A Comparison of Class Imbalance Techniques for Real-World Landslide Predictions

Kapil Agrawal[1], Yashasvi Baweja[2], Deepti Dwivedi[3], Ritwik Saha[1], Prabhakar Prasad[1], Shubham Agrawal[4], Sunil Kapoor[1], Pratik Chaturvedi[5], Naresh Mali[1], Venkata Uday Kala[1], Varun Dutt[1]

[1] Applied Cognitive Science Lab, Indian Institute of Technology Mandi (IIT Mandi), Himachal Pradesh, India

[2]Indraprastha Institute of Information Technology (IIIT Delhi), New Delhi, India

[3] National Institute of Technology Srinagar (NIT Srinagar), Jammu and Kashmir, India

[4] Center for Converging Technologies, University of Rajasthan, Jaipur, Rajasthan, India

[5]Defence Terrain Research Laboratory, Deference Research and Development Organization (DRDO), New Delhi, India

***Abstract*—Landslides cause lots of damage to life and property world over. There has been research in machine-learning that aims to predict landslides based on the statistical analysis of historical landslide events and its triggering factors. However, prediction of landslides suffers from a class-imbalance problem as landslides and land-movement are very rare events. In this paper, we apply state-of-the-art techniques to correct the class imbalance in landslide datasets. More specifically, to overcome the class-imbalance problem, we use different synthetic and oversampling techniques to a real-world landslide data collected from the Chandigarh – Manali highway. Also, we apply several machine-learning algorithms to the landslide data set for predicting landslides and evaluating our algorithms. Different algorithms have been assessed using techniques like the area under the ROC curve (AUC) and sensitivity index (d'). Results suggested that random forest algorithm performed better compared to other classification techniques like neural networks, logistic regression, support vector machines, and decision trees. Furthermore, among class-imbalance methods, the Synthetic Minority Oversampling Technique with iterative partitioning filter (SMOTE-IPF) performed better than other techniques. We highlight the implications of our results and methods for predicting landslides in the real world.**

***IndexTerms*—Landslides, Class-imbalance, SMOTE, SMOTE-IPF,Random Forest, Sensitivity index, AUC.**

I. Introduction

The Himalayan Region has been prone to landslide hazards mainly due to its precarious topographic characteristics and tectonic dynamism [14]. Landslide causes damage to property, blocks roads for days and kills and traps people. For places at very high altitude where everything from food to clothing is imported from cities, blocking roads is a huge issue. These reasons become very supporting evidence for combating this problem at the earliest.

Recently, landslide prediction using machine-learning techniques have been gaining a lot of popularity [13]. Machine-learning algorithms have enabled researchers to not only predict landslides in advance; but, also improve the understanding of causal factors that trigger these natural disasters [13]. While there are several factors responsible for triggering landslides; rainfall and land geology form the most important factors compared to other factors like earthquakes and anthropogenic influences [14]. Sixty percent of the total landslides that occurred in the Himalayas in 2010 were in the monsoon months extending from mid-June to mid-September [14].

Prior research [14, 21, 29] has investigated the performance of several machine-learning classifiers on landslide predictions using the historic rainfall-intensity as a primary predictor. However; little research has been done towards the problem of class-imbalance that landslide datasets inherently suffer from. Landslide activity is a rare event: In a typical dataset, the numbers of landslide days are likely to be very few compared to non-landslide days and this fact makes the positive (landslide occurrence) class very small compared to the negative class. This imbalance among the two classes would likely make the classifier biased towards the majority class leading to classifying all the instances in the dataset as belonging to the majority class [11]. Thus, the accuracy of the classifier due the majority class is very high and this problem is widely known as the class-imbalance problem. The main objective of this paper is to evaluate and compare several class-imbalance techniques based upon synthetic or random oversampling to reduce the problem of class-imbalance in predicting landslides.

To the best of authors' knowledge, up to now, class-imbalance and its influence has not been studied for landslide predictions. Although the application of class-imbalance techniques for landslide prediction is new, attempts have been made in prior literature on applying these techniques to other real-world applications (fraud detection,lung cancer detection,

emotion classification, text classification) for mitigating class-imbalance using a number of class-imbalance techniques. The simplest class-imbalance technique is to resample the original dataset. In resampling, the dataset is modified before applying machine-learning algorithms. The simplest form of resampling is random oversampling [24]. Random oversampling generates new minority class-instances in the dataset by randomly selecting any one minority instance and duplicating it till the dataset is balanced, i.e., there is equal number of instances for both the positive and negative class [24]. Beyond random oversampling, there are other sophisticated forms of resampling proposed in the literature which use a more-focused approach for synthesizing new instances near existing minority class instances. Synthetic Minority Oversampling Technique (SMOTE) and Synthetic Minority Oversampling Technique-Iterative Partitioning Filter (SMOTE-IPF) [23, 15] are popular sophisticated forms of resampling techniques that use the k-nearest neighbour algorithm to synthesize new instances [23]. These techniques have been used in prior literature in improving emotion classification problems and lung-cancer detection and have proven to improve the accuracy of the classifier [22, 31]. In this study, we investigate the performance of machine-learning algorithms with and without resampling, where these techniques are applied to a real-world landslide prediction dataset. Furthermore, we evaluate the optimal value of the k parameter in the k-nearest neighbour algorithm for both SMOTE and SMOTE-IPF techniques.

In what follows, we provide a brief overview of the research that has been conducted in the past for addressing the class-imbalance followed by the research involving machine-learning for landslide prediction. Next, we discuss the study area and data that has been used for this study. Then, we provide a brief explanation of each of the resampling techniques that we have employed in this study. Finally, we apply machine-learning algorithms with and without class imbalance techniques and close the paper by highlighting the implications of our results for landslide predictions using class-imbalance methods.

II. Previous work

The problem of learning from imbalanced data sets has been intensively researched in recent years, and several methods have been proposed to address it [15]. For example, resampling is a classifier-independent method that modifies the data distribution considering local characteristics of instances to change the balance between two or more classes. The most common method is random oversampling [24]. Random oversampling works by selecting a minority class instance randomly and replicating it till the desired balance between classes is reached. In [24], it is suggested that random oversampling has been effective across 125 different synthetic datasets. Beyond random oversampling, SMOTE is one of the most well-known classes imbalance techniques: it generates new artificial minority class examples by interpolating among several existing minority class examples that are similar to each other. In [23], it is showed that among the nine datasets considered in their study from different domains, SMOTE performed better than the random oversampling method in six datasets. SMOTE does not blindly generate random instances in a dataset, and it uses a more focused resampling method that helps it improve overall classification performance. While SMOTE outperforms random oversampling, however; some researchers have shown that the class-imbalance is not a problem itself [15]. The classification performance degradation is usually linked to other factors related to data-distributions. Among them, the influence of noisy and borderline examples on classification performance in an imbalanced dataset has been observed [17]. Borderline instances are defined as instances located either very close to the decision boundary between the minority and majority classes or located in the area surrounding class boundaries where classes overlap. The authors refer to noisy examples as those from one class located deep inside the region of the other class of [17, 19]. The SMOTE-IPF method addresses the problem with these borderline and noisy examples present in the dataset, which SMOTE ignores [15]. SMOTE-IPF is a two-step approach. First, instances are generated using SMOTE algorithm. Second, all the noisy and borderline instances are cleaned from the dataset by using ensemble-filtering based technique called the Iterative Partitioning Filter.

Furthermore, there have been several studies that have used the state-of-the-art machine learning technique in the landslide domain [9, 13, 14]. In [14] the author uses the logistic regression model in their study to predict the slope-failure initiation using the antecedent 30-day and 15-day rainfall. This logistic regression model is further validated through the Receiver Operating Characteristic (ROC) curve analysis using a set of samples which had not been used for training the classifier. The model showed an accuracy of 95.1%. Furthermore, many prior investigations have compared several machine-learning classifiers for prediction of landslides [4, 9, 10, 13]. However, little research has taken place that investigates how certain class-imbalance techniques improve predictions in real-world landslide datasets. Since landslide occurrence is a rare event, the positive (landslide occurred) class has comparatively fewer instances compared to negative (landslide did not occur) class. This can be a major problem, especially in the landslides prediction domain because a classifier can get a very high accuracy simply by predicting all samples to belong to the majority class.

The primary objective of this paper is to compare different oversampling techniques in their ability to reduce class-imbalance and ensure the accuracy of landslide predictions. As part of our evaluation, we consider three different methods, Random Oversampling, SMOTE, and SMOTE-IPF, to reduce the class-imbalance problem. We expect that SMOTE-IPF should perform better than the other two class-imbalance methods

because of its filtering of noisy and borderline instances [15]. Also, we expect that both SMOTE variants would perform better compared to the random oversampling technique as these SMOTE variants use a more focused oversampling approach compared to the random oversampling method. Finally, we also expect that the landslide classification performance would be higher in conditions where class-imbalance methods are used compared to conditions where these methods are not used. For the landslide prediction task, we use some machine-learning algorithms that include Logistic Regression [1], Decision Trees [16], Support Vector Machines [30], Random Forests [32] and Multilayer Perceptron [4]. The choice of these machine-learning algorithms is based upon their use for landslide predictions in the literature [4, 9, 10, 13].

In the next section, we explain the study area and the data used. In the subsequent sections, we provide details of the compilation of the data set followed by a brief overview of different class-imbalance techniques and the various machine-learning algorithms used in this paper.

III. Methodology

A. Data Compilation and Preparation

The dataset used in this paper corresponds to the National Highway-21 (NH-21) of India in a stretch extending from Mandi to Manali. As part of the dataset, we attempt to predict landslides based on the antecedent 30-day rainfall and land susceptibility. These attributes were also used by [14]; however, for a different region corresponding to the Uttarakhand state in India (along with NH 58). The dataset has been built from three distinct sources: Himachal Pradesh Public Works Department (HPPWD), Indian Meteorological Department (IMD), and Indian Space Research Organization (ISRO).

The HPPWD provided the historical landslide occurrence data between 2011 and 2015 along NH-21 between Mandi and Manali towns. The IMD provided the rainfall information in the region of interest in a latitude-longitude format. Finally, for each location corresponding to the HPPWD landslide dataset, a landslide susceptibility was computed from Very Low to Severe using Indian Space Research Organization' Bhuvan website [3].

Figure 1 shows the frequency of landslide along the NH 21 highway between Mandi and Manali towns using the HPPWD data. The horizontal axis denotes the distance (in km) of the landslide occurrence event from Chandigarh, i.e., the origin of NH-21. The vertical axis shows the frequency of landslides between 2011 and 2015 along NH 21. The 200-km milestone refers to the inter-state bus terminal at Mandi, and the 305-km milestone corresponds to the inter-state bus terminal at Manali.

We changed this milestone-based dataset into a latitude-longitude – based dataset since the rainfall and susceptibility data were in the latitude-longitude format. A two-step approach was followed for achieving this data conversion. First, all the milestones along the NH-21 were mapped to their corresponding latitude and longitude positions. Second, we found the mid-point milestone for all landslides occurring along NH-21. Next, based on the mapping in the first step and a landslide's mid-point milestone in the second phase, we interpolated latitude and longitude position for the landslide's mid-point. This technique of interpolation has been followed in the literature by IMD [6].

Next, we used IMD rainfall dataset between 2011 and 2015 and mapped the precipitation occurring at each landslide's midpoint's latitude and longitude. In the IMD dataset, the latitude and longitude positions vary in increments of 0.25° from 6.5° to 38.5°and66.5° to 100°, respectively. The last 30-days rainfall at the latitude and longitude point that was closest to the landslide's mid-point was tagged to the landslide's mid-point.

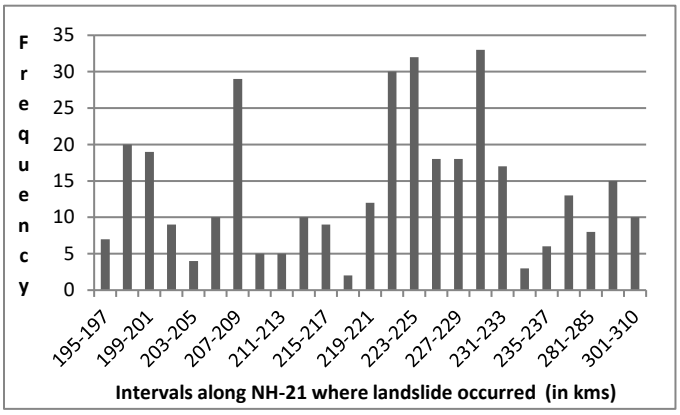

Figure 1: Landslides frequency along the National Highway 21 fromHPPWD landslide occurrence dataset.

Finally, susceptibility data was retrieved from ISRO's Bhuvan website [3] for each landslide's mid-point using the zonation map along NH-21. Figure 2 shows the zonation map along NH-21 between Mandi and Kullu towns. The color on the map indicates the susceptibility of the location to landslides. The red, orange and pink colors correspond to severe, very high and high landslide susceptibility, respectively; whereas, the yellow, green and sky-blue colors correspond to moderate, low, and very low landslide susceptible. The navy-blue color in the map corresponds to the Beas River flowing along NH-21.

For machine-learning analyses, we need at least two classes: landslide (positive) and no-landslide (negative).In the dataset, on a specific day and at certain latitude and longitude point, a landslide could occur only once. These landslide occurrences were marked as the landslide class. Furthermore, if at a latitude-

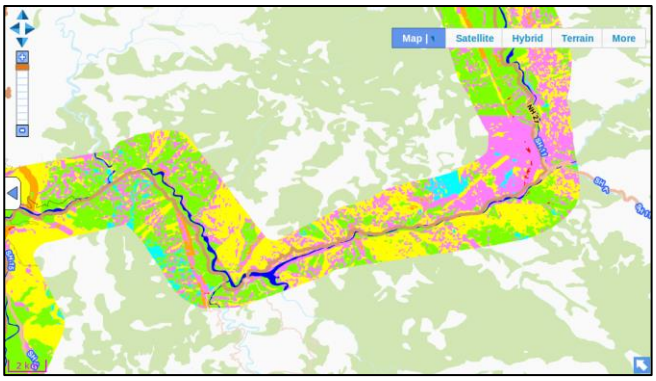

Figure 2: Zonation map along NH-21 between Mandi andKullu towns (Source: ISRO's Bhuvan website).

-Longitude point landslide occurred n times (or over n days) in a month, and then this implied that no landslide took place on any other days in the same month at the same point. So, we generated the no-landslide class by considering all days in the month for a latitude – longitude point where a landslide did not take place (this latitude – longitude point did experience landslides for n days in the same month). Our dataset consisted of 381 landslides (positive) class instances and 9885 instances corresponded to no-landslides (negative) class. As the dataset contained only 3.7% positive class instances, it is severely imbalanced. In the next section, we describe the different class-imbalance techniques that help us correct the imbalance in the dataset.

B. Oversampling Techniques

In this paper, we have categorized oversampling techniques based on their nature of generating new instances for the minority class. The two categories discussed in this study are Random Oversampling and Synthesized Oversampling. The primary difference between these techniques is how new instances are generated in each of this method.

Random Oversampling: In this method, instances of the minority class are selected randomly and duplicated [24] The generated instances do not contribute any new information in the new data set but only increase the magnitude of the number of minority class by just replicating the same information. In our dataset, we have amplified the positive landslide class instances from 381 to 9885 (= to the negative class instances).

Synthesized Oversampling: This technique differs from Random Oversampling, as it contributes new information to the dataset by generating new examples in the proximity of the existing minority class examples. We have discussed two algorithms in this paper namely, SMOTE and SMOTE-IPF, where the former is a widespread technique for resolving the class-imbalance problems; while, the latter is the latest variant of the former technique that performs better than other variants [15].

1. SMOTE [23]: Synthetic Minority Oversampling TEchnique (SMOTE) synthesizes new examples which are not a mere duplication of the existing instances, but, they are generated near of these current instances. This generation is done using the k nearest neighbor algorithm. For every example x in minority class, k-nearest neighbors are found. Among these neighbors, one of the instances is randomly chosen, and a new instance is created along the line from x to the randomly selected neighbor. The ratio required for dividing the two lines is randomly generated. The number of samples duplicated for each x is decided from an input parameter called amplification. If amplification is 100%, one example is created for every x. Similarly, the k is also an input parameter which can be specified by the user. For our experimental framework, we have chosen the amplification as 2594% such that the preprocessed data set has an equal number of positive and negative instances. Apart from amplification, we have generated multiple class-imbalance corrected data sets by varying the k in k-nearest neighbors; from 1 to sqrt(n), i.e., 101 in steps of two. It is important to note that k has been incremented by 2 units to avoid the even values of k because even values are generally vulnerable to a tie and an additional tie-breaker is needed. This range of variation of k has been suggested in the literature [28].

2. SMOTE-IPF [15]: SMOTE – Iterative Partitioning Filter (IPF) is an extension to SMOTE, as it not only generates new examples; but, it removes noisy examples that were created by SMOTE. First, SMOTE is used to oversample the minority class instances by making synthetic examples. Second, Iterative Partitioning Filter [33] cleans the oversampled data set by removing the instances which are noisy. The algorithm has been explained below:
   I. The processed data set E after applying SMOTE has divided into n equal subsets.
   II. At every iteration, a decision tree is trained for each subgroup, creating n-decision trees (C4.5 [8] algorithm (discussed below) in total.
   III. Each decision tree predicts the value of an instance, i.e., either as a landslide or no-landslide.
   IV. An example is labeled noisy using two of the voting schemes: majority or consensus. In the majority scheme, if more than 50% of the classifiers classify an example as no-landslide while it was a landslide or vice-versa, then that case is considered noisy. For the consensus scheme, all classifiers must classify an

instance as the opposite of what it is. These noisy examples are added to data set S.

V. The iteration stops if for consecutive t iterations the number of instances in S is less than p%.

VI. This set S is then removed from the data set E.

For our experimental framework, we have fixed the parameters n, t and p as 9, 3 and 1% respectively as suggested by [15]. We have varied the k parameter in the k-nearest neighbors' algorithm in a similar way as mentioned for SMOTE above.

C. Analysis Methodology

The most straightforward way to evaluate the performance of a classifier is its accuracy or error rate. Accuracy is defined as the ratio of instances (both positive and negative) correctly classified by the total number of instances present in the dataset. However, accuracy can give misleading conclusions for class-imbalanced data [11]. For example, if a binary classifier is trained on a dataset where 99% of instances are of one class, then the classifier can simply get 99% accuracy by classifying every instance as of that class. Thus, it is important to resort to other performance measures that evaluate the classifier's performance class-wise. For our framework, we have chosen true-positive rate, false-positive rate,area under the receiver operating characteristiccurve (AUC), and sensitivity index (d') as the evaluation metrics for the performance of different machine-learning algorithms. True-positive (TP) rate, or precision,is the percentage of landslide instances correctly classified by the classifier as landslides. False-positive (FP) rate is the percentage of no-landslide instances that are classified as landslides. In mostcases, there is a trade-off between these two rates [11]. The choice of TP rate and FP rate in this study is because these measures together provide a complete picture of a classifier's accuracy. Furthermore, the AUC represents the performance of an algorithm as a scalar value. The higher the AUC, the better the model in prediction. Prior research [15, 23, 24]has used AUC as the primary evaluation metric for performance of an algorithm. In addition to the AUC, we have also used the sensitivity index (d') as an additional performance metric in this paper. Sensitivity Index is the separation between the means of the correctly classified instances and the incorrectly classified instances [25]. It is calculated as per the following equation:

$$d' = Z(TP\ rate) - Z(FP\ rate) \qquad (1)$$

Where function $Z(p), p \in [0,1]$, is the inverse of the cumulative distribution function of the Gaussian distribution. The higher the sensitivity index (ord') of a classifier, the better the classifiercompared to other classifiers.

D. Data Classification

Prior research [13] compared several machine-learning algorithms based on prediction performance, interpretability, and high-dimensional prediction. Similarly, we have compared machine-learning algorithms for prediction of landslides. In this paper, we have compared logistic regression [1], decision trees [16], random forests [32], support vector machines (SVMs) [30], and neural networks (or Multilayer Perceptrons) [4] using a 10-fold cross-validation procedure [28].

Logistic regression has been particularly used in modeling landslides as it provides a probability of landslide occurrence against every data point using the logit model [1]. It has been widely used in landslide susceptibility mapping [18]. A decision tree is a hierarchical model composed of decision rules that recursively split independent variables into zones such that each time maximum balance in each split is achieved [16]. The advantage of decision trees is that they can handle categorical as well as numeric variables and can incorporate them without strict assumptions on data [7]. In this study, we have used the J48 algorithm which is a Java implementation of the C4.5 algorithm [8]. The C4.5 uses an entropy-based measure as the attribute selection criteria on the tree nodes, and it is the same as the ID3 algorithm [16]. Given a training dataset T with subsets T_i, i = 1,2,...,s, the C4.5 algorithm constructs a decision tree using the top-down and recursive-splitting technique starting with attributes with the maximum gain [16].

Random forest (RF) is an ensemble technique that utilizes many classification trees (a 'forest') to stabilize the model predictions [32]. The RF algorithm exploits random binary trees which use a subset of the attributes through bootstrapping techniques: From the original data set a random selection of the attributes is performed and used to build the model, the data not included is referred to as "out-of-bag" (OOB) [32]. Each tree is developed to minimize classification errors; but, the random selection influences the results, making a single-tree classification very unstable. For this reason, the RF method makes use of an ensemble of trees (the so-called "forest") thereby ensuring model stability [9]. The RF algorithm has been used in landslide predictions domain and susceptibility modeling by several studies [9, 13, and 20].

Support vector machine is a supervised learning method based on statistical learning theory and the structural risk minimization principle [30]. Using the training data, SVM implicitly maps the original input space into a high-dimensional feature space. Subsequently, in the feature space, the optimal hyper plane is determined by maximizing the margins of class boundaries. We chose the Polynomial Kernel function in this paper since it has outperformed other kernels in prior research [7].

The Multilayer Perceptron (MLP) is an artificial neural network that has been employed widely in many fields including landslide susceptibility assessment [4]. We chose the back propagation algorithm for our framework as it is a popular algorithm for training MLPs.

While each of these machine-learning algorithms could be used with a variety of settings and procedures for model selection, we chose configurations that we have considered typical based upon prior applications. All the techniques mentioned above were run in the Java-written Weka package with default parameter settings [8]. In the next section, the classifiers' accuracy has been discussed. Subsequently, the AUC results and sensitivity index for varying k in case of SMOTE and SMOTE-IPF have been shown. Finally, the performance of both preprocessed and original datasets have been mentioned in training with the Random forests classifier.

IV. RESULTS

Table 1 shows the performance of the different machine-learning classifiers on the landslide dataset. As the dataset is class-imbalanced, each classifier's accuracy is very high. Thus, we need to investigate each classifier's performance on other performance measures beyond accuracy. Among other measures, the TP rate was the highest for Random Forest algorithm followed by the C4.5 Decision Tree algorithm. This suggests that Random Forest did a good job in predicting the landslide instances correctly. In contrast, Logistic Regression and MLP had the lowest TP rate of 0.02 and 0.01, respectively, indicating that they could predict the landslides class instances only 1% of the time. In addition to the TP rate, a low FP rate is preferred because the FP rate corresponds to the ratio of instances that belong to the no-landslide class; however, these instances were incorrectly classified as belonging to the landslide class by the algorithm. Both a low value of TP rate and FP rate is indicative of the fact that the classifier is biased. For example, the FP rate and TP rate in SVM were 0.001 and 0.001 respectively. This result suggested that all the instances were classified as belonging to the majority no-landslide class by the SVM.As seen in Table 1, all classifiers showed good results on the FP rate. The AUC and Sensitivity Index allowed us to observe the combined effect of the TP and FP rates. Based on the AUC and sensitivity index, the Support Vector Machine (SVM) algorithm performed poorly in predicting the landslide class Furthermore, the AUC and sensitivity index was the highest in case of the Random Forest algorithm. These results suggest that the Random Forest algorithm performed better than other classifiers on the landslide dataset. Thus, we took the Random Forest classifier as the base classifier for the class-imbalance techniques.

As mentioned above, the SMOTE and SMOTE-IPF techniques used the k-nearest neighbor algorithm to reduce the class imbalance. For each k value in SMOTE and SMOTE-IPF,

TABLE I.
TEN-FOLD CROSS VALIDATION COMPARISONS OF DIFFERENT CLASSIFIERS ON THE LANDSLIDE DATASET BEFORE PREPROCESSING

Classifier	Accuracy	TP rate	FP rate	AUC	Sensitivity Index
Logistic Regression	96.25%	0.02	0.001	0.79	1.67
C4.5	97.17%	0.44	0.01	0.77	2.18
Random Forests	97.28%	0.58	0.01	0.90	2.53
SVM	96.28%	0.001	0.001	0.50	0.00
MLP	96.85%	0.01	0.01	0.79	1.99

-a new dataset was obtained, and we evaluated the best k value, one that would maximize the sensitivity index. For determining the sensitivity index, the Random Forest algorithm, the best among the different machine-learning algorithms, was applied to the resulting SMOTE or SMOTE-IPF processed data sets in a10-fold cross-validation for several values of k. As seen in Figure 3, for all values of k, the SMOTE-IPF technique outperformed the SMOTE technique for the sensitivity index. Furthermore, the sensitivity index was maximized for SMOTE-IPF and SMOTE techniques at k=1. This k-value implied that the synthesis being done in SMOTE and SMOTE-IPF performed the best when considering only 1-neighbor in the close vicinity of existing landslide class points.

Next, we used the random oversampling, SMOTE, and SMOTE-IPF techniques with Random Forest algorithm to generate 10-fold cross validation predictions on the landslide dataset. First, all the three over-sampling methods improved the performance of the minority class. Comparing with Random Forest algorithm without the use of over-sampling methods, the best improvements of TP rate was in both SMOTE and SMOTE-IPF techniques. Both SMOTE and SMOTE-IPF increased the TP rate from 58% to 99% on the same dataset. Random oversampling was also comparable to the SMOTE and SMOTE-IPF results, where the improvement in TP rate was from 58% to 98%. Secondly, across all class-imbalance techniques, there was very little change in the FP rate. This result suggested that none of these algorithms deteriorated the FP rate. Also, the AUC obtained was 1.0 in case of SMOTE-IPF with Random Forest classifier. This result indicated that almost all instances, both synthetic and original, had been correctly classified by the Random Forest classifier after pre-processing data using SMOTE-IPF.

Furthermore, SMOTE and random oversampling techniques performed equally well regarding the AUC. Finally, the sensitivity index also preferred SMOTE-IPF to other class-

imbalance techniques. We also performed a Wilcoxon signed-rank test to check if the results obtained in SMOTE-IPF were significantly better compared to those obtained in SMOTE.A

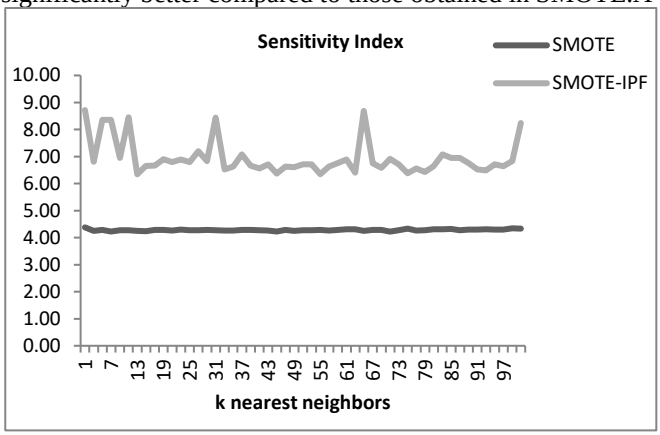

Figure 3: Sensitivity Index when k nearest neighbors are varied in SMOTE and SMOTE-IPF

Wilcoxonsigned-rank test indicated thatthe sensitivity index(d') of SMOTE-IPF was significantly higher than that of SMOTE (Z = -6.22, p< .001). Furthermore, a Wilcoxon signed-rank test indicated that the AUC of SMOTE-IPF was significantly higher than that of SMOTE (Z = -6.81, p< .001).

TABLE II.

TEN-FOLD CROSS VALIDATION COMPARISONS OF DIFFERENT CLASS-IMBALANCE TECHNIQUES WITH RANDOM FORESTS AS THE BASELINE CLASSIFIER

Resampling Technique	Accuracy	TP rate	FP rate	AUC	d'
Without Resampling	97.28%	0.58	0.01	0.90	2.53
Random Oversampling	99.11%	0.98	0.001	0.99	7.30
SMOTE	98.2%	0.99	0.02	0.99	4.38
SMOTE-IPF	99.99%	0.99	0.001	1.00	8.72

V. DISCUSSION AND CONCLUSIONS

The basic aim of this research has been to improve the prediction power of different machine-learning algorithms by removing class-imbalance in a real-world dataset concerning the landslide problem. We tried different class-imbalance techniques consisting of random oversampling, SMOTE, and SMOTE-IPF. All three class-imbalance techniques improved the classification results, which is consistent with previous findings in the literature [15, 23, 24].However, the most recent class-imbalance technique, SMOTE-IPF [15], outperformed other techniques including its predecessor, the original SMOTE [23]. Thus, our expectation that using SMOTE-IPF helps to correct the class-imbalance problem was met, and the accuracy was further improved. In contrast, SMOTE does not perform better than random oversampling. That is because the newly synthesized instances by SMOTE likely add more noisy and borderline instances, deteriorating the FP rate. However, random oversampling simply duplicated the existing instances, preserving the FP rate. Additionally, it is important to note that class-imbalance techniques, which use k-nearest neighbors like SMOTE and SMOTE-IPF, could generate very good prediction results even for small values of k. The smaller k value ensured that it was less likely that the instances belonging to the no-landslide (negative) class don't contribute to the instance generation process.

Furthermore, the Random Forest algorithm performed better than others. This performance of Random forests algorithm could be attributed to its ensemble approach. Ensemble models create multiple classifiers on different subsets of the original dataset. The aggregate opinion of multiple classifiers is likely to be less noisy than one single classifier leading to better and more stable predictions. This result also suggests that ensemble techniques could be suitable for classifying phenomena like landslides, where non-ensemble approaches do not do so well. However, other machine-learning techniques may perform differently on newer datasets. This study was meant to serve as a preliminary study for applying class-imbalance techniques to real-world landslide datasets. As part of future research, we plan to perform an exhaustive study by training other machine-learning classifiers like MLP, SVM, Logistic Regression, and C4.5 on the generated dataset by SMOTE-IPF and other techniques.

While this paper introduced the class-imbalance problem in a landslides prediction task and tried to correct the problem using oversampling techniques, one could also try other techniques as proposed in the literature. One of these techniques is called cost-sensitive learning [34], where a non-zero cost is associated with false positives and false negatives to reduce the number of misclassified instances [31]. Furthermore, newer techniques like SPIDER [17] have been proposed in the literature for mitigating the class-imbalance problem which not only amplifies the minority class instances; but, relabels the existing noisy majority class instances to the minority class. In future, we plan to incorporate some of these techniques with the techniques discussed in this paper and analyze their performance for predicting landslides.

VI. REFERENCES

[1] A. Brenning, "Spatial prediction models for landslide hazards: review, comparison and evaluation", Natural Hazards and Earth System Science,9 vol. 5, no. 6, pp. 853-862, 2005.

[2] A. Stumpf and N. Kerle, "Combining Random Forests and object-oriented analysis for landslide mapping from very high resolution imagery", Procedia Environmental Sciences, vol. 3, pp. 123-129, 2011.

[3] Bhuvan | ISRO's Geoportal | Gateway to Indian Earth Observation | Disaster Services", Bhuvan-noeda.nrsc.gov.in, 2017. [Online]. Available:http://bhuvan-

noeda.nrsc.gov.in/disaster/disaster/disaster.php?id=landslide#. [Accessed: 13- Sep- 2017].

[4] B. T. Pham, D. Tien Bui, H. Pourghasemi, P. Indra and M. Dholakia, "Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods", Theoretical and Applied Climatology, vol. 128, no. 1-2, pp. 255-273, 2015.

[5] B. Üstün, W. Melssen and L. Buydens, "Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel", Chemometrics and Intelligent Laboratory Systems, vol. 81, no. 1, pp. 29-40, 2006.

[6] D. S. Pai, L. Sridhar, M. Badwaik and M. Rajeevan, "Analysis of the daily rainfall events over India using a new long period (1901–2010) high resolution (0.25° × 0.25°) gridded rainfall data set", Climate Dynamics, vol. 45, no. 3-4, pp. 755-776, 2014.

[7] Tien Bui, B. Pradhan, O. Lofman and I. Revhaug, "Landslide Susceptibility Assessment in Vietnam Using Support Vector Machines, Decision Tree, and Naïve Bayes Models", Mathematical Problems in Engineering, vol. 2012, pp. 1-26, 2012.

[8] E. Frank, M. Hall and I. Witten, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Morgan Kaufmann, 2016.

[9] F. Catani, D. Lagomarsino, S. Segoni and V. Tofani, "Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues", Natural Hazards and Earth System Science, vol. 13, no. 11, pp. 2815-2831, 2013.

[10] F. Guzzetti, A. Carrara, M. Cardinali and P. Reichenbach, "Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy", Geomorphology, vol. 31, no. 1-4, pp. 181-216, 1999.

[11] G. E. Batista, R. Prati and M. Monard, "A study of the behavior of several methods for balancing machine learning training data", ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, p. 20, 2004.

[12] IMD - Surface Meteorological Instrumentation", Imd.gov.in, 2017. [Online]. Available: http://www.imd.gov.in/pages/services_sid.php. [Accessed: 13- Sep- 2017].

[13] J. N. Goetz, A. Brenning, H. Petschko and P. Leopold, "Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling", Computers & Geosciences, vol. 81, pp. 1-11, 2015.

[14] J. Mathew, D. Babu, S. Kundu, K. Kumar and C. Pant, "Integrating intensity–duration-based rainfall threshold and antecedent rainfall-based probability estimate towards generating early warning for rainfall-induced landslides in parts of the Garhwal Himalaya, India", Landslides, vol. 11, no. 4, pp. 575-588, 2013.

[15] J. Sáez, J. Luengo, J. Stefanowski and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering", Information Sciences, vol. 291, pp. 184-203, 2015.

[16] J. Quinlan, "Induction of decision trees", Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[17] K. Napierała, J. Stefanowski and S. Wilk, "Learning from Imbalanced Data in Presence of Noisy and Borderline Examples", Rough Sets and Current Trends in Computing, pp. 158-167, 2010.

[18] L. Wang, K. Sawada and S. Moriguchi, "Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy", Computers & Geosciences, vol. 57, pp. 81-92, 2013.

[19] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection", In Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179-186, 1997.

[20] M. Ließ, B. Glaser and B. Huwe, "Functional soil-landscape modelling to estimate slope stability in a steep Andean mountain forest region", Geomorphology, vol. 132, no. 3-4, pp. 287-299, 2011.

[21] M. Marjanović, M. Kovačević, B. Bajat and V. Voženílek, "Landslide susceptibility assessment using SVM machine learning algorithm", Engineering Geology, vol. 123, no. 3, pp. 225-234, 2011.

[22] M. Naseriparsa and M. Mansour RiahiKashani, "Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset", International Journal of Computer Applications, vol. 77, no. 3, pp. 33-38, 2013.

[23] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", J. Artif. Int. Res., vol. 16, no. 1, pp. 321-357, 2002.

[24] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study", Intelligent Data Analysis, vol. 6, no. 5, pp. 429-449, 2002.

[25] N. Macmillan and C. Creelman, Detection theory. New York, NJ [u.a.]: Psychology Press, 2010.

[26] P. Atkinson and R. Massari, "GENERALISED LINEAR MODELLING OF SUSCEPTIBILITY TO LANDSLIDING IN THE CENTRAL APENNINES, ITALY", Computers & Geosciences, vol. 24, no. 4, pp. 373-385, 1998.

[27] R. Dubey, J. Zhou, Y. Wang, P. Thompson and J. Ye, "Analysis of sampling techniques for imbalanced data: An n=648 ADNI study", NeuroImage, vol. 87, pp. 220-241, 2014.

[28] R. Duda, Pattern Classification 2nd Edition with Computer Manual 2nd Edition Set. John Wiley & Sons, 2004.

[29] S. Tripathi, V. Srinivas and R. Nanjundiah, "Downscaling of precipitation for climate change scenarios: A support vector machine approach", Journal of Hydrology, vol. 330, no. 3-4, pp. 621-640, 2006.

[30] V. Vapnik, Statistical learning theory. New York: J. Wiley, 1998.

[31] P. Sarakit, T. Theeramunkong and C. Haruechaiyasak, "Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm", Advanced Informatics: Concepts, Theory and Applications (ICAICTA), 2015 2nd International Conference on, pp. 1-5, 2015.

[32] L. Breiman, "Random forests", Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[33] T. Khoshgoftaar and P. Rebours, "Improving software quality prediction by noise filtering techniques", Journal of Computer Science and Technology, vol. 22, no. 3, pp. 387-396, 2007.

[34] J. Zhang, H. Lu, W. Chen and Y. Lu, "A Comparison Study of Cost-Sensitive Learning and Sampling Methods on Imbalanced Data Sets", Advanced Materials Research, vol. 271-273, pp. 1291-1296, 2011.