Mapping Soil Erosion Classes using Remote Sensing Data and Ensemble Models

Ayomide Oraegbu[1], Emmanuel Jolaiya[1]

[1]Universitat Jaume I (UJI), Castellón, Spain - (ayomide.oraegbu, emmanuel.jolaiya)@uji.es

Keywords: Soil Erosion, Ensemble Models, Remote Sensing, Cross-Validation, F1-Score.

Abstract

Soil loss by water erosion is projected to increase by 13 - 22.5% in the European Union (EU) and United Kingdom (UK) by 2050, leading to loss of cultivable land and soil structure degradation. Accurate mapping of soil erosion is crucial for identifying vulnerable areas and implementing sustainable land management practices. In this study, we introduce machine learning (ML) models to map soil erosion, leveraging their capabilities in categorical mapping. Unlike previous applications that primarily mapped the absence or presence of a soil erosion class, we propose an ensemble strategy using three ML ensemble models (CatBoost, LightGBM, XGBoost) with remote sensing data to map four classes of soil erosion (i.e No Gully/badland, Gully, Badland, Land-slides). The proposed model effectively captures spatiotemporal variations over Europe in the period of 2000 - 2022, with particular precision in mapping Land-slides. The proposed method advances soil erosion mapping across different spatial and temporal scales particularly in the EU, contributing to the development of targeted conservation strategies and sustainable land management practices.

1. Introduction

Soil erosion is a general problem that affects all continents, with some experiencing its negative impacts more severely than others (Svoray, 2022), resulting in both on-site and off-site impacts on soil conservation and land management (Segarra et al., 1991). On-site impacts of soil erosion directly affect the productive capacity of the land. This means the soil becomes less fertile and unable to sustain plant growth. Off-site impacts, on the other hand, contribute to environmental degradation, particularly through water pollution. As eroded soil particles are washed away, they can carry pollutants into waterways, harming aquatic ecosystems (Issaka and Ashraf, 2017). Understanding the extent and severity of soil erosion is crucial for implementing effective mitigation strategies. Thankfully, much study has been done in this area, focusing on mapping and monitoring soil erosion across different regions and over time. Several mathematical models have been developed and applied to model soil erosion, categorized as either empirical, conceptual, or process-based (Bagarello et al., 2018). Among these models, the RUSLE (Revised Universal Soil Loss Equation) family is the most extensively used globally for soil erosion prediction (Borrelli et al., 2021). The extensive use of RUSLE models is because they are empirical models that apply a simple technique. They utilize measurable meteorological, soil, or erosion data as factors to model soil erosion losses. Additionally, RUSLE models can be applied to any spatial context once these factors can be determined (Renard and Freimund, 1994). However, it's important to acknowledge that RUSLE, as an empirical model, has limitations (Smith, 1999). For instance, they cannot effectively predict complex processes like gully erosion or mass wasting events such as landslides , as these models weren't designed for such intricate phenomena (Alewell et al., 2019). To account for these limitations, physically based models, also to be regarded as process based models (Webster, 2005), have been developed to model particularly hydrological dynamics and soil erosion processes such as sediment transport and deposition, surface runoff, rill and interrill erosion (**?**, Webster, 2005, Petkovek, 2002). However, these physically based models have their limitations as they often underpredict or overpredict the soil erosion process and as such need to be calibrated before use (Webster, 2005). Though many physically-based models including the Ephemeral Gully Erosion Model (Gudino-Elizondo et al., 2018, Woodward, 1999), and a spatially distributed adaptation of version 2 of the Revised Universal Soil Loss Equation combined with an ephemeral gully erosion estimator (Dabney et al., 2015) have been developed to model gully erosion. However, these physically-based models are unable to explain the process of gully erosion formation (Bennett and Wells, 2019).

The current improvement in computing power, modelling techniques and vast collection of datasets suggest that it is possible to overcome the limitations of currently used soil erosion models (Epple et al., 2022). The outcome of this would help in better understanding the process of formation of soil erosion, and could further lead to the development of advanced methods for mapping and monitoring soil erosion on a global scale. The use of remote sensing data in studying soil erosion has surged in recent years, with many studies leveraging this technology. Remote sensing satellite data offers new information which might not be possible through conventional field studies, as it proves to serve as an efficient source of data particularly when studying soil erosion on a large scale (Seutloali et al., 2018, Wang et al., 2023). Given the spatiotemporal nature of these satellite data, these data are continuously captured, making them valuable for long-term soil erosion monitoring (Wang et al., 2023, Zhu et al., 2024). Accessibility has improved over time, with open-satellite images becoming more accessible (Vrieling, 2007). As a result, these open satellite images serve as a valuable resource for developing cost-effective soil erosion models (Chappell et al., 2006, Mwaniki et al., 2015, Xu et al., 2019).

In response to the limitations of empirical and physically based models, there has been a notable trend in recent studies towards the use of ML to better model soil erosion dynamics (Vu Dinh et al., 2021), as this technique can better capture the relationship that exists between soil erosion controlling factors and the erosion process (Sahour et al., 2021). To overcome the limitation of limited number of parameters and lack of evaluation criteria for empirical soil erosion models, (Avand et al., 2023) integrated a RUSLE model with four(4) ML models: random

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W12-2024-135-2024 | © Author(s) 2024. CC BY 4.0 License.

forest (RF), artificial neural network (ANN), classification tree analysis (CTA), and generalised linear model (GLM) to map soil erosion hotspots in a data scarce watershed region. The RF model displayed the highest prediction performance with an AUC of 0.97 and succesfully classified the watershed region into 5 risk levels. Among the thirteen parameters used, slope angle, land cover, elevation, and rainfall erosivity are the most influential factors to determine the likelihood of soil erosion in the watershed. (Arabameri et al., 2020) compared the performance of three decision tree ML models: Alternating Decision Tree (ADTree), Naïve-Bayes tree (NBTree), and Logistic Model Tree (LMT) in mapping gully erosion susceptibility. The LMT outperformed the other two models by giving more realistic predictions of potential locations for the formation of gully erosion. In a study by (Sahour et al., 2021), ML techniques were applied in mapping annual soil erosion in a water-induced region. Coefficient of determination (R-squared), normalised root mean squared error (NRMSE), and Nash-Sutcliffe efficiency (NSE) were used as evaluation criteria. The study applied Boosted regression trees (BRT), Deep Learning (DL), and Multiple Linear Regression (MLR) to identify the factors influencing soil erosion. The BRT outperformed the other models, highlighting that models based on decision trees might outperform other ML models.

This study aims to improve upon these limitations by employing a model ensemble strategy in conjunction with remote sensing data. Ensemble models combine the predictions of multiple individual models, potentially leading to more robust and accurate predictions than a single model alone. Furthermore, remote sensing data offers vast spatial and temporal coverage, providing valuable insights for studying soil erosion across large areas (Seutloali et al., 2018, Wang et al., 2023). By leveraging the strengths of both ensemble modeling and remote sensing data, we aim to develop a model capable of mapping these erosion classes with improved accuracy across larger spatial and temporal scales.

This paper will address the following research questions:

1. How do the performances of various ML models, as documented across reviewed research studies, compare in accurately predicting soil erosion classes using remote sensing data ?
2. How much can the performance of the first three best models be improved ?
3. How can these models be combined to form an ensemble to further improve the accuracy of mapping each soil erosion class ?

By addressing these questions, this study seeks to contribute to the development of more robust and efficient soil erosion modeling techniques.

2. Data and Methods

In this section, we elucidate the naming convention and description of the various thematic groups and columns utilised in our research. The comprehensive naming convention is essential for understanding the variables and procedures involved in our analysis.

2.1 Dataset Description

The dataset used in this study constitutes a subset of an ongoing larger dataset under development (Borrelli et al., 2022). Each data point in the dataset contains a Universally Unique Identifier (UUID) which allows for easy identification. The observation year indicates when the data is collected, coordinates (longitude and latitude) are also included along with other variables discussed below. The dataset classifies erosion into four categories: No Gully/badland (0), Gully (1), Badland (2) Landslides (3).

2.2 Thematic Groups

The dataset used in this study can be categorized into the following thematic groups:

- Landsat Band (Dynamic): Includes bands like blue, green, red, NIR, SWIR1, and SWIR2. Data is derived from Landsat ARD with seasonal overlays by year.
- Lithology (Static): Represents different rock and soil types based on the (Hengl, 2018) dataset.
- Landform and Landscape Parameters (Static): Encompasses various landforms and terrain classifications based on the (Hengl, 2018) dataset.
- BioClim v1.2 (Static): Represents climate variables aggregated over the period 1981-2010, derived from the CHLSA-climate dataset (Karger et al., 2017).
- Vegetation Index (Dynamic): Includes annual EVI (MODIS) (ht-tps://modis.gsfc.nasa.gov/data/dataprod/mod13.php) data with overlays by year.
- Climate Variables (Dynamic): Incorporates dynamic climate variables with overlays by year, sourced from various datasets such as (Parente et al., 2023).
- Human Footprint (Dynamic): Represents indicators of human footprint, such as average night light intensity, derived from datasets like (Hengl et al., 2017), with overlays by year.
- Land Cover (Dynamic): Encompasses data related to land cover dynamics, including cropland percentage and forest cover percentage, with overlays by year.

2.3 Exploratory Data Analysis (EDA)

The initial stage of our analysis involved an Exploratory Data Analysis (EDA) of the dataset using Python libraries like pandas and seaborn. This EDA aimed to achieve three main objectives:

2.3.1 Identify Irrelevant Variables: During this process, it was determined that the UUID and sample id served as "unique ids" solely for identifying individual data points and as such contributed no reasonable information to mapping of soil erosion classes. As a result, these variables were excluded from further analysis.

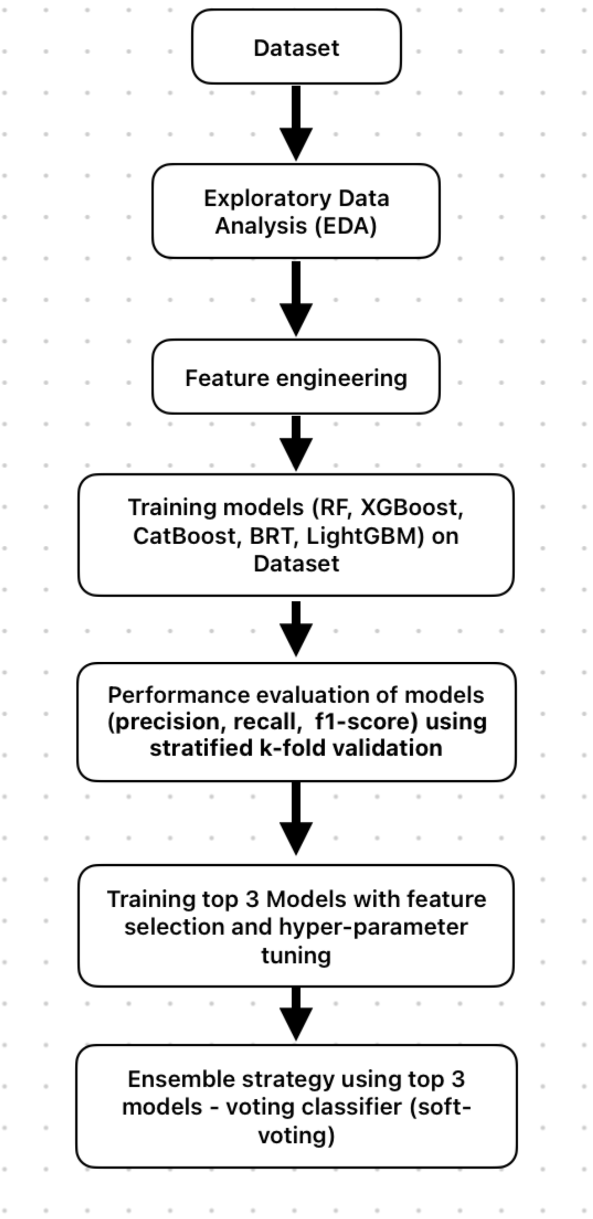

Dataset
Exploratory Data Analysis (EDA)
Feature engineering
Training models (RF, XGBoost, CatBoost, BRT, LightGBM) on Dataset
Performance evaluation of models (precision, recall, f1-score) using stratified k-fold validation
Training top 3 Models with feature selection and hyper-parameter tuning
Ensemble strategy using top 3 models - voting classifier (soft-voting)

Figure 1. Methodology workflow.

2.3.2 Understand Class Distribution: The distribution of each soil erosion class (No Gully/badland, Gully, Badland, Landslide) within the dataset was analyzed. This involved techniques like frequency distribution, which is visually represented in Figure 2. Understanding the class distribution is crucial for assessing potential model biases and ensuring the model performs well across all erosion categories.

2.4 Evaluate Spectral Bands for Erosion Mapping:

In numerous studies, researchers have used a range of data visualisation tools to uncover patterns and relationships between variables. Perhaps one of the widely used tools for representing relationships between two variables is the scatterplot. In this study, we adopt a similar approach using the bands of the Landsat satellite images. This is somewhat similar to what is known as a feature space, which is the plotting of the values of one band against another (Shivakumar and Rajashekararadhya, 2017). Feature space graphs can help find areas that contain lots of spectral information (Shivakumar and Rajashekararadhya, 2017), and this could be used to better understand what band combinations could be used for better identifying each soil erosion class. This feature space approach was implemented using a pairwise plot in the Python seaborn library. However, to get a better representation of the data, each band was multiplied by its percentile and then added to other existing similar bands of the same channel. The resulting band was then divided by 3. This was then used as a mean band when plotting the pairwise plot as shown in Figure 3.

The findings from this EDA stage provided valuable insights into the data structure, class distribution, and potential spectral features that could be informative for building robust soil erosion classification models.

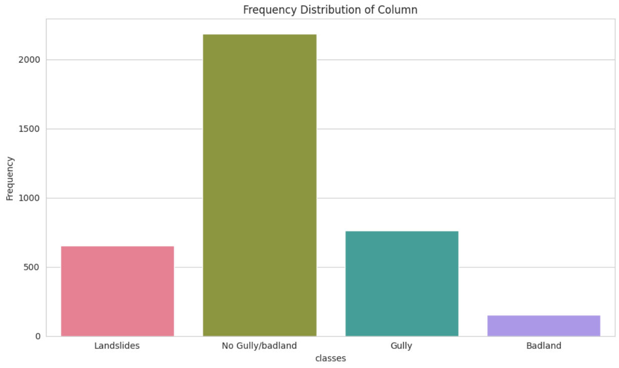


Figure 2. frequency distribution of soil erosion classes.

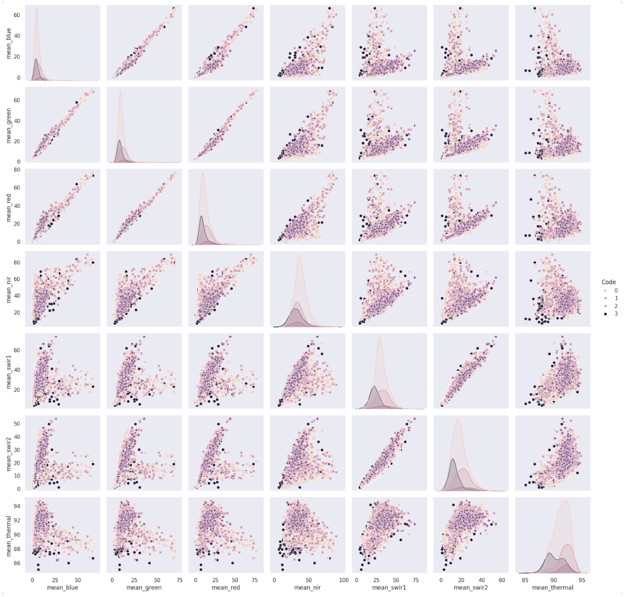


Figure 3. Exploratory Data Analysis.

2.5 Feature Engineering

A number of soil erosion studies commonly adopt an approach in which specific environmental factors perceived as influential to the soil erosion process are selected for modelling purposes. For instance, research by (Arabameri et al., 2020, Folharini et al., 2023) uses morphometric factors as key elements in soil erosion mapping, following recommendations from previous studies. Vegetation indices play a crucial role in mapping soil erosion, as demonstrated by (Moraes et al., 2018). The author used vegetation indices to predict the amount of soil loss and compared their predictive performance against fraction images derived from linear spectral mixture analysis. Notably, the findings revealed that vegetation indices outperform fraction images in accurately predicting soil erosion.

For this study, we adopted an exhaustive approach by deriving new factors from the existing dataset that are relevant to the soil erosion process, based on our findings from previous studies. Three topographic factors were derived: (1) Topographic Wetness Index (Seutloali et al., 2017, Ma et al., 2010); (2) Slope length and slope steepness factor (Seutloali et al., 2017, Huang et al., 2022, Panagos et al., 2015); (3) Stream Power Index (**?**). We also considered vegetation indices as seen in previous studies. Nine vegetation indices were derived: (4) Soil Adjusted Vegetation Index (Almutairi et al., 2013, Xue and Su, 2017); (5) Weighted Difference Vegetation Index - WDVI (Qi et al., 1994); (6) Normalised Difference Water Index (Casamitjana et al., 2020); (7) Normalised Difference Infrared Index (Sriwongsitanon et al., 2016); (8) Shortwave Infrared Water Stress Index (Fensholt and Sandholt, 2003); (9) Tasseled Cap Transformation Brightness (Zhang et al., 2002, Eniolorunda and Jibrillah, 2020); (10) Tasseled Cap Greenness (Zhang et al., 2002, Eniolorunda and Jibrillah, 2020); (11) Tasseled Cap Wetness (Zhang et al., 2002, Eniolorunda and Jibrillah, 2020). (Meng et al., 2017) suggests that low temperature coupled with enough rainfall may result in soil loss. We therefore added (12) Land Surface Temperature as a factor to model this possibility.

Observing the pairplot in Figure 3, It became evident that the relationships between certain landsat bands, attributed to their spectral information, better separate the soil erosion classes. These band combinations are; (1) near infrared and shortwave infrared 1; (2) near infrared and shortwave infrared 2; (3) near infrared and thermal; (4) shortwave infrared 1 and shortwave infrared 2; (5) shortwave infrared 1 and thermal; (6) shortwave infrared 2 and thermal. We then made these band combinations into indices using a formula similar to WDVI. We selected the WDVI because of its reduced sensitivity to soil background, which helps reduce the effects of soil noise (Qi et al., 1994). The WDVI results from the following equation;

$$WDVI = Pn - YPr \tag{1}$$

where Pn = near infrared band
Pr = red band
Y = slope line

Where Pn is the near infrared band, Pr is the red band and Y is the slope line. However, during the implementation of the formula, both Pn and Pr were substituted with the respective bands from the band combinations. Finally, we applied the following aggregation functions (mean, maximum, minimum, standard deviation, count of distinct numbers, skew, and kurtosis) to factors exhibiting comparable nomenclature.

2.6 Training Models on Dataset

The models (RF, XGBoost, CatBoost, BRT and LightGBM) were trained using the Python Language within a Google Colab Notebook using Google's free T4 GPU resources. The training of the models took an estimated time of about seven (7) minutes. Before training, the dataset underwent a stratified split into five folds. Each iteration involved training each model on four folds and validating its performance on the fifth fold. This approach of the stratified kfold cross validation enabled each soil erosion class to be adequately represented in each fold, allowing for a robust assessment of the model generalisation across different subsets of the data.

2.7 Performance Evaluation of Models

The performance of the models were evaluated using precision, recall and F1 score as suggested by (Vujovic, 2021) for classification models. The precision is calculated from the from the following equation:

$$precision = tp/(tp + fp) \tag{2}$$

where tp = number of true positives
fp = number of false positives

The precision is intuitively the ability of the classifier not to label as positive a sample that is negative (Afrah Mousa, 1970). The recall is intuitively the ability of the classifier to find all the positive samples (Afrah Mousa, 1970). The recall is calculated from the from the following equation:

$$recall = tp/(tp + fn) \tag{3}$$

where tp = number of true positives
fn = number of false negatives

The F1 score can be interpreted as a harmonic mean of precision and recall, where an F1 score reaches its best value at 1 and worst score at 0 (Afrah Mousa, 1970). F1 score is calculated from the following equation:

$$F1 = 2 * (precision * recall)/(precision + recall) \tag{4}$$

2.8 Training Top Three (3) Models with Feature Selection and HyperParameter Tuning

Among the five models considered, the top three (3) models were selected based on their performance across the three evaluation metrics (precision, recall and F1 score). Subsequently, we conducted a feature selection process to select the most effective factors contributing to the soil erosion process as well as to remove redundant factors, as highlighted by (Sánchez-Maroño et al., 2009). (Ahmadpour et al., 2021) assessed gully erosion susceptibility in a watershed region and used the Boruta algorithm for feature selection. The Boruta algorithm uses a Z-score to quantify how the removal of a factor impacts the accuracy of a model, it then uses the Z-score in eliminating the lower important factors (Kursa and Rudnicki, 2010). In a study conducted by (Farhana et al., 2023), the effectiveness of the Boruta algorithm in identifying important features for ML detection was assessed. The findings of the study revealed that the Boruta algorithm offered better accuracy by eliminating redundant features. However, the study highlighted a drawback of the Boruta algorithm, of longer time particularly on larger dataset.

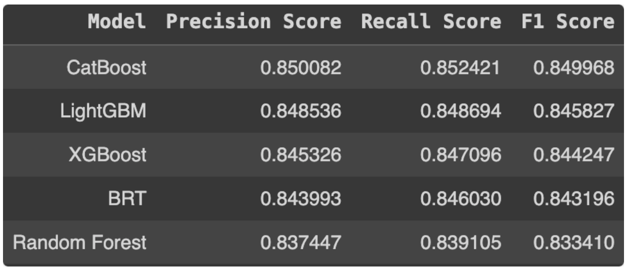
Given the large size of our dataset, we choose the Recursive Feature Elimination (RFE) algorithm. RFE algorithm is a wrapper feature selection method that uses a ML algorithm to recursively select a reduced set of features based on an input parameter "number of features to select" defined by the user - after numerous iterations, we determined this parameter to be 155 for optimal feature selection. We selected the XGBoost model as the ML algorithm to be used based on its training speed (Bentéjac et al., 2020). Hyperparameter tuning of the models was done manually, focusing on key hyperparameters such as the number of estimators, maximum depth and learning rate. The top 3 Models were trained using the factors selected by the RFE algorithm alongside their hypertuned parameters.

2.9 Ensemble Strategy

While many studies have compared the performance of individual ML models in mapping soil erosion (Nguyen et al., 2021, Fernández et al., 2023, Mohammed et al., 2023), there has been little exploration of combining models to enhance overall predictive performance. This approach is known as ensemble learning. Ensemble learning works by minimising the disparities made by the predictions from each classifier through mutual learning. In our study, we implement a unique form of ensemble learning using a voting classifier. The voting classifier trains each model on the training set, and afterwards presents the validation set to each model to predict a class label for each data point. Following this, a voting process is conducted for each data point. The voting classifier employs two voting methods - hard voting which uses a majority rule for the predicted class labels, and soft voting which sums the predicted probabilities of each model and averages it. We used soft voting as it performed better.

3. Results and Discussion

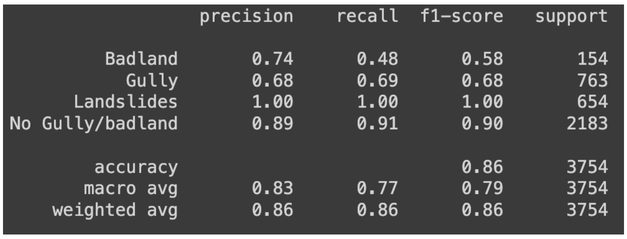
The performance of the five models (RF, XGBoost, LightGBM, BRT and CatBoost) in the validation stage using the evaluation metrics (precision, recall, F1 score) is presented in Figure 4. The CatBoost model consistently outperformed all other models across all evaluated metrics. LightGBM demonstrated the second-best performance, while XGBoost and BRT ranked third and fourth, respectively. Random Forest presented the least favourable results among the models assessed.



Model	Precision Score	Recall Score	F1 Score
CatBoost	0.850082	0.852421	0.849968
LightGBM	0.848536	0.848694	0.845827
XGBoost	0.845326	0.847096	0.844247
BRT	0.843993	0.846030	0.843196
Random Forest	0.837447	0.839105	0.833410

Figure 4. Performance evaluation of the Models (RF, XGBoost, CatBoost, BRT and LightGBM).

The ensemble approach using the top 3 Models namely; CatBoost, LightGBM, and XGBoost trained on the RFE selected factors by the voting classifier resulted in an approximate weighted average precision score, recall score and F1 score of 0.86 each, as shown in the classification report in Figure 5.



```
                   precision    recall  f1-score   support

         Badland       0.74      0.48      0.58       154
           Gully       0.68      0.69      0.68       763
      Landslides       1.00      1.00      1.00       654
No Gully/badland       0.89      0.91      0.90      2183

        accuracy                           0.86      3754
       macro avg       0.83      0.77      0.79      3754
    weighted avg       0.86      0.86      0.86      3754
```

Figure 5. Classification Report.

A confusion matrix plot was used to analyse the error matrix of the ensemble model across the four classes. The matrix provides an analysis of the actual classes to that predicted by the model. As shown in (Figure 6, it can be observed that the model could accurately predict landslides with no misclassifications. However, the model exhibited a slightly lower performance in correctly classifying no gully/badland, and significantly lower performance in classifying badland and gully.

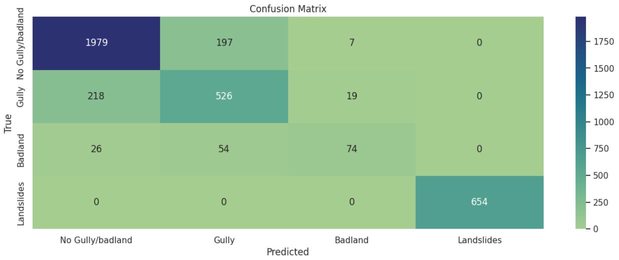


Figure 6. Confusion matrix.

