Comparative Analysis of Machine Learning Models and Hybrid Ensemble Approach's for Landslide Prediction

Fayaz Mohsin* and Kanti Tushar
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, INDIA
*mohsinkhanday93@gmail.com

Abstract
The NH-44 Jammu Srinagar National Highway in India is susceptible to landslides, rock falls and shooting stones due to its geological characteristics and steep slopes. This study aims to compare the performance of various Machine Learning (ML) algorithms and hybrid models in predicting landslides using historical data. Seven optimized ML approaches Support Vector Classifier (SVC), Logistic Regression (LR), Decision Tree (DT), K Nearest Classifier (KNC), Random Forest (RF), GaussianNB (GNB) and AdaBoost Classifier (ABC) are used.

Additionally, two Hybrid Ensemble methods, the Voting Hybrid Model (VHM) and Stacking Hybrid Model (SHM), are introduced. The performance of each model is evaluated using accuracy, precision, recall, F1-score and AUC metrics. Results indicate that all the models performed well, with hybrid ensemble models surpassing all individual algorithms. The Stacking Hybrid Model (SHM) excels achieving 99.4% accuracy and 98.5% AUC, outperforming the Voting Hybrid Model (VHM). Hybrid models consistently outperform individual models in accuracy and AUC. These proposed methods exhibit robustness and enhanced results in addressing this issue. This framework can help to predict the landslides with high accuracy which can save the lives through timely evacuations from high-risk areas.

Keywords: Landslide, Disaster, Prediction, Machine Learning.

Introduction

Landslides are natural catastrophes that can inflict substantial property damage and significant loss of life. Landslide-prone areas are habitat to approximately 66 million people across world with soaring numbers in Asia[1]. According to the Centre for Research on Epidemiology of Disasters (CRED), landslides claim for nearly 17% of all natural disaster fatalities worldwide[2]. Landslides have been responsible for considerable losses on a global scale. According to a survey spanning from 1998 to 2017, an estimated 4.8 million people were affected and tragically, 18,000 lives were lost due to these devastating events. The profound impact of landslides on communities and individuals highlights the urgent need for comprehensive measures to enhance prediction, preparedness and response strategies to mitigate their toll on human life and infrastructure.

It is anticipated that such occurrences may rise in future due to climate change and mass human migration towards such perturbed areas and an increase in built-up around these areas[3].

In recent years, South Asian countries particularly the Hindu-Kush-Himalaya foothills are witnessing a rise in landslide occurrences because of its overall dynamic geological conditions[4]. Jammu Srinagar National Highway (NH-44) passes through huge, barren, perturbed and steep slopes of the Great Himalayan range and Pir-Panjal range which are highly prone to landslides, shooting stones and rock falls. The range has a group of huge and steep mountains with the elevation ranging from 1400m to 4100m ASL[5]. The anthropogenic activities, unchecked deforestation, natural and human activities have made the area highly prone and susceptible to landslides[6].

According to Brabb[2], 1991 losses due to landslides can be reduced up to 90% if it is predicted or recognized earlier[7]. An early prediction method can reduce the loss of life caused by the disaster, as people can evacuate before a landslide hit the area. Modern early warning and forecasting models use the IOT to monitor various physical and ecological factors in order to anticipate calamities while others use ML techniques to analyze various responsible internal and external parameters to predict these disasters. ML an application of AI develops the model to represent the relation between data and target variables. It laces the system with automatic learning capability from the historical data without being explicitly programmed. It uses prior knowledge as input to predict future output values.

Dataset used in this study consists of eight responsible variables (rainfall precipitation, snowfall precipitation, land surface temperature, soil moisture, average subsurface runoff, average snowmelt, slope and average near-surface wind speed) with a single response (output) variable. Three-year time-series data from January 2018 to December 2020 for each factor was obtained from different sources.

The responsible variables of landslide events from the same period were obtained from various databases which include the open landslide data portal of NASA and historical landslide reports from GSI. Some events were collected from the local and national media. The data was integrated into a single dataset with one response variable landslide chances (LSC) where 1 signifies yes and 0 represents no.

The present study uses optimized ML approaches: Support Vector Classifier (SVC), Logistic Regression (LR), Decision Tree (DT), K Nearest Classifier (KNC), Random Forest (RF), GaussianNB (GNB) and AdaBoost Classifier (ABC). Additionally, two hybrid ensemble methods are introduced: the Voting Hybrid Model (VHM) and the Stacking Hybrid Model (SHM). These methodologies are applied to predict the occurrences of landslides. The results underscore the commendable overall performance displayed by all models.

However, the hybrid ensemble models outperformed all the individual algorithms. Notably, the stacking hybrid model (SHM) garners attention for its superior performance compared to the voting hybrid model (VHM). The SHM achieves an impressive accuracy rate of 99.4%, accompanied by a notable AUC of 98.5%. Consistently, the hybrid models demonstrate higher accuracy and AUC values in comparison to the individual models. The proposed methods demonstrate exceptional robustness and yield superior overall outcomes in the domain of landslide prediction. Consequently, the hybrid models VHM and SHM models emerges as the most promising contender for advancing the Landslide Early Warning Systems (LEWS).

