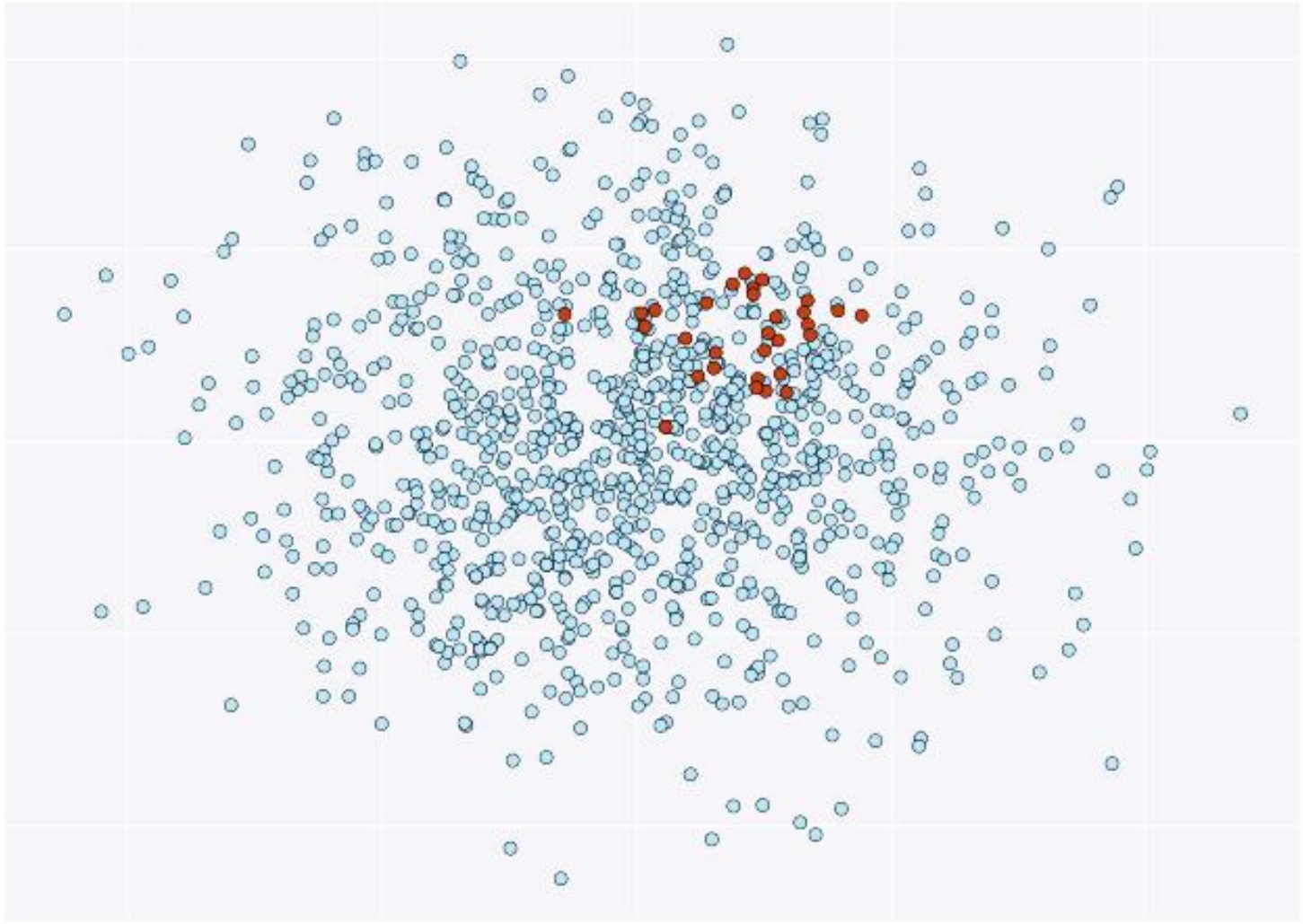
Solving Data Imbalance in Landslide Susceptibility Zonation

Sharad K. Gupta, Dericks P. Shukla

INTRODUCTION

- Disproportionate ratio of observations in various classes.
- The data is imbalanced when the class ratio is of the order of 1:100, 1:1000 and 1:10000 (i.e., number of points in one-class are 100 times or 1000 times or 10000 times less than that in another class).



Blue – Non-Landslide
Red – Landslide Point

METHODS

Two major data balancing techniques

Over-sampling of a minority class	Under-sampling of majority class

Under sampling Methods (Liu et al. 2009)

Random under-sampling	Tomek links	Cluster centroids	Balance Cascade	Easy Ensemble

- EasyEnsemble Algorithm
- BalanceCascade Algorithm

} **Data Balancing**

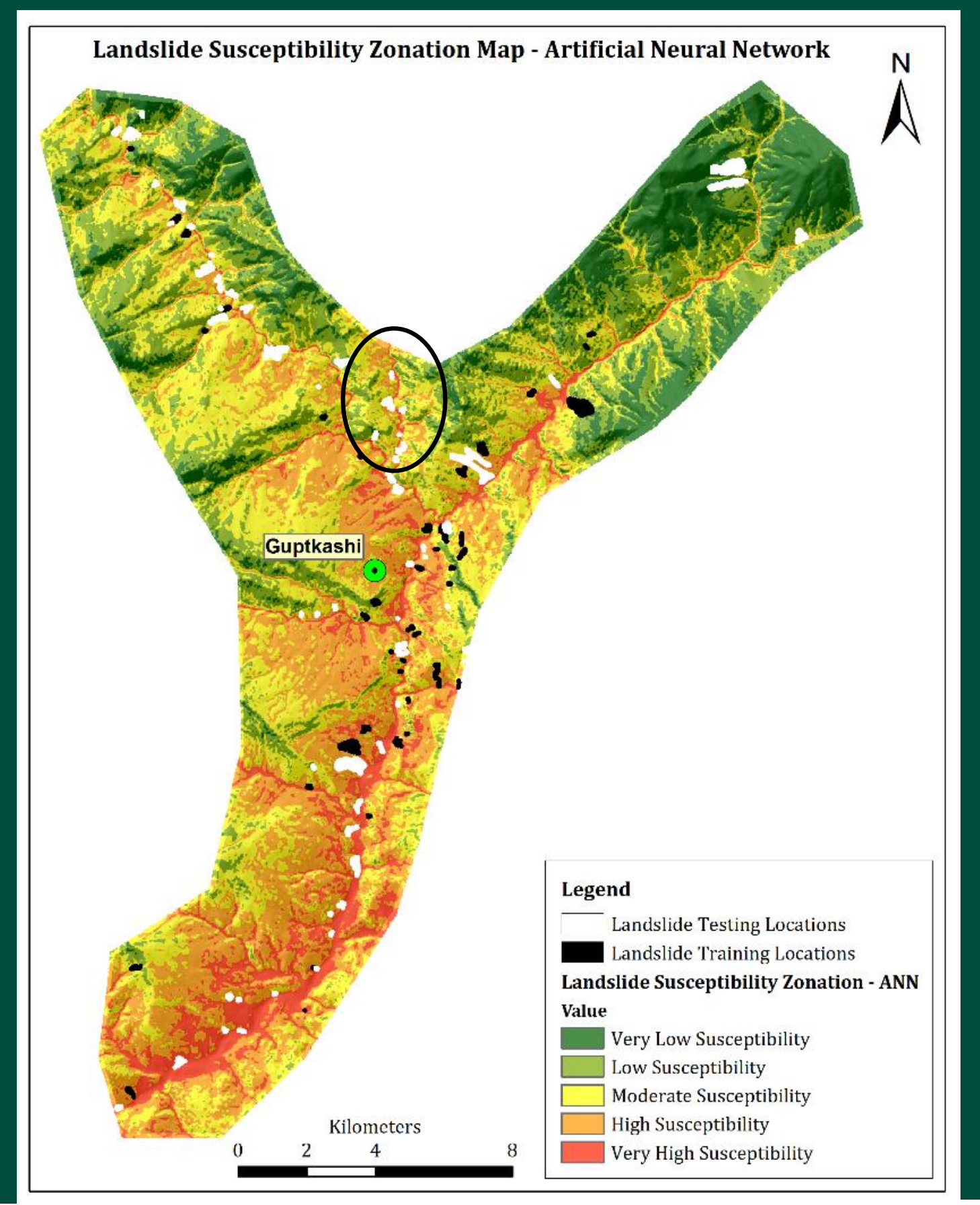
- Fisher Discriminant Analysis
- Logistic Regression
- Artificial Neural Network