The F1 score achieved by the ensemble model is approximately 0.8591, which can be rounded to 0.86. A comparison with previous studies (Avand et al., 2023) reveals that while our ensemble model's performance does not surpass theirs, it is important to note that their study focused on mapping a single soil erosion class. In contrast, our model can classify four distinct classes: no gully/badland, gully, badland, and landslides. This versatility demonstrates our model's ability to adapt well to different soil erosion scenarios, particularly in accurately identifying mass wasting events like landslides. Additionally, our model exhibits high accuracy in mapping regions with no erosion, achieving an F1 score of 0.90. However, the gully erosion class posed the greatest challenge for our model, aligning with findings from existing research that highlight the difficulty in accurately mapping gully erosion.

4. Conclusion

Mapping soil erosion class is important for any soil conservation or land management plan. However, the accuracy of mapping these soil erosion classes isn't impressive, particularly for gully erosion and mass wasting events such as landslides. We proposed a methodology that maps soil erosion on a larger spatiotemporal scale using remote sensing data and ML algorithms to map the erosion classes. Factors controlling the soil erosion process were identified and processed using remotely sensed data. Five ML algorithms (RF, XGBoost, LightGBM, BRT and CatBoost) were used to model the soil erosion process, and the performance of each model was assessed using precision score, recall score and F1 score. The top 3 models were selected namely; CatBoost, LightGBM and XGBoost. These Gradient Boosting Models were used in an ensemble strategy for the mapping of soil erosion classes. The results showed that the combination of ML models in an ensemble learning approach is effective for the spatiotemporal mapping of soil erosion classes particularly for identifying locations of landslides, and the methodology can be implemented elsewhere. A major hurdle with implementing this methodology for soil erosion mapping is its reliance on substantial data inputs for training the models. ML techniques demonstrate improved performance when trained on larger datasets. To mitigate issues associated with small datasets and to accurately quantify the performance of our trained models, we employed a stratified kfold cross validation technique, distinct from those used in model training. Our assessment of the ensemble model across both training and validation sets reveals the stability of the model in mapping soil erosion across different spatial and temporal scales, especially for mass waste events. Hence, we recommend implementing the same methodology for soil erosion mapping in comparable situations.

5. Acknowledgements

The authors extend their sincere appreciation to the Erasmus Mundus GeoTech scholarship program (mastergeotech.info), as well as Dr. Christoph Brox, for their generous support, which has been instrumental in the realization of this research.

The author would also like to acknowledge the Kaggle Competition platform for providing a valuable platform for data science exploration and experimentation.

Finally, the authors gratefully acknowledges the reference dataset, ”Monitoring gully erosion in the European Union: A novel approach based on the Land Use/Cover Area frame survey (LUCAS)”, available at https://www.sciencedirect.com/science/article/pii/S2095633921000897. This dataset provided the foundation for the research conducted in this work.

References

Afrah Mousa, Thorsten Auth, A. S. S. O., 1970. Random Walk Generation and Classification Within an Online Learning Platform. *The International Arab Journal of Information Technology (IAJIT)*, 19(3), 536 - 543. doi.org/10.34028/iajit/19/3A/14.

Ahmadpour, H., Bazrafshan, O., Rafiei-Sardooi, E., Zamani, H., Panagopoulos, T., 2021. Gully Erosion Susceptibility Assessment in the Kondoran Watershed Using Machine Learning Algorithms and the Boruta Feature Selection. *Sustainability*, 13(18). doi.org/10.3390/su131810110.

Alewell, C., Borrelli, P., Meusburger, K., Panagos, P., 2019. Using the USLE: Chances, challenges and limitations of soil erosion modelling. *International Soil and Water Conservation Research*, 7(3), 203-225. doi.org/10.1016/j.iswcr.2019.05.004.

Almutairi, B., El Battay, A., Belaid, M., Mohamed, N., 2013. Comparative Study of SAVI and NDVI Vegetation Indices in Sulaibiya Area (Kuwait) Using Worldview Satellite Imagery. *International Journal of Geosciences and Geomatics*, 1, 50 – 53.

Arabameri, A., Chen, W., Loche, M., Zhao, X., Li, Y., Lombardo, L., Cerda, A., Pradhan, B., Bui, D. T., 2020. Comparison of machine learning models for gully erosion susceptibility mapping. *Geoscience Frontiers*, 11(5), 1609-1620. doi.org/10.1016/j.gsf.2019.11.009.

Avand, M., Mohammadi, M., Mirchooli, F., Kavian, A., Tiefenbacher, J. P., 2023. A New Approach for Smart Soil Erosion Modeling: Integration of Empirical and Machine-Learning Models. *Environmental Modeling & Assessment*, 28(1), 145-160. doi.org/10.1007/s10666-022-09858-x.

Bagarello, V., Ferro, V., Flanagan, D., 2018. Predicting plot soil loss by empirical and process-oriented approaches. A review. *Journal of Agricultural Engineering*, 49(1), 1–18. doi.org/10.4081/jae.2018.710.

Bennett, S. J., Wells, R. R., 2019. Gully erosion processes, disciplinary fragmentation, and technological innovation. *Earth Surface Processes and Landforms*, 44(1), 46-53. doi.org/10.1002/esp.4522.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2020. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937–1967. doi.org/10.1007/s10462-020-09896-5.