Study Area

The Degital Elevation Model (DEM) based study area shown in figure 1 is extended over 65 km stretch starting from Jawahar Tunnel Banihal district of Jammu and Kashmir to Chandarkot and covers a total area of 401sqkm[8]. A DEM is an electronic model of the Earth's surface which is processed and manipulated using ArcGis 10.3.

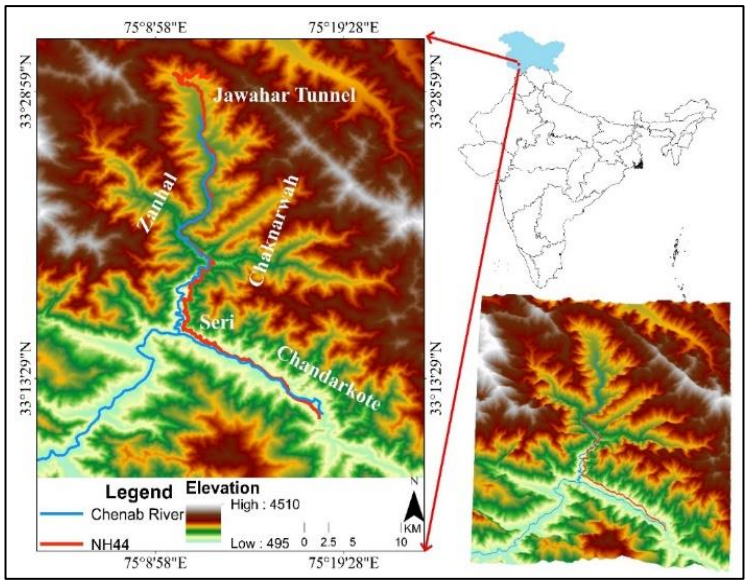

Figure 1: Study Area

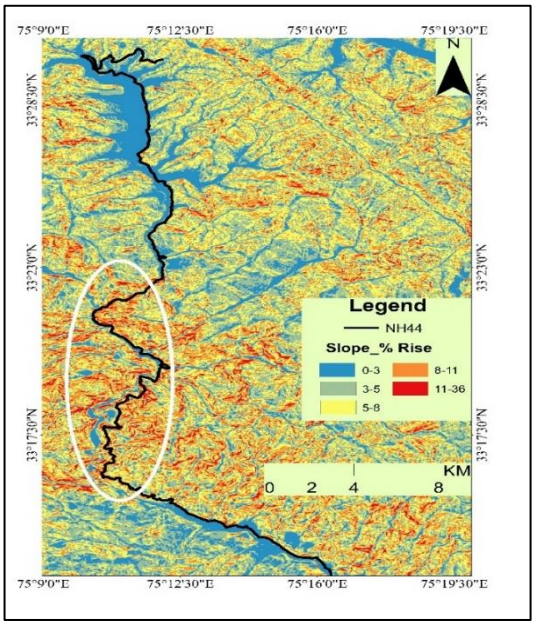

Figure 2: Slope Map of the Study Area

It offers greater features than the nominal and qualitative characterization of topography, it can provide a variety of important types of data such as slope aspect, contour lines, curvature, elevation rise and drainage. The data derived can be used to analyze the area for susceptibility mapping and landslide prediction modelling.

The area has extensive steep hilly topography as shown in figure 2, with an average slope angle greater than 19° and an average altitude of 2741m ASL. Slopes along NH44 have undergone huge deformations due to heavy traffic, road widening and tectonic activities, resulting in deadly landslides and rock-falls at various locations. The slope angle of the study area varies enormously and shows a huge rise of 0-36%.

The encircled area that ranges from 'Nachlana' to 'Seri' shows a rise in slope with an average rise of 28%. The region is highly susceptible to landslides and rock-falls with several active landslides present along the same stretch. The occurrence of both rock falls and landslips is highly correlated to rainfall precipitation and snowfall precipitation which increase the moisture content in the soil and that in turn decreases the sturdiness of soil and causes slope failures. The area receives an average precipitation of 63 mm/month with both extended and intensive rainfall events resulting from the western disturbances and huge elevation. It was revealed from the previous studies that extended and intensive precipitation events have a direct effect in causing slope failures by over saturating the slopes in the area[9].

Material and Methods

The entire approach as depicted in figure 3 employed in this research is explained here. Several data preprocessing approaches along with machine learning algorithms used in this study are discussed. The assessment metrics had been used to evaluate each model's performance. Python version 3.8.5 was employed for constructing the requisite machine learning models due to its reputation for user-friendliness and its potency within the domain of machine learning modeling.

Data Collection and Pre-processing: The dataset used in this study, as detailed in table 1, was sourced from multiple origins. This dataset comprises a total of 1096 instances, with 763 instances labeled as '0,' representing 'No Landslide,' and 333 instances labeled as '1,' indicating 'Landslide,' as visually depicted in figure 4 (a). The dataset encompasses a three-year time-series data spanning from January 2018 to December 2020. The data pertaining to landslide events (LSC) for this same timeframe was acquired from diverse sources including the open landslide data portal of NASA, Geological Survey of India, in addition to relevant events gathered from local and National media outlets. The data was integrated into a single dataset with eight explanatory variables and a single response variable. Several data mining techniques were implemented to preprocess the data which includes three primary stages: data cleaning, data integration and data transformation.

Data cleaning was performed to remove unwanted, redundant and incomplete data from the dataset. The incorrect information and noisy data can lead to poor prediction and decision making. Data integration, an important data preprocessing technique is used to merge the data from several heterogeneous sources into a single coherent dataset. Data transformation is performed to transform the data into the desired format using data smoothing and data aggregation techniques to make it better organized for better results.

Data Cleaning: The KNN imputation method was employed to handle the missing data values, it is an optimal technique to handle the missing values in a dataset.

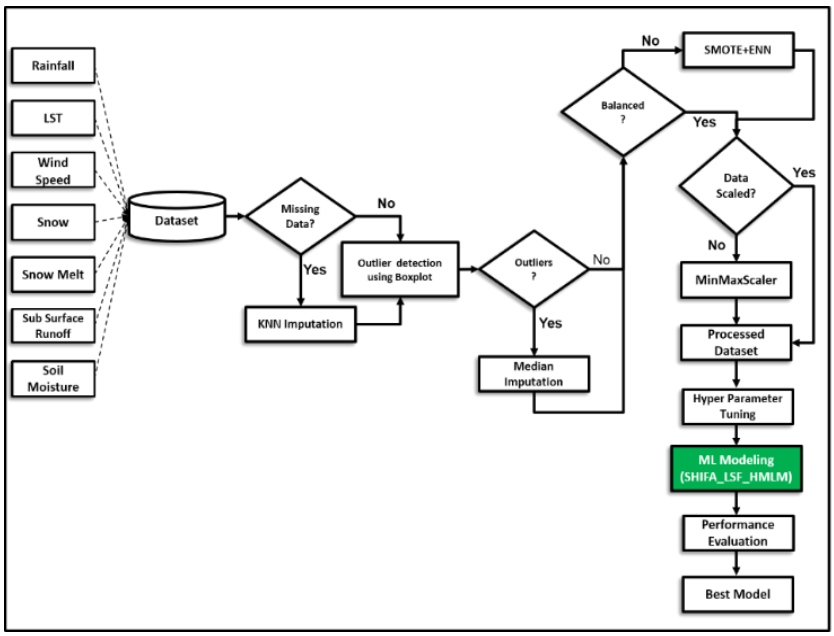

Figure 3: Workflow Diagram

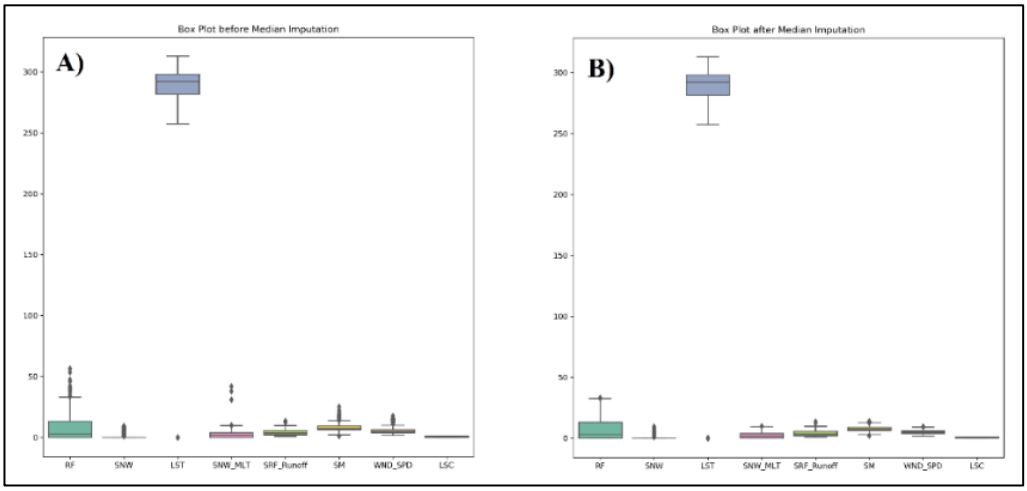

Figure 4: (a) Outlier Detection (b) Outliers removed using Median Imputation Method

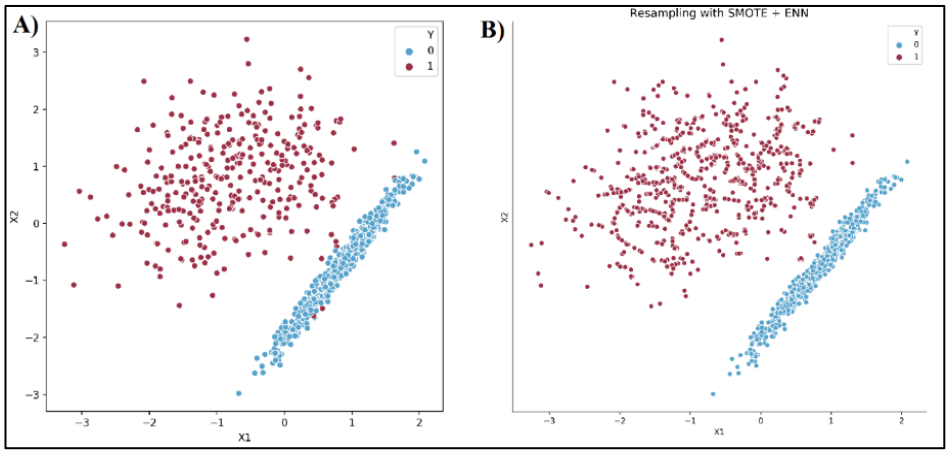

Figure 5: A) Imbalanced Data B) Resampling with SMOTE+ENN

The box plot method is used to identify the outliers present in dataset and subsequently, these outliers were removed and replaced using median imputation method which is proven as better imputation technique than other methods[10]. Figure 4 clearly exhibits how well the technique has performed in dealing with the outliers present in the dataset.

The dataset exhibits a notable class imbalance with a considerably lower number of instances representing landslide events compared to those representing no landslide events. To address this imbalance, a hybrid data resampling method is employed, which combines Synthetic Minority Over-sampling Technique (SMOTE) with Edited Nearest Neighbor (ENN) to achieve a balanced dataset. SMOTE employs the nearest neighbor algorithm to generate new instances of minority categories by convexly combining neighboring instances. The ENN technique serves as an undersampling approach, estimating the nearest neighbors of instances within the majority class.

Incorporating this technique alongside SMOTE enhances the process of thorough data refinement. This combined approach results in the removal of samples from both classes that were inaccurately classified by the nearest neighbors. Consequently, the distinction and clarity between classes become more pronounced and concise, as depicted in figure 5.

SMOTE combined with ENN undersampling technique is proven effective resampling technique[11]. SMOTE helps to balance the class distribution by adding the new data points while the ENN removes the irrelevant data points over the boundary of two classes in order to increase the separation between the two classes. The dataset is made more purposeful and streamlined using these techniques by substituting the sample imbalance rate with the sample misclassification rate in view of the shortcomings of the conventional SMOTE algorithm.

Machine Learning Modelling for Landslide Prediction: ML modeling which is an efficient and effective approach to predict the future events, is used to predict the future landslides by recognizing the important patterns in the historical data. In the present study, the following seven ML approaches Support Vector Classifier (SVC), Logistic Regression (LR), Decision Tree (DT), K Neighbors Classifier (KNC), Random Forest (RF), Naïve Bayes (NB), AdaBoost Classifier (ABC) and two ensemble hybrid methods Stacking Hybrid Model (SHM) and Voting Hybrid Model (VHM) are employed to develop and find the better prediction model for the prediction of landslides occurrence.

Algorithmic Steps for Landslide Prediction

Step 1: Import Dataset.
Step 2: Import Required Libraries.
Step 3: Data Partitioning using Random Splitting Method.
Step 4: ML Algorithms used in the Model.
mn=[SVC(), LR, DT( ), NB(), RF(), ABC(), KNN( ),SHM(),VHM()]
Step 5: for(i=1; i<=9; i++) do
Model= mn[i];
Model.fit();
model.predict();
Step 6: print(Accuracy(i),confusion_matrix, auc, classification_report);
Step 7: End

KNC Model for Landslide Prediction: KNN is a supervised Machine Learning technique that aids in solving both regression and classification problems[12]. The algorithm works by assuming the resemblance between the current data and previously known instances and placing the newly added instance to the category that is closest to the previous instance in its nearest neighbor 'K'. Here K is the number of nearby neighbors chosen from a set of classes. The Euclidean distance is employed in order to determine the proximity between the two points.

DT for Landslide classification and Prediction: The ML approach DT is the biggest contributor to predictive modeling with a wide range of applications in disaster predictions. It is a basic method of classification with a tree-like structure from a root node to leaves (target value). The final output comprises of a discrete set of values with branches representing features and leaves representing class labels. At each phase of the tree building process, the information gain (IG) is used to choose which is characteristic to split in building and extending the tree. The entropy at each split is calculated as:

Entropy = - (class0 * log2(class0) + class1 * log2(class1))

NB for Landslide Prediction: The supervised ML method NB is used to classify the landslide possibilities yes (1) or no (0). It is based on the Bayes theorem which is used to determine the conditional probabilities i.e. the chances of one event occurring given that another has previously occurred[13].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where P (A|B) represents the posterior probability, P (B|A) is possibility probability, P (A) is the previous probability and P (B) denotes predictor prior probability.

RF for Landslide Prediction: RF is a meta estimator that employs averaging to increase the predictive accuracy and reduces overfitting by fitting various decision tree classifiers to distinct subsamples of the landslide dataset. The sub-sample size is controlled in Python with "max_samples" while "n_estimators" is used to decide the number of trees to build before the final prediction. The tree depth is controlled by "max_depth," which decides the number of splits a decision tree can make. The algorithmic flowchart is shown in figure 7.

