Dear Anonymous Referee #1,

Thank you for your positive and constructive comments. Below is a documented list of changes we have made to the manuscript (marked R: in blue font). We have shortened the introduction of the ML algorithms, detailed the majority and minority class sampling used in the training and testing, revised the results by 5-fold cross-validation, and refined the discussion and conclusion of the findings. We hope these clarifications will improve the reader's understanding of our work.

Kind Regards,
Praveen Kumar

Anonymous Referee #1 Comments

1. The different ML algorithms are individually reported and described in perhaps too much detail.

   R: We appreciate the reviewer's feedback on our ML algorithm descriptions. In response, we have revised and shortened the introductions of each algorithm. We now provide a brief overview of each algorithm in three to four sentences, followed by an explanation focused solely on the crucial parameter variations and their significance in our experimental setup.

2. The monitored landslide is presented only in geographical terms. It might be useful to provide more details on the characteristics of the landslide.

   R: Thank you for this suggestion. We have revised the main manuscript and added the paragraph in the Data Collection and Description section on page 4 as follows, along with a new reference:

   "The monitored landslides are characterized as shallow landslides with debris flow, occurring at elevations ranging from 1450 m to 1920 m. The slopes in the landslide zones in the upper parts are made up of weathered limestone and dolomitic limestone, whereas the lower slopes exhibit black carbonaceous slate. The slates are highly weathered and leached, adorned with white and yellow encrustation. These are covered with a thin veneer of debris, mainly consisting of pebble- and cobble-sized limestone, sandstone, and slate embedded in a sand–silt–clay matrix. Additional context includes an annual rainfall of 4190 mm in the area, as reported by Gupta et al. (2015)."

   We added a new reference.

   Gupta, V., Bhasin, R. K., Kaynia, A. M., Tandon, R. S., & Venkateshwarlu, B. (2016). Landslide hazard in the Nainital township, Kumaun Himalaya, India: the case of September 2014 Balia Nala landslide. *Natural Hazards*, *80*, 863-877.

3. However, it is not clear if the other two minority classes were oversampled, or if they were removed in subsequent analyses.

R: Thank you for your comment. We appreciate your feedback and have addressed this concern in the revised manuscript.

In the Class Labeling section on page 5, we now provide detailed information on the distribution of classes, explicitly stating the percentages for each category:

"The majority of the dataset (97.8%) falls under the 'No Movement' category, indicating a lack of significant movement. On the other hand, the 'High Movement' category represents only a small fraction (1.1%) of the dataset. Additionally, the 'Moderate Movement' category comprises 0.7% of the samples, while the 'Low Movement' category accounts for 0.4% of the dataset."

In the Oversampling section on page 6, we clarify the representation of all classes:

"All other classes, including "High Movement," "Moderate Movement," and "Low Movement," represent minority classes, each constituting only 1%, 0.7%, and 0.4% of the total data, respectively."

Furthermore, on page 6, in the Oversampling section, we now explicitly state:

"By utilizing the characteristics of existing samples from the minority classes, we created new data points, thereby increasing the representation of the 'High Movement,' 'Moderate Movement,' and 'Low Movement' classes."

4. The main results are synthesised in Tables 5 and 6. In my opinion, these two tables are not enough to convey the effect of oversampling. In most cases the data without oversampling returns better scores than the oversampled data in all the metrics, not allowing the reader to understand the cause. Furthermore, scores so close to 1 might suggest a data leakage between training and model testing. It could be worth it to revise the data-splitting procedure and implement the pipeline with cross-validation to avoid this issue.

R: We appreciate your valuable feedback and have taken your comments into careful consideration. To address your concerns regarding the impact of oversampling, we have revised our methodology by incorporating a 5-fold cross-validation approach. This enhancement ensures a more robust evaluation of model performance, minimizing the risk of data leakage between the training and testing phases.

Upon implementing this cross-validation technique, we re-evaluated the results and observed a consistent improvement in the performance of models utilizing K-Means SMOTE for oversampling. The revised Tables 5 and 6 now accurately reflect the effectiveness of oversampling techniques, particularly highlighting the superiority of K-Means SMOTE in enhancing predictive accuracy.

We have updated the manuscript to include this important modification in the "Model Execution, Minimization, and Handling Class Imbalance" section on page 9, providing a clear description of the revised methodology.

“A rigorous process was followed to develop an effective model for predicting the intensity of soil movement. The dataset was partitioned into a 70:30 ratio, with 70% allocated for training and the remaining 30% for testing. To tackle the class imbalance issue in the training data, oversampling techniques were applied exclusively to the training set, ensuring a balanced representation of all three classes. The oversampling methods were not extended to the testing data, preserving its original distribution. Following the balancing process, a suite of ML models underwent training using a 5-fold cross-validation (5-CV) approach to the training data (Kumar et al., 2023). The models were optimized by employing grid search methodology, systematically exploring various parameter combinations that maximized the average cross-validation accuracy during training. The training performance, assessed through 5-CV, reflected the models' effectiveness with the optimized parameters. Subsequently, the models with the best parameters found during training were tested on the independent testing data, and their performance metrics were reported as indicative of their predictive capabilities. The evaluation primarily focused on accuracy metrics to determine how effectively the models predicted the intensity of soil movement.”

We have also added a new reference.

Kumar, P., Priyanka, P., Dhanya, J., Uday, K. V., & Dutt, V.: Analyzing the Performance of Univariate and Multivariate Machine Learning Models in Soil Movement Prediction: A Comparative Study. *IEEE Access*, 11, 62368–62381, 2023

Additionally, the results section on page 11 has been amended to present the latest findings obtained through 5-fold cross-validation.

“Table 5 presents the training results of different classification models combined with various oversampling techniques for landslide prediction. These results provide valuable insights into the performance of each model when trained on the training dataset with and without oversampling. The dynamic ensemble model with K-Mean SMOTE emerges as the best model in training, achieving outstanding accuracy, precision, recall, and F1 scores of 0.996, 0.996, 0.996, and 0.996, respectively. The dynamic ensemble model with SMOTE, Borderline SMOTE, and ADASYN techniques also showed similar performance with 0.995 F1 scores. It demonstrates remarkable predictive capability by achieving perfect accuracy in oversampling scenarios. When the model is trained without oversampling, its accuracy, precision, recall, and F1 score are notably lower, with values of 0.981, 0.557, 0.386, and 0.436, respectively.

Table 6 presents the test results of various classification models combined with different oversampling techniques for landslide prediction. Among them, the dynamic ensemble model utilizing the K-Mean SMOTE technique demonstrates exceptional performance in accurately predicting landslides on unseen data. It achieves impressive accuracy, precision, and recall rates of

0.994, 0.882, and 0.945, respectively, along with an F1 score of 0.911. These outstanding results confirm the effectiveness of the dynamic ensemble approach when combined with K-Mean SMOTE for accurate soil movement prediction. Notably, it is crucial to highlight the impact of oversampling on the performance of the dynamic ensemble model. When the model is tested without oversampling, its accuracy, precision, recall, and F1 score are notably lower, with values of 0.981, 0.557, 0.386, and 0.436, respectively. The best-performing model is highlighted in bold in Table 6.

Additionally, the dynamic ensemble model incorporating SMOTE emerges as the second-best model in the test phase, showcasing high accuracy, precision, and recall rates of 0.993, 0.872, and 0.950, respectively, along with an F1 score of 0.907. Moreover, it is noteworthy that K-Means SMOTE consistently outperformed other oversampling techniques across all models during the test performance evaluations, establishing itself as the optimal technique. In addition, the SMOTE technique consistently secured the second-best position across all models. This underscores the discernible effectiveness of K-Means SMOTE in generating oversampling for the soil movement dataset. The success of K-Means SMOTE can be attributed to its ability to identify clusters within the minority class and select similar features for oversampling. The IR employed by K-Means SMOTE aids in determining the appropriate degree of oversampling for the minority class, ensuring a balanced representation of classes in synthetic samples.

Moreover, the absence of oversampling techniques negatively impacted the models' performance in both training and testing. Without oversampling, the models exhibited lower accuracy, precision, recall, and F1 scores during training and testing, emphasizing the challenges posed by class imbalance. In the absence of balanced representation through oversampling, the models struggled to effectively learn and generalize from the imbalanced dataset. Consequently, this underscores the pivotal role of oversampling in mitigating class imbalance issues, leading to substantial enhancements in predictive accuracy and overall model robustness during both training and testing evaluations."

Table 5. The results of the ML models from the training dataset.

Model	Oversampling Technique	Accuracy	Precision	Recall	F1 Score
AdaBoost	SMOTE	0.632	0.638	0.632	0.632
	K-Means SMOTE	0.641	0.646	0.641	0.631
	Borderline SMOTE	0.663	0.670	0.663	0.659
	ADASYN	0.618	0.622	0.618	0.618
	Without Oversampling	0.980	0.556	0.357	0.393
XGBoost	SMOTE	0.921	0.921	0.921	0.921
	K-Means SMOTE	0.926	0.926	0.926	0.926
	Borderline SMOTE	0.973	0.973	0.973	0.973
	ADASYN	0.915	0.916	0.915	0.915
	Without Oversampling	0.994	0.983	0.814	0.882
Light GBM	SMOTE	0.920	0.920	0.920	0.920
	K-Means SMOTE	0.939	0.940	0.939	0.939
	Borderline SMOTE	0.963	0.963	0.963	0.963
	ADASYN	0.915	0.916	0.915	0.915
	Without Oversampling	0.991	0.845	0.791	0.807
CatBoost	SMOTE	0.860	0.860	0.860	0.859

	K-Means SMOTE	0.876	0.876	0.876	0.876
	Borderline SMOTE	0.932	0.932	0.932	0.932
	ADASYN	0.859	0.859	0.859	0.859
	Without Oversampling	0.983	0.797	0.399	0.469
RF	SMOTE	0.731	0.742	0.731	0.728
	K-Means SMOTE	0.734	0.748	0.734	0.729
	Borderline SMOTE	0.795	0.806	0.795	0.797
	ADASYN	0.732	0.747	0.732	0.728
	Without Oversampling	0.982	0.905	0.325	0.372
MLP	SMOTE	0.902	0.903	0.902	0.901
	K-Means SMOTE	0.944	0.945	0.944	0.944
	Borderline SMOTE	0.961	0.962	0.961	0.962
	ADASYN	0.942	0.943	0.942	0.942
	Without Oversampling	0.979	0.635	0.309	0.339
LSTM	SMOTE	0.747	0.750	0.747	0.745
	K-Means SMOTE	0.767	0.769	0.767	0.766
	Borderline SMOTE	0.779	0.781	0.779	0.778
	ADASYN	0.756	0.759	0.756	0.755
	Without Oversampling	0.758	0.760	0.758	0.756
Dynamic Ensemble	SMOTE	0.995	0.995	0.995	0.995
	K-Means SMOTE	**0.996**	**0.996**	**0.996**	**0.996**
	Borderline SMOTE	0.995	0.995	0.995	0.995
	ADASYN	0.995	0.995	0.995	0.995
	Without Oversampling	0.981	0.557	0.386	0.436

Table 6. The results of the ML models from the test dataset.

Model	Oversampling Technique	Accuracy	Precision	Recall	F1 Score
AdaBoost	SMOTE	0.798	0.301	0.646	0.313
	K-Means SMOTE	0.865	0.313	0.610	0.342
	Borderline SMOTE	0.804	0.313	0.598	0.326
	ADASYN	0.788	0.293	0.625	0.300
	Without Oversampling	0.979	0.419	0.313	0.340
XGBoost	SMOTE	0.957	0.509	0.872	0.610
	K-Means SMOTE	0.957	0.486	0.807	0.576
	Borderline SMOTE	0.954	0.480	0.770	0.560
	ADASYN	0.958	0.517	0.876	0.619
	Without Oversampling	0.981	0.618	0.402	0.461
Light GBM	SMOTE	0.951	0.495	0.858	0.591
	K-Means SMOTE	0.952	0.475	0.796	0.561
	Borderline SMOTE	0.954	0.474	0.735	0.548
	ADASYN	0.951	0.488	0.865	0.586
	Without Oversampling	0.978	0.511	0.447	0.467
CatBoost	SMOTE	0.944	0.439	0.831	0.524
	K-Means SMOTE	0.945	0.443	0.793	0.528
	Borderline SMOTE	0.948	0.431	0.747	0.510
	ADASYN	0.948	0.433	0.791	0.517
	Without Oversampling	0.980	0.664	0.389	0.442
RF	SMOTE	0.829	0.342	0.737	0.367
	K-Means SMOTE	0.831	0.336	0.687	0.355
	Borderline SMOTE	0.833	0.321	0.632	0.346
	ADASYN	0.828	0.350	0.689	0.364
	Without Oversampling	0.978	0.477	0.268	0.280
MLP	SMOTE	0.887	0.414	0.945	0.498
	K-Means SMOTE	0.928	0.499	0.959	0.602
	Borderline SMOTE	0.907	0.423	0.742	0.473

	ADASYN	0.909	0.454	0.942	0.547
	Without Oversampling	0.978	0.635	0.308	0.339
LSTM	SMOTE	0.856	0.318	0.684	0.352
	K-Means SMOTE	0.940	0.402	0.736	0.473
	Borderline SMOTE	0.925	0.384	0.720	0.448
	ADASYN	0.887	0.326	0.556	0.361
	Without Oversampling	0.827	0.312	0.710	0.339
Dynamic Ensemble	SMOTE	0.993	0.872	0.950	0.907
	K-Means SMOTE	**0.994**	**0.882**	**0.945**	**0.911**
	Borderline SMOTE	0.993	0.900	0.869	0.880
	ADASYN	0.993	0.854	0.952	0.898
	Without Oversampling	0.982	0.695	0.434	0.506

5. Chapter 7 is just conclusions; the critical investigation of results (i.e., the discussion) is completely missing.

R: We sincerely appreciate your thorough review of Chapter 7. Your insightful comments have guided us in making important revisions to ensure the completeness of the document. We have now addressed this concern by incorporating a comprehensive discussion section on page 13, covering critical investigation, outcomes of the experiment, implications of oversampling techniques, limitations, and key findings.