Borrelli, P., Alewell, C., Alvarez, P., Anache, J. A. A., Baartman, J., Ballabio, C., Bezak, N., Biddoccu, M., Cerdà, A., Chalise, D., Chen, S., Chen, W., De Girolamo, A. M., Gessesse, G. D., Deumlich, D., Diodato, N., Efthimiou, N., Erpul, G., Fiener, P., Freppaz, M., Gentile, F., Gericke, A., Haregeweyn, N., Hu, B., Jeanneau, A., Kaffas, K., Kiani-Harchegani, M., Villuendas, I. L., Li, C., Lombardo, L., López-Vicente, M., Lucas-Borja, M. E., Märker, M., Matthews, F., Miao, C., Mikoš, M., Modugno, S., Möller, M., Naipal, V., Nearing, M., Owusu, S., Panday, D., Patault, E., Patriche, C. V., Poggio, L., Portes, R., Quijano, L., Rahdari, M. R., Renima, M., Ricci, G. F., Rodrigo-Comino, J., Saia, S., Samani, A. N., Schillaci, C., Syrris, V., Kim, H. S., Spinola, D. N., Oliveira, P. T., Teng, H., Thapa, R., Vantas, K., Vieira, D., Yang, J. E., Yin, S., Zema, D. A., Zhao, G., Panagos, P., 2021. Soil erosion modelling: A global review and statistical analysis. *Science of The Total Environment*, 780, 146494. doi.org/10.1016/j.scitotenv.2021.146494.

Borrelli, P., Poesen, J., Vanmaercke, M., Ballabio, C., Hervás, J., Maerker, M., Scarpa, S., Panagos, P., 2022. Monitoring gully erosion in the European Union: A novel approach based on the Land Use/Cover Area frame survey (LUCAS). *International Soil and Water Conservation Research*, 10(1), 17-28. doi.org/10.1016/j.iswcr.2021.09.002.

Casamitjana, M., Torres-Madroñero, M. C., Bernal-Riobo, J., Varga, D., 2020. Soil Moisture Analysis by Means of Multispectral Images According to Land Use and Spatial Resolution on Andosols in the Colombian Andes. *Applied Sciences*, 10(16). doi.org/10.3390/app10165540.

Chappell, A., Zobeck, T. M., Brunner, G., 2006. Using bi-directional soil spectral reflectance to model soil surface changes induced by rainfall and wind-tunnel abrasion. *Remote Sensing of Environment*, 102(3), 328-343. doi.org/10.1016/j.rse.2006.02.020.

Dabney, S. M., Vieira, D. A. N., Yoder, D. C., Langendoen, E. J., Wells, R. R., Ursic, M. E., 2015. Spatially Distributed Sheet, Rill, and Ephemeral Gully Erosion. *Journal of Hydrologic Engineering*, 20(6), C4014009. doi.org/10.1061/(ASCE)HE.1943-5584.0001120.

Eniolorunda, N., Jibrillah, A., 2020. Application of Tasselled-Cap Transformation to Soil Textural Mapping of a Semi-Arid Environment: A Case of Usmanu Danfodiyo University Main Campus, Sokoto, Nigeria. *Nigerian Journal of Environmental Sciences and Technology*, 4, 97-110. doi.org/10.36263/nijest.2020.01.0158.

Epple, L., Kaiser, A., Schindewolf, M., Bienert, A., Lenz, J., Eltner, A., 2022. A Review on the Possibilities and Challenges of Today's Soil and Soil Surface Assessment Techniques in the Context of Process-Based Soil Erosion Models. *Remote Sensing*, 14(10). doi.org/10.3390/rs14102468.

Farhana, N., Firdaus, A., Darmawan, M. F., Ab Razak, M. F., 2023. Evaluation of Boruta algorithm in DDoS detection. *Egyptian Informatics Journal*, 24(1), 27-42. doi.org/10.1016/j.eij.2022.10.005.

Fensholt, R., Sandholt, I., 2003. Derivation of a shortwave infrared water stress index from MODIS near- and shortwave infrared data in a semiarid environment. *Remote Sensing of Environment*, 87(1), 111-121. doi.org/10.1016/j.rse.2003.07.002.

Fernández, D., Adermann, E., Pizzolato, M., Pechenkin, R., Rodríguez, C. G., Taravat, A., 2023. Comparative Analysis of Machine Learning Algorithms for Soil Erosion Modelling Based on Remotely Sensed Data. *Remote Sensing*, 15(2). doi.org/10.3390/rs15020482.

Folharini, S., Vieira, A., Bento-Gonçalves, A., Silva, S., Marques, T., Novais, J., 2023. Soil Erosion Quantification using Machine Learning in Sub-Watersheds of Northern Portugal. *Hydrology*, 10(1). doi.org/10.3390/hydrology10010007.

Gudino-Elizondo, N., Biggs, T. W., Bingner, R. L., Yuan, Y., Langendoen, E. J., Taniguchi, K. T., Kretzschmar, T., Taguas, E. V., Liden, D., 2018. Modelling Ephemeral Gully Erosion from Unpaved Urban Roads: Equifinality and Implications for Scenario Analysis. *Geosciences (Basel)*, 8(4), 137. doi.org/10.3390/geosciences8040137.

Hengl, T., 2018. Global landform and lithology class at 250 m based on the USGS global ecosystem map. doi.org/10.5281/zenodo.1464846.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2), 1-40. doi.org/10.1371/journal.pone.0169748.

Huang, D., Su, L., Fan, H., Zhou, L., Tian, Y., 2022. Identification of topographic factors for gully erosion susceptibility and their spatial modelling using machine learning in the black soil region of Northeast China. *Ecological Indicators*, 143, 109376. doi.org/10.1016/j.ecolind.2022.109376.

