i,j}$ (for $j = 1 \ldots K$)
  - For $j = 1 \ldots K$
    - Let $X_{i,j}$ be the data set X for the features in $\mathbf{F}_{i,j}$
    - Eliminate from $X_{i,j}$ a random subset of classes
    - Select a bootstrap sample from $X_{i,j}$ of size 75% of the number of objects in $X_{i,j}$. Denote the new set by $X'_{i,j}$
    - Apply PCA on $X'_{i,j}$ to obtain the coefficients in a matrix $C_{i,j}$
  - Arrange the $C_{i,j}$, for $j = 1 \ldots K$ in a rotation matrix R_i as in equation (1)
  - Construct R_i^a by rearranging the the columns of R_i so as to match the order of features in $\mathbf{F}$.
- Build classifier D_i using (XR_i^a, Y) as the training set

Classification Phase

- For a given $\mathbf{x}$, let $d_{i,j}(\mathbf{x}R_i^a)$ be the probability assigned by the classifier D_i to the hypothesis that $\mathbf{x}$ comes from class ω_j. Calculate the confidence for each class, ω_j, by the average combination method:

$$\mu_j(\mathbf{x}) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x}R_i^a), \quad j = 1, \ldots, c.$$

- Assign $\mathbf{x}$ to the class with the largest confidence.

Fig 4: Pseudo Code for Rotation Forest ensemble method

3) *Tool :* The tool used to build this system is the WEKA tool. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre- processing, classification, regression, clustering, association rules, and visualization. Weka uses the Attribute Relation File Format for data analysis, by default. But listed below are some formats that Weka supports, from where data can be imported:

a) CSV- Comma Separated Value
b) ARFF- Attribute Relation File Format
c) Database using ODBC

D. Ensemble Classifiers

The combination of classifiers gives much better performance than individual classifiers. We are implementing an ensemble of Random Forest and Rotation Forest that would form decision trees based on the values forming class. This algorithm provides over fitting hence unstructured data also forms deep tree. Rotation Forest algorithm gives much smoother boundary to take decision. They are also much faster. Thus, when both the classifiers are used as an ensemble, they give much faster performance and accurate results.

V. RESULTS AND DISCUSSION

The modus operandi used in this system is ensemble, where two classifiers: Rotation Forest and Random Forest are combined for implementation. The ensemble of classifiers has proven to improve accuracy of model than a single classifier. In ensemble classifier, each classifier's output must be considered for the result. This can be accomplished using widely known methods like voting, stacking, averaging, bagging and boosting. In our system voting method is used as it works best for classification in our

system. In this method, multiple classification models are created using training dataset. In our system different dataset with different algorithm is used to create base model. After these predictions are done for each model and saved in 'List' collections of Java. The List consists of prediction class with 1 or 0 value. Here, 1 signifies earthquake predicted for the given instance and 0 signifies no earthquake. Following table shows the various scenario:

Table III Details of Prediction Class

Random Forest predicted class value	Rotation Forest predicted class value	Result
0	0	No Earthquake
0	1	Warning
1	0	Warning
1	1	Earthquake possible

this system for a single instance the output of both the classifiers is considered where the following combination would give the specified output with the help of ensemble method majority voting. In majority voting, final output prediction is one that receives major votes. In our programming 'counter' variables are used for keep count of votes for each instance prediction and final result is computed comparing the number of votes.