The Discussion and Conclusion Section is revised as follows:

In summary, the threat posed by landslides requires the development of effective prediction frameworks, although modeling the chaotic nature of natural data remains challenging. The analyzed dataset exhibited a significant class imbalance, with the majority class dominating the samples. This distribution imbalance necessitated careful consideration and appropriate techniques to address the issue.

Various oversampling techniques, including SMOTE and its extensions (K-Means SMOTE, Borderline SMOTE, and ADASYN), were employed to tackle the class imbalance. ADASYN, which focuses on the minority class boundary, effectively generated synthetic data points and improved the class distribution balance.

Multiple classification models, such as ADABoost, XGBoost, Light GBM, CatBoost, RF, MLP, LSTM, and a dynamic ensemble, were evaluated to predict soil movement. The grid search approach and 5-CV were employed to optimize the hyperparameters of each model. The training results highlight the significant impact of oversampling on model performance. The dynamic ensemble model, particularly when coupled with K-Means SMOTE, emerges as the standout performer in the training phase. Achieving remarkable accuracy, precision, recall, and F1 scores of 0.996, 0.996, 0.996, and 0.996, respectively, this model demonstrates superior predictive capabilities.

Furthermore, these models were tested to assess their ability to generalize well to unseen data. The testing results showcased the dynamic ensemble model with K-Means SMOTE as the top performer, achieving an outstanding accuracy of 0.994, precision of 0.882, recall of 0.945, and an

F1 score of 0.911. This confirms that the exceptional performance observed in training extends to the testing phase, emphasizing the robustness and reliability of the dynamic ensemble approach with K-Means SMOTE. Moreover, the dynamic ensemble model incorporating SMOTE emerges as the second-best model in the test phase, showcasing high accuracy, precision, and recall rates of 0.993, 0.872, and 0.950, respectively, along with an F1 score of 0.907. This result reinforces the reliability and robustness of the model in tackling landslide prediction tasks.

Furthermore, the dynamic ensemble model incorporating SMOTE emerges as the second-best model in the test phase, showcasing high accuracy, precision, and recall rates of 0.993, 0.872, and 0.950, respectively, along with an F1 score of 0.907. This result reinforces the reliability and robustness of the model in tackling landslide prediction tasks.

The superior performance of the K-Means SMOTE technique can be attributed to its ability to identify clusters within the minority class and generate synthetic samples that maintain the underlying structure of the data. By considering the IR, K-Means SMOTE ensures a balanced representation of classes in the synthetic samples, contributing to improved model generalization and predictive accuracy. Furthermore, the lack of oversampling adversely affected both training and testing performances. The models faced challenges in learning and generalizing from the imbalanced dataset without a balanced representation.

On the other hand, the success of the dynamic ensemble model, comprising AdaBoost, XGBoost, Light GBM, CatBoost, and Random Forest, can be attributed to the complementary strengths of these diverse algorithms. Ensemble methods leverage the collective decision-making power of multiple models, each capturing different aspects of the underlying data patterns. The combination of boosting algorithms like AdaBoost, gradient boosting methods like XGBoost, tree-based models like Light GBM and CatBoost, and the robustness of RF creates a robust and versatile ensemble that excels in handling various aspects of the dataset, contributing to its overall superior performance.

In summary, the findings underscore the critical role of oversampling techniques, especially K-Means SMOTE, in enhancing the predictive performance of landslide prediction models. The success of the dynamic ensemble model further highlights the importance of ensemble techniques in aggregating diverse model predictions for improved accuracy.

Despite these achievements, it is crucial to acknowledge the study's limitations. The generalizability of the findings to different geological conditions or regions may be restricted due to the specificity of the dataset. The synthetic data points generated through oversampling, while effective, may only capture part of the complexity inherent in real-world landslide occurrences. The choice of classification models and hyperparameter settings introduces a level of bias, with alternative configurations potentially yielding different results. Additionally, relying on historical data may limit the model's ability to account for future changes or unforeseen events, such as changes in rainfall intensity, seismic activity, or human influences.

In future work, the exploration of encoder-decoder models or transformer models on the class-imbalanced movement dataset is planned. These models, known for their success in sequence-to-sequence tasks, may offer improvements in classification accuracy and address class imbalance challenges. This avenue of experimentation aims to provide valuable insights into the suitability of advanced models for analyzing and modeling imbalanced movement data.

To sum up, the study contributes to the understanding of landslide risks and supports the development of effective preventive measures. The combination of robust oversampling techniques, ensemble modeling, and a systematic approach to hyperparameter tuning yields a promising framework for accurate landslide prediction. The work presented lays the groundwork for future research aimed at refining models and addressing the inherent challenges in landslide prediction tasks.