ABC for Landslide Prediction: Adaptive boosting, referred to as ABC, is an ensemble ML technique consisting of a forest of stumps made with a node and two leaves. Stumps are technically weak learners which when combined, make a good ensemble model. The error that a first stump makes influences, how the second stump is made and vice versa.

Algorithm - ABC

Step 1: Initialize the dataset with the equal weight assigned to each data point.
Step 2: Identify the wrongly classified data points.
Step 3: Increase the weight of the data points that were wrongly classified.
Step 4: (if result is satisfactory), Goto step 5 else, Goto step 2.
Step 5: End

Support Vector Classifier (SVC): A support vector machine (SVM) belonging to the Kernel method family is a modern machine-learning method for classification problems[14]. These methods make use of linear algorithms to fabricate hyper-planes in a high or even infinite-dimensional space. The Vapnik-Chervonenkis dimension of statistical learning theory serves as the conceptual foundation for SVM theory, which implements structural risk minimization (SRM). It has a higher generalization capacity than traditional techniques that always bid in the empirical risk minimization method (ERM)[15]. The basic SVC is the basic SVC classifier that can be used to handle binary classification problems while the input sample is mapped to a high-dimensional feature space using Kernel function where the original space's crisscrossed samples become linearly scissile[16].

Logistic Regression (LR): Logistic regression (LR) performs categorical classification by assigning observations to a distinct set of classes. An output in such a classification belongs to either of the two classes either 1 or 0. LR returns a probability value by adjusting its output with the logistic sigmoid function[17]. It gives a description of the data and elucidates the connection between the dependent binary variable and one or more ordinal, nominal interval, or ratio-level Independent variables. Figure 8 depicts the algorithmic flow chart of the LR model.

Hybrid Ensemble Modeling: Two hybrid ensemble models have been developed to enhance landslide prediction. The first model, SHM, is created through the utilization of a stacking ensemble technique while the second model, VHM, employs a voting ensemble technique. Stacking, a powerful

ensemble method, allows the training of multiple base learners to address a common problem. SHM, in particular, leverages this stacking technique to merge the predictions from various base learners into a single meta-classifier. At the primary stage, the refined landslide prediction outcomes, derived from all base learners, are harmonized and employed as input data for the subsequent stage. This iterative process enhances predictive accuracy, ultimately yielding refined landslide predictions.

Base-level models are systematically trained through k-fold cross-validation where the value of k is set at 5. Moreover, at the secondary level, denoted as level '1', the meta-model is meticulously trained, incorporating the refined and significant new features extracted in prior stages. The ensemble of base-level models encompasses K-Nearest Neighbors (KNC), Random Forest (RF), AdaBoost Classifier (ABC) and Decision Trees (DT). It is worth noting that logistic regression serves as the meta-model at the secondary level, as delineated in figure 6. This comprehensive ensemble framework ensures a marked enhancement in predictive performance, establishing it as an indispensable tool for scientific research within the domain of landslide risk assessment.

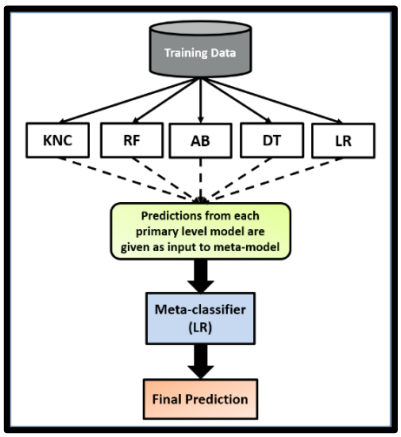

Figure 6: SHM Landslide Prediction Model

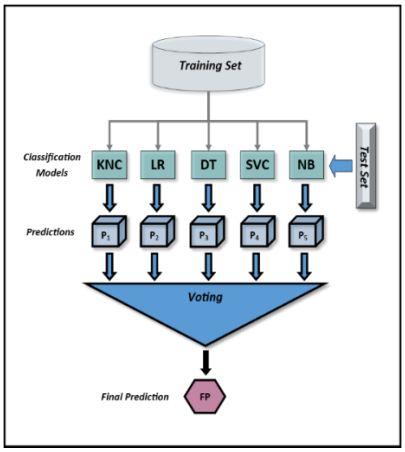

Figure 7: VHM Landslide Prediction Model

Voting hybrid model (VHM) which is an ensemble machine learning model called a voting ensemble, commonly referred to as a "majority voting ensemble," aggregates landslide predictions from various models and produces a decision using the majority voting method as shown in figure 7. It is a method for enhancing model performance, potentially surpassing the performance of any individual algorithm used in the ensemble model.

The majority vote forecast for classification has two different approaches which are as follows:

- **Hard Voting:** Predict the class with the largest sum of votes from models
- **Soft Voting:** Predict the class with the largest summed probability from models.

Four ML models KNC, LR, DT, SVC and NB, are used as the classification models to predict the likelihood of landslide using the same remote sensing training data set. The final landslide prediction is made with the largest sum of votes (Hard Voting) form all the base models.

Results and Discussion

The prediction of landslides over the Jammu Srinagar National Highway (NH44) kicks off with the collection of data spawning landslips. The data used to design the ML models has a total of 1096 instances with two attainable classes 1 and 0 where 1 represents ‘Landslide’ and 0 represents ‘No Landslide’. The missing values present in the dataset were imputed using the KNN imputation method. The anomalies and outliers were observed and replaced using the median imputation method, which is the foremost way to handle such type of errors present in the dataset.

The dataset was further balanced and scaled using SMOTE+ENN to make it more meaningful for better prediction outcome. After data preprocessing, the next step is to develop the ML models and establish the best fit model for better results. The metrics used to evaluate the performance of the models are Recall, F1 Score, Accuracy, Precision, and, ROC/AUC.

Performance Matrix: Model evaluation and performance measures are the primary conditions for adopting any model. Different performance measurements are used in this study to evaluate the various supervised ML algorithms. The accompanying evaluation techniques used are Confusion Matrix and Area under Curve (AUC).

Confusion Matrix: A confusion matrix is used to describe how a classification model performs on known true values in a set of test data. It is used to evaluate the accuracy, misclassification rate, precision, prevalence, F1 score and overall performance of the model. In this study, the actual landslide instances predicted correctly are labeled as true positive (TP), the actual non-failure instances predicted correctly are denoted by true negative (TN), actual non-failure instances that were classified as true (1) are referred to as false positives (FP). Finally, the false negative (FN)

instances demonstrate that the model predicted landslides as false (0) but they were true (1). Figure 8 shows the confusion matrices of each model used for landslide prediction.

Area under Curve (AUC) Model Evaluation Metrics: Figures 9 and 10 illustrate the ROC area under curve of the all individual classification models and both ensemble hybrid models. The area under the curve (AUC) represents the area beneath the ROC curve. It is a measure of not only how well a parameter can distinguish between the TPR and FPR but it shows to how well the model can identify different classes. The AUC evaluates a model's ability to correctly predict "No landslides" as 0 and "landslides" as 1. The higher is the AUC, the more accurate the model is; lower is the AUC, the worse the model is at predicting.

SHIFA_LSF-SVC	AP (1)	AN (0)
PP (1)	72 True Positive	128 False Positive
PN (0)	4 False Negative	125 True Negative

SHIFA_LSF-LR	AP (1)	AN (0)
PP (1)	174 True Positive	14 False Positive
PN (0)	6 False Negative	135 True Negative

SHIFA_LSF-DT	AP (1)	AN (0)
PP (1)	181 True Positive	8 False Positive
PN (0)	9 False Negative	131 True Negative

SHIFA_LSF-KNC	AP (1)	AN (0)
PP (1)	182 True Positive	6 False Positive
PN (0)	7 False Negative	134 True Negative

SHIFA_LSF-NB	AP (1)	AN (0)
PP (1)	176 True Positive	7 False Positive
PN (0)	9 False Negative	137 True Negative

SHIFA_LSF-ABC	AP (1)	AN (0)
PP (1)	183 True Positive	7 False Positive
PN (0)	6 False Negative	133 True Negative

SHIFA_LSF-RF	AP (1)	AN (0)
PP (1)	176 True Positive	12 False Positive
PN (0)	3 False Negative	138 True Negative

SHIFA_LSF-VHM	AP (1)	AN (0)
PP (1)	229 True Positive	0 False Positive
PN (0)	4 False Negative	96 True Negative

SHIFA_LSF-SHM	AP (1)	AN (0)
PP (1)	230 True Positive	0 False Positive
PN (0)	2 False Negative	97 True Negative

Figure 8: Confusion Matrices of the Models used in SHIFA_LSF-HMLM

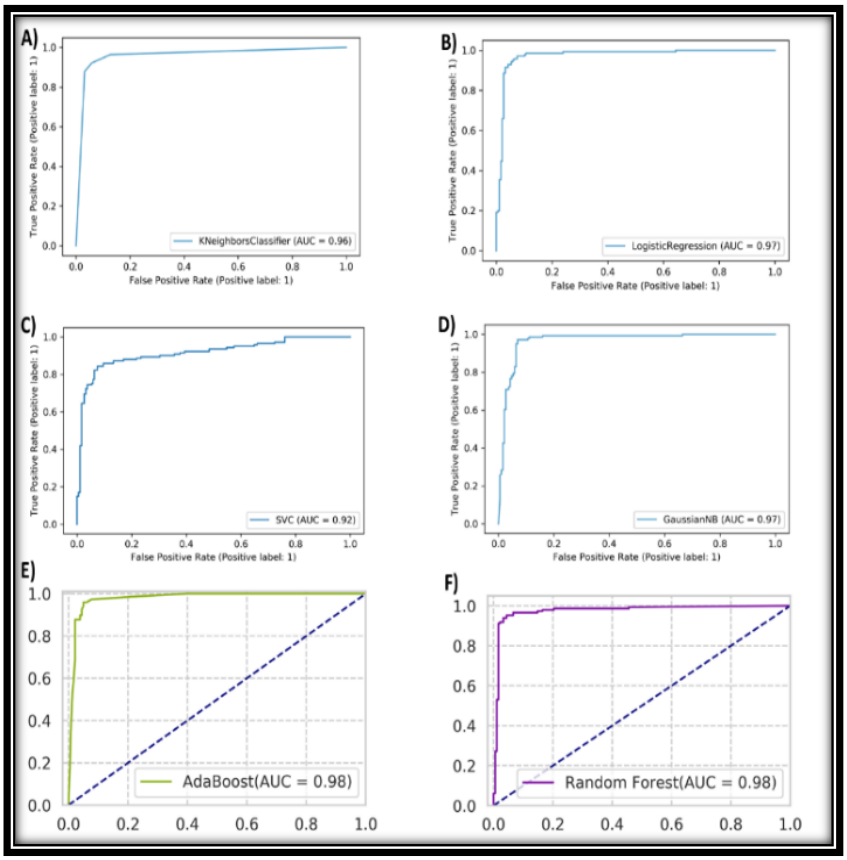

Figure 9: AUC of a) KNC, b) LR, c) SVC, d) NB, e) ABC and f) RF

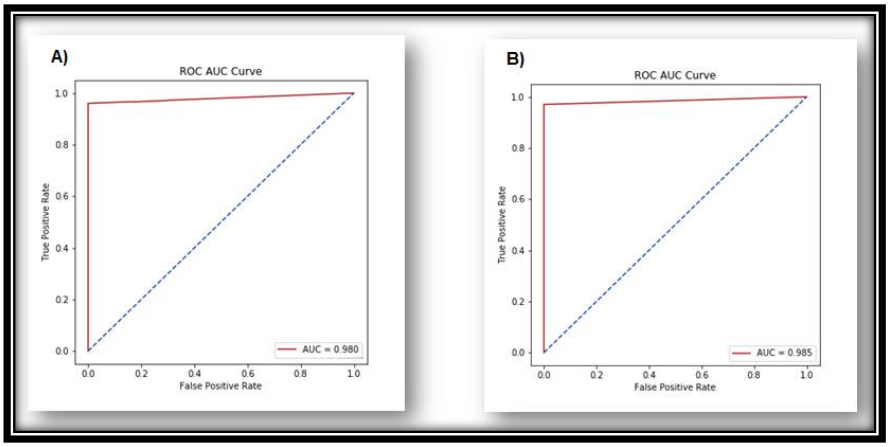

Figure 10: ROC AUC Curve a) SHIFA_LSF-VHM b) SHIFA_LSF-SHM

Table 1
Performance Evaluation of VHM Model

Model	Precision	Recall	F1 score	Accuracy	AUC
SVC	0.36	0.94	0.52	59.8	0.92
LR	0.93	0.96	0.94	93.9	0.97
DT	0.96	0.95	0.95	94.8	0.95
KNC	0.96	0.96	0.97	96.5	0.97
NB	0.96	0.95	0.96	95.2	0.97
VHM	***1***	***0.98***	***0.99***	***98.7***	***0.98***

Table 2
Performance Evaluation of SHM Model

Model	Precision	Recall	F1 score	Accuracy	AUC
ABC	0.96	0.97	0.96	96.5	0.98
DT	0.96	0.95	0.95	94.8	0.95
KNC	0.96	0.96	0.97	96.5	0.97
LR	0.93	0.96	0.94	93.9	0.97
RF	0.94	0.98	0.96	95.4	0.97
SHM	***1***	***0.99***	***0.99***	***99.4***	***0.985***

As shown in figure 13, the hybrid ensemble methods show the highest AUC than all individual algorithms while stacking outperformed the voting ensemble model with an AUC of 98.5.

Model Evaluation

The additional performance metrics employed to assess the effectiveness of each model include recall, precision and the F1 Score. Recall denotes the fraction of accurately predicted instances in each class relative to the corresponding actual instances. The results indicate that KNC accurately predicted 96% of all actual classes, RF predicted 98%, ABC predicted 97% and NB predicted 95%. Precision reflects the percentage of accurate predictions for each specific class in relation to the total predicted instances. In simpler terms, it signifies the portion of all predicted instances that were correctly predicted. RF demonstrated superior precision compared to the other individual models, achieving a precision rate of 98%. On the other hand, KNC and ABC exhibited higher accuracy rates compared to the other models. Recall and precision exhibit an inverse relationship. To render them comparable when both are important, the F1 score is introduced.

The F1 score, which calculates their harmonic mean, serves to make precision and recall comparable when both are crucial. The F1 score is used to evaluate the overall translation quality generated by the machine learning engine. The KNC model achieved an F1 score of 97% while NB, KNC, RF and DT achieved an F1 score of 96%. However, both individual algorithms were outperformed by the voting and stacking ensemble techniques. The stacking ensemble model exhibited superior performance compared to the voting ensemble model, achieving an accuracy rate of 99.4% and an AUC of 98.5%. The summarized prediction results of each model used in VHM and SHM are provided in tables 1 and 2.