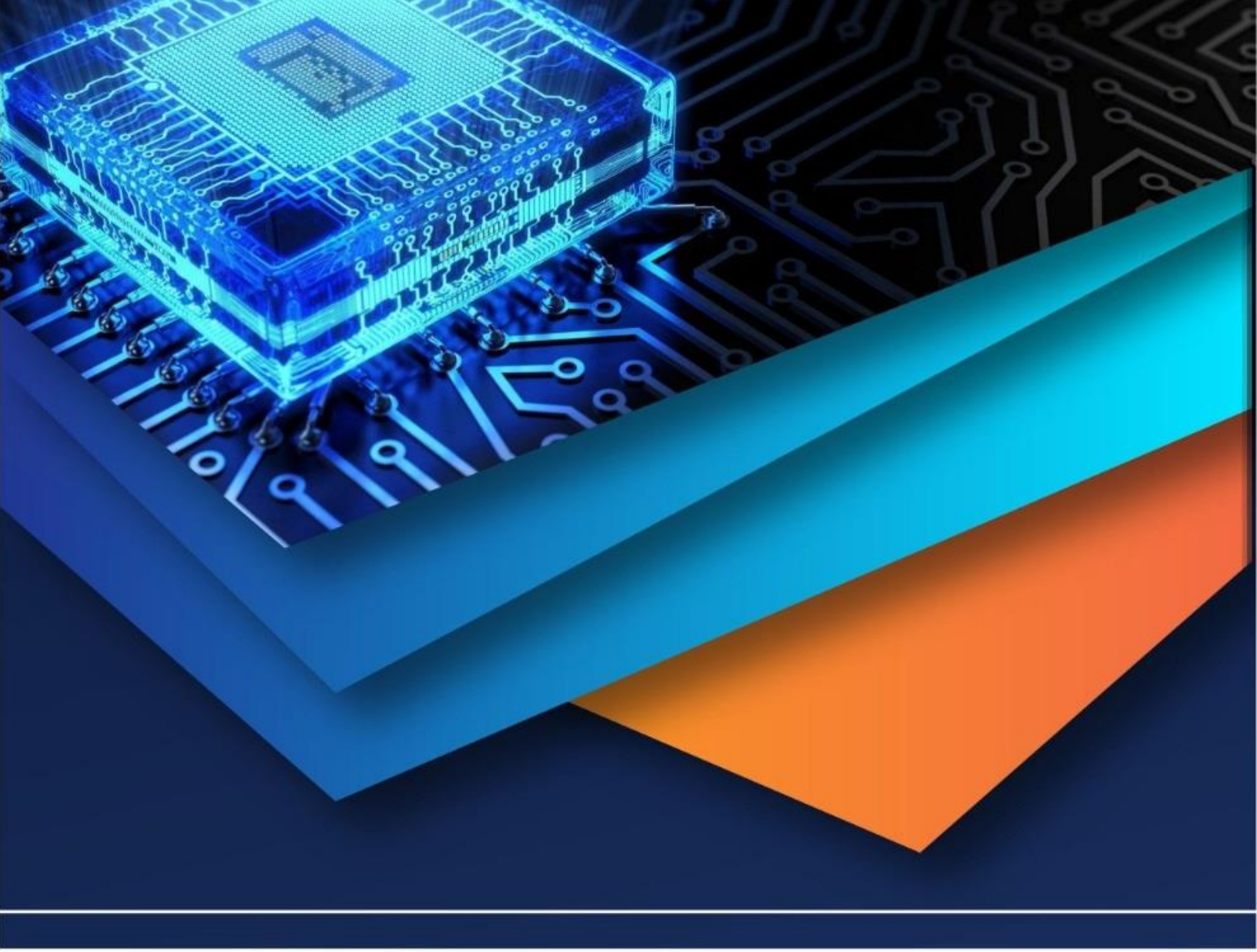
VI. CONCLUSION

Results obtained from this implementation depict that machine learning approaches can play considerable role in predicting seismic activities. In this paper, we have proposed a prediction algorithm for dangerous seismic events in coal mines using a combination of Random Forest (RF) and Rotation Forest (RoF) classifiers, working with support of WEKA Tool. It was observed that Random Forest provides an efficient prediction and Rotation Forest provides an accurate one. The project is highly reusable as many more features can be added to system architecture and a system can be developed towards more accurate prediction. Using physical attributes we can predict Risk Score.

REFERENCES

[1] Zdravevski, E., Lameski, P., & Kulakov, A. (2016). Automatic feature engineering for prediction of dangerous seismic activities in coal mines. Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, 8, 245–248. https://doi.org/10.15439/2016F152
[2] Kurach, K., & Pawlowski, K. (2016). Predicting dangerous seismic activity with Recurrent Neural Networks. Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, 8, 239–243. https://doi.org/10.15439/2016F13
[3] Grzegorowski, M. (2016). Massively Parallel Feature Extraction Framework Application in Predicting Dangerous Seismic Events, 8, 225–229. https://doi.org/10.15439/2016F90
[4] Güzel, B. E. K., & Karaçalı, B. (2016). Fisher's Linear Discriminant Analysis Based Prediction using Transient Features of Seismic Events in Coal Mines. Proceedings of the Federated Conference on Computer Science and Information Systems, 8, 231–234
[5] Boullé, M. (2016). Predicting Dangerous Seismic Events in Coal Mines under Distribution Drift,8 221– 224.
[6] Netti, K., & Radhika, Y. (2016). An efficient Naive Bayes classifier with negation handling for seismic hazard prediction. Proceedings of the 10th International Conference on Intelligent Systems and Control, ISCO2016
[7] Marek Sikora^{1,2} (marek.sikora '@' polsl.pl), Lukasz Wrobel^{1} (lukasz.wrobel '@' polsl.pl) (1)Institute of Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland (2)Institute of Innovative Technologies EMAG, 40- 189 Katowice, Poland(input data set
[8] Bogucki, R., Lasek, J., Milczek, J. K., & Tadeusiak, M. (2016). Early Warning System for Seismic Events in Coal Mines Using Machine Learning,8,213–220.
[9] Bontempi, G., & Group, M. L. (2013). Machine Learning Strategies for Time Series Prediction.
[10] 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.

[11] 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.
[12] 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.L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.
[13] 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.
[14] Kumar, S. L., Paper presented at the Punjab Engineering Congress, 1933.
[15] Rai, D. C., Prasad, A. M. and Jain. S. K., In 2001 Bhuj, India earthquake reconnaissance report (eds Jain, S. K. et al.), Earthquake Spectra, supplement A to volume 18, Earthquake Engineering Research Institute, Oakland, CA, July 2002, pp. 265–277.
[16] Jain, S. K., Murthy, C. V. R., Rai, D. C., Malik, J., Sheth, A. and Jaiswal, A., Curr. Sci., 2005, 88, 357–359.
[17] Spence, R., In Keeping Schools Safe in Earthquakes, Organization for Economic Co-operation and Development, Paris, 2004, pp. 217–228.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY