# DATA IMBALANCE IN LANDSLIDE SUSCEPTIBILITY ZONATION: UNDER-SAMPLING FOR CLASS-IMBALANCE LEARNING

S. K. Gupta[1,*], Muskan Jhunjhunwalla[2], A. Bhardwaj[1], D. P. Shukla[1]

[1] School of Engineering, Indian Institute of Technology Mandi, India (*sharadgupta27@gmail.com)
[2] Department of Computer Science, National Institute of Technology Hamirpur, India

KEY WORDS: Landslide susceptibility zonation (LSZ), Imbalanced learning, Under-sampling, Artificial neural network (ANN), Fisher Discriminant Analysis (FDA), Logistic Regression (LR)

ABSTRACT:

Machine learning methods such as artificial neural network, support vector machine etc. require a large amount of training data, however, the number of landslide occurrences are limited in a study area. The limited number of landslides leads to a small number of positive class pixels in the training data. On contrary, the number of non-landslide pixels (negative class pixels) are enormous in numbers. This under-represented data and severe class distribution skew create a data imbalance for learning algorithms and suboptimal models, which are biased towards the majority class (non-landslide pixels) and have low performance on the minority class (landslide pixels).

In this work, we have used two algorithms namely EasyEnsemble and BalanceCascade for balancing the data. This balanced data is used with feature selection methods such as fisher discriminant analysis (FDA), logistic regression (LR) and artificial neural network (ANN) to generate LSZ maps The results of the study show that ANN with balanced data has major improvements in preparation of susceptibility maps over imbalanced data, where as the LR method is ill-effected by data balancing algorithms. The FDA does not show significant changes between balanced and imbalanced data.

## 1. INTRODUCTION

Landslides are amongst the most devastating natural disasters, which cause billions of dollars in property damage and thousands of deaths every year worldwide. India has more than 15% of its land area prone to landslides, hence mapping of these areas for the presence/absence of landslides is of utmost importance. Numerous studies have contributed to reduce the damage from landslides through modelling and production of susceptibility maps (Roodposhti et al., 2019, Jhunjhunwalla et al., 2019, Gupta et al., 2018, Shukla et al., 2016). The susceptibility mapping can be a crucial tool for a wide range of end-users, from both private and public sectors, aimed at hazard mitigation purposes at both local and international levels. Landslide susceptibility zonation (a.k.a. LSZ) maps give approximate information about the occurrence of landslides. The susceptibility mapping requires data of various factors responsible for slope instability. In this work we have considered seven causative factors such as aspect, elevation, plan curvature, profile curvature, slope, tangential curvature, topographic wetness index.

In recent years, there is an increasing application of machine learning techniques to complex real-world problems. The application ranges from daily life problems to nation's security, processing of the information to decision making support system and from micro-scale analysis of data to macro-scale discovery of knowledge (Stumpf et al., 2012). Most standard machine learning algorithms presume or expect balanced class distributions or equal misclassification costs (He, Garcia, 2009) and suffers data imbalance (Stumpf et al., 2014). Data imbalance refers to a scenario where majority classes dominate or overpower minority classes. In simple words, there is disproportionate distribution of observations in each class. It leads to the classifier being more biased towards the dominating class (Pradhan et al., 2014). Generally, the data is imbalanced when the class ratio is of the order of 1:100, 1:1000 or 1:10000 etc. i.e.number of points in one-class are 100 times or 1000 times or 10000 times less than that in another class (Liu et al., 2009). The imbalance level can be as high as $10^6$. In this research, the class ratio is 1:300, i.e. for each landslide pixel we have more than 300 non-landslide pixels. When we use various machine-learning approaches for the generation of LSZ maps then the algorithms do not classify the landslide pixels correctly. Therefore, it is necessary to reduce the imbalance in the susceptibility mapping data. There are two major data balancing techniques, which are oversampling of a minority class and under-sampling of majority class (He, Garcia, 2009). The minority oversampling cannot be applied, as it will create false landslide pixels. We under-sample the majority class (i.e., non-landslide pixels) using Balance Cascade and Easy Ensemble methods. Some of the techniques used by various authors to overcome data imbalance are random oversampling and under sampling, informed under sampling, Synthetic sampling with data generation etc. (Haixiang et al., 2017, Stumpf et al., 2014, Stumpf et al., 2012, Galar et al., 2012, Chawla, 2010, Liu et al., 2009, He, Garcia, 2009).

This work aims at first, balancing data using two different data balancing techniques i.e. EasyEnsemble and BalanceCascade. This balanced data is used for computing the weights using various methods methods such as fisher discriminant analysis (FDA), logistic regression (LR) and artificial neural network(ANN) to generate LSZ maps. Furthermore visual analysis, statistical quantities, Heidke Skill Score (HSS) and Recall is used to assess the quality of susceptibility maps. Based on these observations, We can make assertions as how accuracy is affected when balancing techniques are applied to our data w.r.t. imbalanced data (Jhunjhunwalla et al., 2019).

## 2. STUDY AREA & DATA RESOURCES

A small part of Mandakini river basin of Garhwal Himalaya in Uttarakhand has been considered for the study as shown in Figure-1. Mandakini river originates from the Chorabari Glacier near Kedarnath in Uttarakhand, India. The study area covers about 275.60 sq. km area and lies between 30°19′00″N to 30°49′00″N latitude and 78°49′00″E to 79°21′13″E longitude. The study area falls in the Survey of India toposheet no. 53J and 53N. This region is highly prone to landslides during the monsoon season. This region has highly rugged topography, deep

*Corresponding author

gorges, high peaks where higher areas are mostly snow-covered forming the U-shaped wide valleys of glacial landscape. The study area is highly prone to landslide as every year many landslides occur in this area. In this research we have used 30m shuttle radar topography mission (SRTM) DEM, which was downloaded from EarthExplorer (www.earthexplorer.usgs.gov).

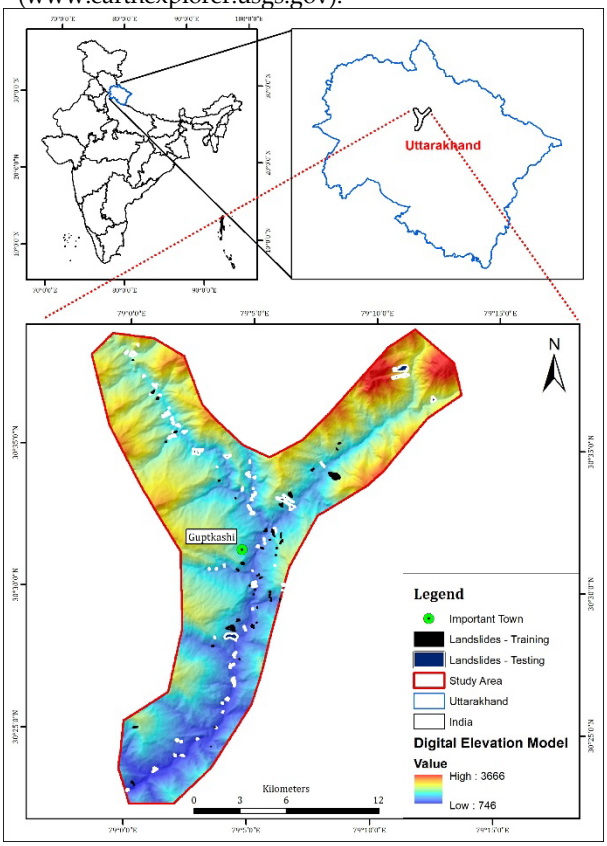


Figure 1. Map of study area (Source: Boundary of India and Uttarakhand provided by Survey of India, DEM provided by USGS)

## 3. DATA PREPARATION & METHODS

There are total 122 landslides, which have occurred between 2004 and 2017. The landslides occurred within 2004 - 2012 (46 landslides with 1203 pixels) are used for training of the models and 2013 - 2017 (76 landslides with 2743 pixels) are used for testing of the models. A landslide inventory is prepared to show landslide occurrences. The landslide inventory is binary in nature where one ('1') shows the occurrences of landslide and zero ('0') shows non-occurrences of landslide. The inventory was prepared manually from the past satellite images in GIS environment. This task is limited by the resolution of satellite images.

The landslide causative factors can be either categorical, which can be classified i nto fi nite nu mber of groups/classes (e.g. soil types in an area), or continuous (e.g. slope or elevation of the mountain). In this study, we have used only continuous data. Seven causative factors/layers i.e. aspect, slope, topographic wetness index, elevation, profile c urvature a nd p lan c urvature are considered for preparation of LSZ maps. These layers have been prepared from 30 m spatial resolution SRTM elevation model using ArcGIS and QGIS software. The size of all the layers is 1028 × 801 pixels. The layers are converted into ASCII format for training of the models. These layers have been transformed to column vector of size 823428× 1 and are stacked to generate a matrix of size 823428×7. In this work, we have used three algorithms i.e. FDA, LR and ANN for finding the weights o f various factors, which are further used for finding susceptibility index using weighted linear combinations.

### 3.1 Data Balancing Algorithms

The following data balancing algorithms have been applied to the initial experimental data set to obtain a balanced data set.

3.1.1 EasyEnsemble This method works on samples of majority class. It reduces the number of observations from majority class to make the data set balanced. This method is best to use when the data set is huge and reducing the number of training samples helps to improve run time and storage troubles. EasyEnsemble is an example of informed under sampling as it explore subsets of majority class by independent replacement of subsets (He, Garcia, 2009, Liu et al., 2009). This method is an unsupervised learning algorithm as it explores subsets of majority class by independent random sampling with replacement.

3.1.2 BalanceCascade In this method we create several subsets of data which are balanced, and a weak classifier is trained for each subset. This method reduces the majority class training sets at every step by removing all the examples that are correctly classified. I t i s d ifferent from Easy Ensemble in two steps. Primarily the weights are adjusted based on false positive rates that a classifier have to achieve. Second, the samples are removed which are correctly classified. This sequential dependence mainly focuses on reducing the redundant information in majority class.
In Balance Cascade, the training can finally b e stopped

when size of majority class( $M$ ) is less than size of minority class( $N$ ), as size of majority class is getting shrunk at every iteration. The main advantage of Balance Cascade is that it generate the restricted sample space to extract as much useful information possible (Liu et al., 2009).

### 3.2 Methods of Weights Assignment

The following filter and wrapper methods have b een used for finding the weights of factors.

3.2.1 Fisher Discriminant Analysis (FDA): This method is used in pattern recognition, statistics and machine learning for finding t he l inear c ombination of features to distinguish two or more classes of events or objects (Jhunjhunwalla et al., 2019). In this method, all the factors/layers are projected in one dimension corresponding to landslide occurrences. A multiplicative factor is required for projection of the data. This multiplicative factor is used for giving weights to all the thematic layers (Gupta et al., 2018).

3.2.2 Logistic Regression (LR): It is a special case of linear regression and predicts the probability of the occur-rence of an event by using logit function (Gupta et al., 2018). In this method, the probability of presence or ab-sence of a binary outcome (1 = landslide and 0 = no land-slide) is modelled based on the values of predictor variables (Shukla et al., 2016). The independent variable can be in-terval or categorical while the dependent variable can be multinomial or binary. The LR coefficients a re u sed for giving weights to all the factors.

3.2.3 Artificial Neural Network (ANN): ANN is a computational system, which is inspired by the human brain. The network consists of set of neurons, an input layer, an output layer and few hidden layers. The layers are interconnected and input data is passed to output layer by means of the hidden layers. There can be one or more hidden layers depending on the complexity of data. It is useful when a complex relationship exists between the data and the responses such as between landslide factors and landslide occurrences (Jhunjhunwalla et al., 2019).

ANN require several architectural and training parameters to be selected prior to analysis. The optimal number of hidden layers and the number of neurons per hidden layer are not known apriori. These parameters are empirically determined through rigorous experimentation and examination of different p arameter settings (Blackard, Dean, 1999).

### 3.3 Landslide Susceptibility Index Computation and Susceptibility Mapping

The weights obtained in previous section are used for computation of landslide susceptibility index (LSI). LSI is calculated using Weighted linear combination method (Jhunjhunwalla et al., 2019, Gupta et al., 2018, Michael, Samanta, 2016). LSI can be calculated as given in Eq-5

$$LSI = \sum attributes * weights \tag{5}$$

LSI can be classified into five different zones i.e LSZ (from very high susceptibility to very low susceptibility) based on natural break in the data. Results obtained with/without data balancing are compared using statistical quantities and visual quality analysis. The methodology used in the current study has been shown in Figure-2.

The EasyEnsemble and BalanceCascade methods was applied on total 30 subsets, whihch were randomly taken from the non-landslide pixels by the algorithm in python. After the data balancing, the FDA, LR and ANN methods were applied on these 30 subsets of data, which resulted in 30 LSI images for each of the methods. The mean and median of these 30 LSI images for each method was taken to generate mean image and median image for all the methods.

### 3.4 Accuracy Assessment

3.4.1 Heidke Skill Score: The accuracy of the LSZ maps is measured using Heidke Skill Score (HSS). It is the measure of skill of prediction and lies between 0 and 1. HSS can be defined as given in the following Eq-6 (NDFD Verification Score Definitions, 2017, Hyvärinen, 2014).

$$HSS = \frac{NC - E}{T - E} \tag{6}$$

where $NC$ is the number of correct predictions, i.e. number of times the prediction and observation match, $E$ is the number of predictions expected to verify based on chance, i.e. incorrect predictions and $T$ is the total number of observations.

3.4.2 Recall: Recall can be defined as the ratio of relevant instances (landslides) predicted correctly by the model to actual number of relevant instances. It is also known as “sensitivity”. This gives us a measure of how accurately the model predicts with respect to the actual instances of that class. We can compute recall using the expression in Eq-7

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{7}$$

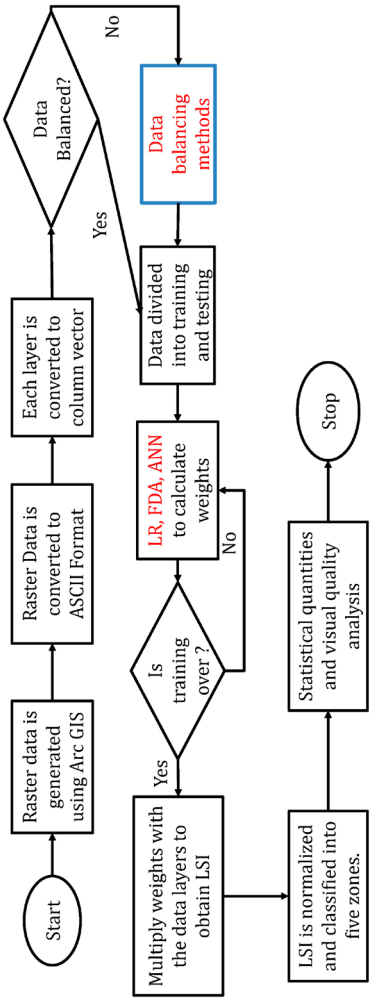


Figure 2. Flowchart of the study

## 4. RESULTS AND DISCUSSIONS

### 4.1 With Imbalanced Data

The weights and ranking of factors is important for preparation of landslide susceptibility zonation maps. Weights and ranking obtained by FDA, LR and ANN for all seven causative factors without using the data balancing algorithms are given in Table-1 (Jhunjhunwalla et al., 2019).

Table 1. Weights calculated from different models using imbalanced data

| Causative Factors | LR | FDA | ANN |
|---|---|---|---|
| Aspect | -1.11 | 0.31 | 0.66 |
| DEM | -3.72 | -3.69 | 0.38 |
| Plan Curvature | 1.00 | 4.59 | 0.17 |
| Profile Curvature | -2.35 | 0.80 | 0.05 |
| Slope | 5.04 | 3.36 | 0.36 |
| Tangential Curvature | 0.51 | -3.49 | -0.43 |
| TWI | 2.42 | 0.72 | 0.72 |

The weights obtained in Table-1 are multiplied with the corresponding causative factor layer to obtain the LSI values. The LSI values are normalized between 0 and 1 and classified into five different zones i.e. from very high susceptibility to very low susceptibility (as shown in Figure-3, 4(a) and 5(a)) based on natural break in the data. The mean and median of LSI values for all the methods (except ANN) is observed to be near 0.55 (refer Table-2) so 0.55 is set as a threshold for classification.

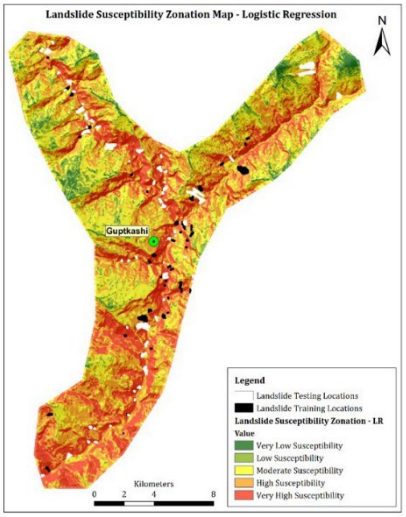


Figure 3. LSZ map obtained using weights from LR, without data balancing

The landslides with LSI greater than 0.55 are considered to be correctly classified and below 0.55 are considered to be falsely classified. As we can see from Table-2 that ANN has mean/median values significantly less than 0.55 and hence the susceptibility maps generated by ANN have maximun area lying in low or very low susceptibility zones. Hence, after data balancing, the mean/median value of LSI for ANN should increase significantly.

Table 2. Statistics (mean, median & standard deviation) for all the three methods using imbalanced data

| Method | Mean | Median | Standard Derivation |
|---|---|---|---|
| LR | 0.58 | 0.58 | 0.11 |
| FDA | 0.55 | 0.56 | 0.12 |
| ANN | 0.43 | 0.42 | 0.17 |

### 4.2 With Balanced Data

The imbalanced data is provided to EasyEnsemble and BalanceCascade algorithms and the data is balanced to match the size of minority class pixels. The feature selection algorithms are applied to balanced data and the weights are obtained. These weights are multiplied with the different l ayers a nd s usceptibility i ndex m ap are generated. The statistical quantities for LSI using all three methods are given in Table-3. The mean and median for ANN has increased significantly u sing balanced data, however, the values using LR has been reduced and mean/median is very small compared to that without balancing the data. The mean/median obtained for LSI values using FDA does not show significant c hanges. The susceptibility maps obtained using balanced data are shown in Figure-4(b) and 5(b).

The results of accuracy assessment using HSS and Recall are given in Table-4 & 5. The threshold value of 0.55 was considered for calculating the number of correctly classified l andslides ( for c omputation o f H SS) a nd n umber of correctly classified p ixels o f t hose l andslides ( for computation of recall). It is evident from Table-4 & 5 that LSZ maps prepared using mean of weights in ANN with the data balanced using EasyEnsemble method gives the highest Score and Recall value and hence the highest accuracy of landslide zonation. Results obtained using FDA

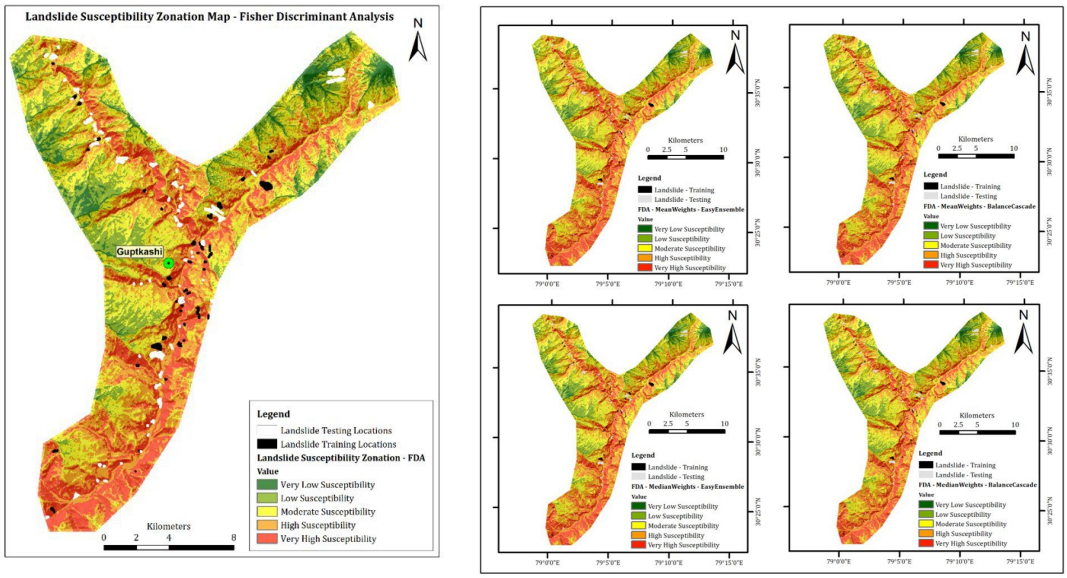


Figure 4. LSZ map obtained using weights from FDA, without data balancing (a) and with data balancing (b)

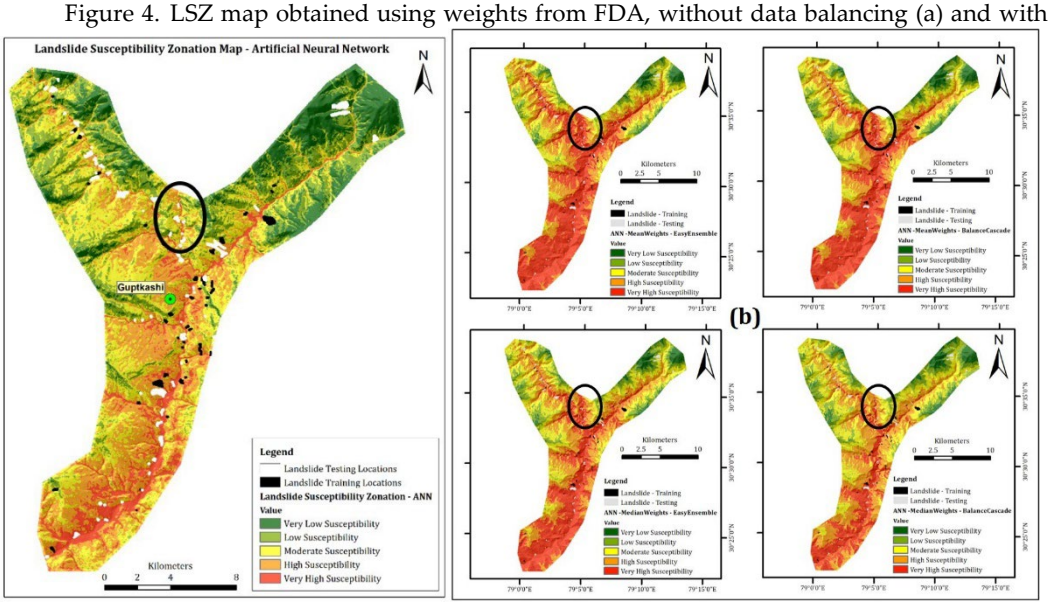


Figure 5. LSZ map obtained using weights from ANN, without data balancing (a) and with data balancing (b). The sample location with significant changes are marked with black circle.

Table 3. Statistics (mean, median and standard deviation) for all the three methods using balanced data

| Balancing Method | Statistical Quantities | LR | | FDA | | ANN | |
|---|---|---|---|---|---|---|---|
| | | Mean Image | Median Image | Mean Image | Median Image | Mean Image | Median Image |
| Easy Ensemble | Mean | 0.2125 | 0.3834 | 0.5368 | 0.5558 | 0.6147 | 0.5822 |
| | Median | 0.2141 | 0.3870 | 0.5684 | 0.5604 | 0.6289 | 0.5948 |
| | Standard Deviation | 0.0348 | 0.0775 | 0.1140 | 0.1163 | 0.1442 | 0.1364 |
| Balance Cascade | Mean | 0.2337 | 0.2934 | 0.5568 | 0.5518 | 0.5926 | 0.5455 |
| | Median | 0.2358 | 0.2960 | 0.5614 | 0.5562 | 0.6064 | 0.5582 |
| | Standard Deviation | 0.0438 | 0.0565 | 0.1149 | 0.1151 | 0.1475 | 0.1268 |

also shows comparable HSS and recall values. The zonation of LSI obtained using LR can not prepared due to the smaller mean values. The natural break limits for different zones can not be applied on it. The results obtained in this study validates the fact that LR does not require the data balancing, given that we have sufficient samples in positive class (Crone, Finlay, 2012, King, Zeng, 2001). The FDA method may or may not not show major changes in the results with/without data balancing, which also agree with the results of few studies (Xue, Hall, 2015, Xue, Titterington, 2008).

## 5. CONCLUSIONS

Data balancing methods improves accuracy for machine learning based methods such as ANN, Support vector machine etc. The EasyEnsemble method coupled with mean of weights seems to overpredict the high susceptibility zones whereas BalanceCascade method with median of

Table 4. Accuracy assessment of FDA and ANN on balanced data using Heidke Skill Score for total 76 landslides from 2013 to 2017

| Balancing Method | Methods | Correctly Classified Landslides (LSI >= 0.55) | Wrongly Classified Landslides (LSI <0.55) | HSS |
|---|---|---|---|---|
| EasyEnsemble | FDA Mean | 73 | 3 | 0.9589 |
| | FDA Median | 70 | 6 | 0.9143 |
| | ANN Mean | 73 | 3 | 0.9589 |
| | ANN Median | 73 | 3 | 0.9589 |
| BalanceCascade | FDA Mean | 69 | 7 | 0.8986 |
| | FDA Median | 69 | 7 | 0.8986 |
| | ANN Mean | 72 | 4 | 0.9444 |
| | ANN Median | 68 | 8 | 0.8824 |

Table 5. Accuracy assessment of FDA and ANN on balanced data using Recall for total 2743 pixels in 76 landslides from 2013 to 2017

| Balancing Method | Methods | Correctly Classified Pixels (LSI >= 0.55) | Wrongly Classified Pixels (LSI <0.55) | Recall |
|---|---|---|---|---|
| EasyEnsemble | FDA Mean | 2192 | 551 | 0.7991 |
| | FDA Median | 2114 | 629 | 0.7707 |
| | ANN Mean | 2254 | 489 | 0.8219 |
| | ANN Median | 2198 | 545 | 0.8013 |
| BalanceCascade | FDA Mean | 2125 | 618 | 0.7747 |
| | FDA Median | 2074 | 669 | 0.7561 |
| | ANN Mean | 2205 | 538 | 0.8039 |
| | ANN Median | 2011 | 732 | 0.7331 |

weights generated by ANN gives most appropriate LSZ map based on visual analysis. Using statistical quantities, the LSI generated from both the data balancing methods show mean and median value greater or very near to 0.55, which also justifies the importance of data balancing before using ANN. The HSS and recall value shows the superiority of LSZ maps prepared using mean of weights in ANN with the data balanced using EasyEnsemble method. The Balanced data do not show good results with Logistic regression as the LR method is not able to model the underlying probability distribution. Balanced data does not effect F DA m ethod a nd h ave a pproximately s ame accuracy as before. The landslide data is highly imbalanced in nature, hence balancing algorithms must be applied before preparation of LSZ maps using machine learning methods. However the data driven methods do not need balancing as seen from the results.

## REFERENCES

Blackard, J. A., Dean, D. J., 1999. Comparative accuracies of artificial n eural n etworks a nd d iscriminant a nalysis in predicting forest cover types from cartographic variables. Computers and Electronics in Agriculture, 24(3), 131–151.

Chawla, N. V., 2010. Data Mining for Imbalanced Datasets: An Overview. O. Maimon, L. Rokach (eds), Data Mining and Knowledge Discovery Handbook, 2nd edn, Springer US.

Crone, S. F., Finlay, S., 2012. Instance sampling in credit scoring: An empirical study of sample size and balancing. International Journal of Forecasting, 28(1), 224–238.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 463–484.

Gupta, S. K., Shukla, D. P., Thakur, M., 2018. Selection of weightages for causative factors used in preparation of landslide susceptibility zonation (LSZ). Geomatics, Nat. Hazards Risk, 9(1), 471–487.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220–239.

He, H., Garcia, E. A., 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

Hyvärinen, O., 2014. A Probabilistic Derivation of Heidke Skill Score. Weather and Forecasting, 29(1), 177-181.

Jhunjhunwalla, M., Gupta, S. K., Shukla, D. P., 2019. Landslide Susceptibility Zonation (LSZ) Using Machine Learning Approach for DEM Derived Continuous Dataset. K. Santosh, R. S. Hegadi (eds), Recent Trends on Image Processing and Pattern Recognition, Springer Singapore, 505–519.

King, G., Zeng, L., 2001. Logistic Regression in Rare Events Data. Political Analysis, 9(2), 137–163.

Liu, X.-Y., Wu, J., Zhou, Z.-h., 2009. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man and Cybernetics, 39(2), 539–550.

Michael, E. A., Samanta, S., 2016. Landslide vulnerability mapping (LVM) using weighted linear combination (WLC) model through remote sensing and GIS techniques. Model. Earth Syst. Environ., 2(2), 1–15.

NDFD Verification Score Definitions, 2017.

Pradhan, B., Abokharima, M. H., Jebur, M. N., Tehrany, M. S., 2014. Land subsidence susceptibility mapping at Kinta Valley (Malaysia) using the evidential belief function model in GIS. Nat. Hazards, 73(2), 1019–1042.

Roodposhti, M. S., Aryal, J., Pradhan, B., 2019. A novel rule-based approach in mapping landslide susceptibility. Sensors (Switzerland), 19(10), 2274.

Shukla, D. P., Gupta, S., Dubey, C. S., Thakur, M., 2016. Geo-spatial Technology for Landslide Hazard Zonation and Prediction. M. Marghany (ed.), Environmental Applications of Remote Sensing, InTech, Rijeka, Coratia, 281–308.

Stumpf, A., Lachiche, N., Kerle, N., Malet, J.-P., Puissant, A., 2012. Adaptive spatial sampling with active random forest for object-oriented landslide mapping. 2012 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 87–90.

Stumpf, A., Lachiche, N., Malet, J.-P., Kerle, N., Puissant, A., 2014. Active Learning in the Spatial Domain for Remote Sensing Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 52(5), 2492–2507.

Xue, J. H., Hall, P., 2015. Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis? IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(5), 1109–1112. http://ieeexplore.ieee.org/document/6906278/.

Xue, J.-H., Titterington, D. M., 2008. Do unbalanced data have a negative effect on LDA? Pattern Recognition, 41(5), 1558–1571.