} **Weightage Determination**

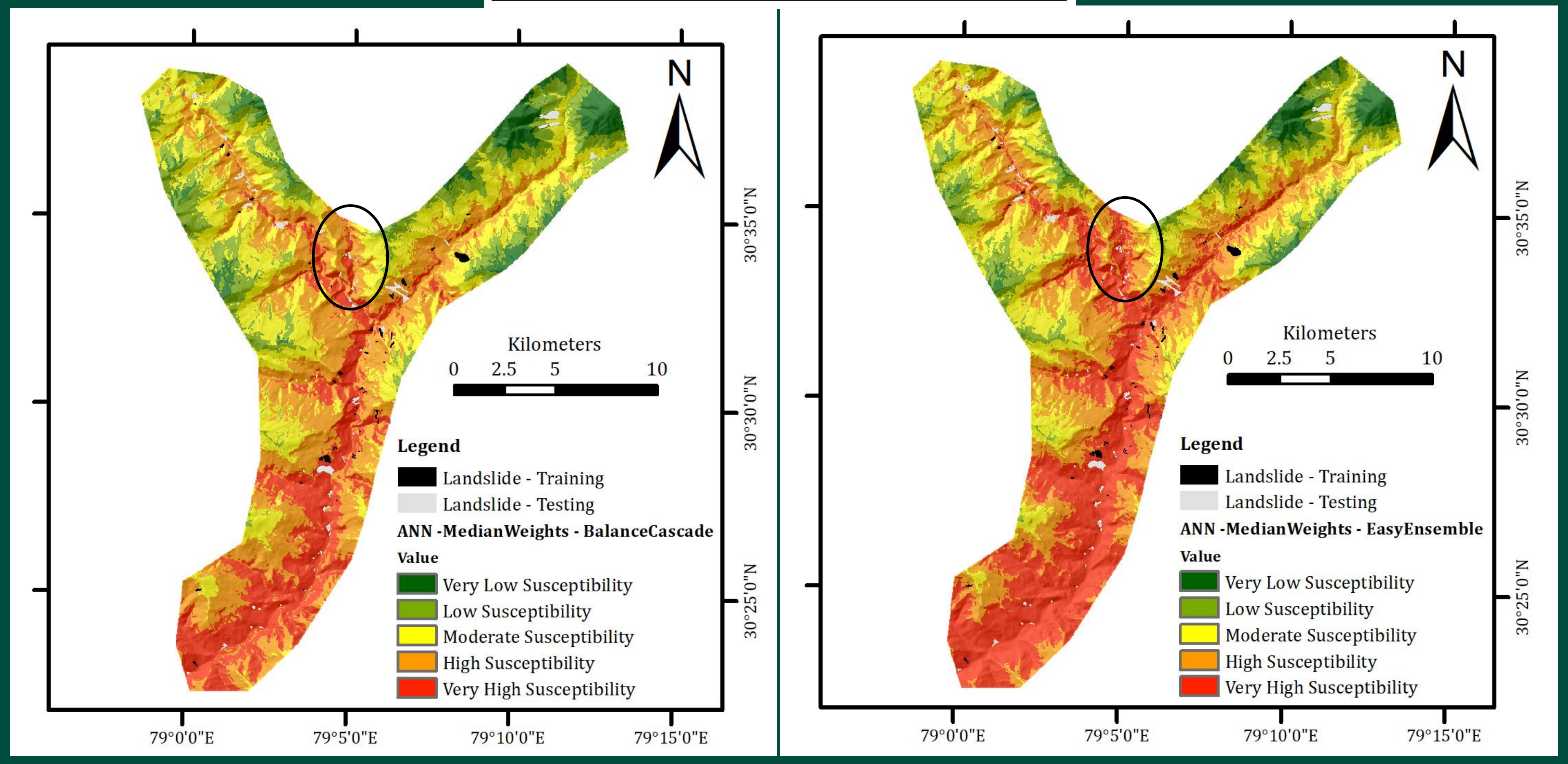
- Heidke Skill Score
- Recall (Sensitivity)

} **Accuracy Assessment**

Machine learning methods require data balancing whereas data driven methods do not need balancing.