Conclusion

In this research, the study area was analyzed using the field investigation and remote sensing techniques such as aerial photography, DEM analysis and slope analysis. The

resultant DEM and slope maps were meticulously crafted utilizing diverse spatial analyst tools within ArcMap 10.3. Seven finely-tuned machine learning methodologies were harnessed: the Support Vector Classifier (SVC), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors Classifier (KNC), Random Forest (RF), Naive Bayes (NB) and AdaBoost Classifier (ABC). Furthermore, two hybrid ensemble models, specifically the Stacking Hybrid Model (SHM) and the Voting Hybrid Model (VHM), were strategically devised to ascertain the most proficient model for landslide prediction.

The results indicate that all the models demonstrated strong overall performance. However, the hybrid ensemble models outperformed the individual algorithms. Notably, the Stacking Hybrid Model (SHM) showed better performance compared to the Voting Hybrid Model (VHM), achieving an accuracy of 99.4% and an AUC of 98.5%. These hybrid models surpassed the performance of all individual models, boasting the highest accuracy and AUC scores. The proposed methods exhibit exceptional robustness and consistently delivered improved results in landslide prediction. Consequently, this framework can be considered the most effective approach for implementing a Landslide Early Warning System (LEWS).

References

1. Au S.W.C., Rain-induced slope instability in Hong Kong, *Engineering Geology*, **51(1)**, 1-36 **(1998)**

2. Brabb E.E., The world landslide problem, *Episodes Journal of International Geoscience*, **14(1)**, 52-61 **(1991)**

3. Cai Y.D. and Lin S.L., Support vector machines for predicting rRNA-, RNA- and DNA-binding proteins from amino acid sequence, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, **1648(1-2)**, 127-133 **(2003)**

4. Fayaz M., Khader S.A. and Rafiq M., Landslides in the Himalayas: Causes, Evolution and Mitigation—A Case Study of National Highway 44, India, In Disaster Management in the Complex Himalayan Terrains: Natural Hazard Management, Methodologies and Policy Implications, Cham: Springer International Publishing, 43-58 **(2022)**

5. Fayaz M., Meraj G., Khader S.A., Farooq M., Kanga S., Singh S.K., Kumar P. and Sahu N., Management of landslides in a rural–urban transition zone using machine learning algorithms—A case study of a National Highway (NH-44), India, in the Rugged Himalayan Terrains, *Land*, **11(6)**, 884 **(2022)**

6. Guha-Sapir D., Vos F., Below R. and Ponserre S., Annual disaster statistical review 2010, Centre for Research on the Epidemiology of Disasters, 1-80 **(2011)**

7. Keller E.A. and DeVecchio D.E., Natural hazards: earth's processes as hazards, disasters and catastrophes, Routledge **(2019)**

8. Misra P. and Yadav A.S., Improving the classification accuracy using recursive feature elimination with cross-validation, *Int. J. Emerg. Technol*, **11(3)**, 659-665 **(2020)**

9. Maniruzzaman M., Rahman M., Al-MehediHasan M., Suri H.S., Abedin M., El-Baz A. and Suri J.S., Accurate diabetes risk stratification using machine learning: role of missing value and outliers, *Journal of Medical Systems*, **42(5)**, 1-17 **(2018)**

10. Mountrakis G., Im J. and Ogole C., Support vector machines in remote sensing: A review, *ISPRS Journal of Photogrammetry and Remote Sensing*, **66(3)**, 247-259 **(2011)**

11. Schuster R.L. and Highland L.M., The Third Hans Cloos Lecture. Urban landslides: socioeconomic impacts and overview of mitigative strategies, *Bulletin of Engineering Geology and the Environment*, **66**, 1-27 **(2007)**

12. Shah B., Sultan Bhat M., Alam A., Sheikh H.A. and Ali N., Developing landslide hazard scenario using the historical events for the Kashmir Himalaya, *Natural Hazards*, **114(3)**, 3763-3785 **(2022)**

13. Shravya C., Pravalika K. and Subhani S., Prediction of breast cancer using supervised machine learning techniques, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, **8(6)**, 1106-1110 **(2019)**

14. Singh O.P., Exploration of sparse representation techniques in language recognition, Doctoral dissertation **(2019)**

15. Vaidya R.A., Shrestha M.S., Nasab N., Gurung D.R., Kozo N., Pradhan N.S. and Wasson R.J., Disaster risk reduction and building resilience in the Hindu Kush Himalaya, The Hindu Kush Himalaya assessment: Mountains, climate change, sustainability and people, 389-419 **(2019)**

16. Wang Q., Garrity G.M., Tiedje J.M. and Cole J.R., Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, *Applied and Environmental Microbiology*, **73(16)**, 5261-5267 **(2007)**

17. Xu Z., Shen D., Nie T. and Kou Y., A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data, *Journal of Biomedical Informatics*, **107**, 103465 **(2020)**.

(Received 29th January 2024, accepted 05th April 2024)