Landslide Susceptibility Zonation Map - Artificial Neural Network
N
Guptkashi
Legend
Landslide Testing Locations
Landslide Training Locations
Landslide Susceptibility Zonation - ANN
Value
Very Low Susceptibility
Low Susceptibility
Moderate Susceptibility
High Susceptibility
Very High Susceptibility
Kilometers
0
2
4
8

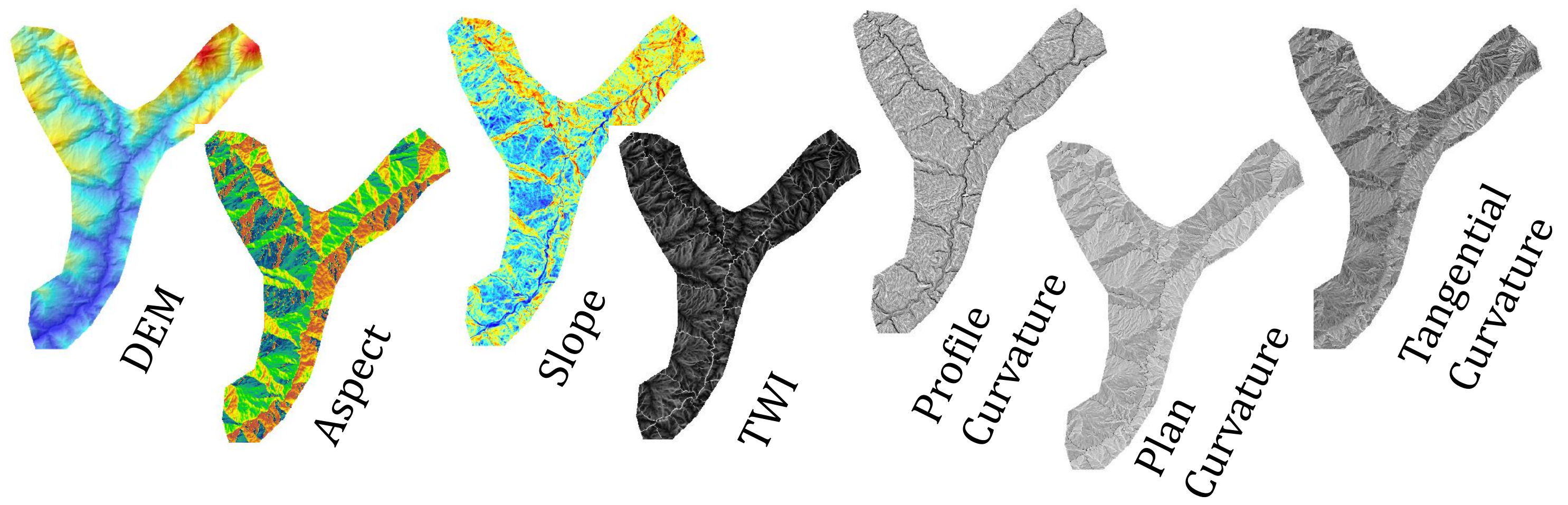
Imbalanced data



Balanced Data

Take a picture to download the **full paper**

DATA PROPERTIES



- The study area comprises of total 122 landslides occurred between 2004 and 2017
- Training - 46 landslides (1203 pixels) occurred from 2004 to 2012, Testing - 76 landslides (2744 pixels) occurred from 2013 to 2017

RESULTS

Table - 1: Statistics for all the three methods **(imbalanced data)**

Method	LR	FDA	ANN
Mean	0.58	0.55	**0.43**
Median	0.58	0.56	**0.42**
Standard Deviation	0.11	0.12	**0.17**

Table 2. Statistics for all the three methods **(Balanced data)**

Balancing Method	Statistical Quantities	LR	FDA	ANN
Easy Ensemble	Mean	0.3834	0.5558	0.5822
	Median	0.3870	0.5604	0.5948
	Std. Dev.	0.0775	0.1163	0.1364
Balance Cascade	Mean	0.2934	0.5518	0.5455
	Median	0.2960	0.5562	0.5582
	Std. Dev.	0.0565	0.1151	0.1268

Decreased No significant change Improved Significantly

CONCLUSIONS

- LR method is not able to model the underlying probability distribution after data balancing.
- The FDA method may or may not show major changes in the results after data balancing.
- 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.

ACKNOWLEDGEMENTS

I would like to heartily thank AGU for providing Austin Travel Grant for attending the fall meeting. I am also very grateful to HIMCOSTE for providing Partial Travel Support for presenting the paper.