Issaka, S., Ashraf, M. A., 2017. Impact of soil erosion and degradation on water quality: a review. *Geology, Ecology, and Landscapes*, 1(1), 1–11. doi.org/10.1080/24749508.2017.1301053.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 4(1), 170122. doi.org/10.1038/sdata.2017.122.

Kursa, M., Rudnicki, W., 2010. Feature Selection with Boruta Package. *Journal of Statistical Software*, 36, 1-13. doi.org/10.18637/jss.v036.i11.

Ma, J., Lin, G., Chen, J., Yang, L., 2010. An improved topographic wetness index considering topographic position. *2010 18th International Conference on Geoinformatics*, 1–4.

Meng, X., Wang, H., Wu, Y., Long, A., Wang, J., Shi, C., Ji, X., 2017. Investigating spatiotemporal changes of the land-surface processes in Xinjiang using high-resolution CLM3.5 and CLDAS: Soil temperature. *Scientific Reports*, 7(1), 13286. doi.org/10.1038/s41598-017-10665-8.

Mohammed, S., Jouhra, A., Enaruvbe, G. O., Bashir, B., Barakat, M., Alsilibe, F., Cimusa Kulimushi, L., Alsalman, A., Szabó, S., 2023. Performance evaluation of machine learning algorithms to assess soil erosion in Mediterranean farmland: A case-study in Syria. *Land Degradation & Development*, 34(10), 2896-2911. doi.org/10.1002/ldr.4655.

Moraes, A. G. d. L., Carvalho, D. F. d., Antunes, M. A. H., Ceddia, M. B., 2018. Relationship between remote sensing data and field-observed interril erosion. *Pesquisa Agropecuária Brasileira*, 53(3), 332–341. doi.org/10.1590/S0100-204X2018000300008.

Mwaniki, M. W., Agutu, N. O., Mbaka, J. G., Ngigi, T. G., Waithaka, E. H., 2015. Landslide scar/soil erodibility mapping using Landsat TM/ETM+ bands 7 and 3 Normalised Difference Index: A case study of central region of Kenya. *Applied Geography*, 64, 108-120. doi.org/10.1016/j.apgeog.2015.09.009.

Nguyen, K. A., Chen, W., Lin, B.-S., Seeboonruang, U., 2021. Comparison of Ensemble Machine Learning Methods for Soil Erosion Pin Measurements. *ISPRS International Journal of Geo-Information*, 10(1). doi.org/10.3390/ijgi10010042.

Panagos, P., Borrelli, P., Meusburger, K., 2015. A New European Slope Length and Steepness Factor (LS-Factor) for Modeling Soil Erosion by Water. *Geosciences*, 5(2), 117–126. doi.org/10.3390/geosciences5020117.

Parente, L., Simoes, R., Hengl, T., 2023. Monthly aggregated Water Vapor MODIS MCD19A2 (1 km): Yearly time-series (2000-2011).

Petkovek, G., 2002. Procesno utemeljeno modeliranje erozije tal process based soil erosion modelling.

Qi, J., Chehbouni, A., Huete, A., Kerr, Y., Sorooshian, S., 1994. A modified soil adjusted vegetation index. *Remote Sensing of Environment*, 48(2), 119-126. doi.org/10.1016/0034-4257(94)90134-1.

Renard, K. G., Freimund, J. R., 1994. Using monthly precipitation data to estimate the R-factor in the revised USLE. *Journal of Hydrology*, 157(1), 287-306. doi.org/10.1016/0022-1694(94)90110-4.

Sahour, H., Gholami, V., Vazifedan, M., Saeedi, S., 2021. Machine learning applications for water-induced soil erosion modeling and mapping. *Soil and Tillage Research*, 211, 105032. doi.org/10.1016/j.still.2021.105032.

Sánchez-Maroño, N., Alonso-Betanzos, A., Calvo-Estévez, R. M., 2009. A wrapper method for feature selection in multiple classes datasets. J. Cabestany, F. Sandoval, A. Prieto, J. M. Corchado (eds), *Bio-Inspired Systems: Computational and Ambient Intelligence*, Springer Berlin Heidelberg, Berlin, Heidelberg, 456–463.

Segarra, E., Ervin, R., Dicks, M. R., Taylor, D. B., 1991. On-site and off-site impacts of soil erosion: their implications for soil conservation policy. *Resources, Conservation and Recycling*, 5(1), 1-19. doi.org/10.1016/0921-3449(91)90036-N.

Seutloali, K. E., Dube, T., Mutanga, O., 2017. Assessing and mapping the severity of soil erosion using the 30-m Landsat multispectral satellite data in the former South African homelands of Transkei. *Physics and Chemistry of the Earth, Parts A/B/C*, 100, 296-304. doi.org/10.1016/j.pce.2016.10.001. Infrastructural Planning for Water Security in Eastern and Southern Africa.

Seutloali, K. E., Dube, T., Sibanda, M., 2018. Developments in the remote sensing of soil erosion in the perspective of sub-Saharan Africa. Implications on future food security and biodiversity. *Remote Sensing Applications: Society and Environment*, 9, 100-106. doi.org/10.1016/j.rsase.2017.12.002.

Shivakumar, B. R., Rajashekararadhya, S. V., 2017. Spectral similarity for evaluating classification performance of traditional classifiers. *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 1999–2004.

Smith, H. J., 1999. Application of Empirical Soil Loss Models in southern Africa: a review. *South African Journal of Plant and Soil*, 16(3), 158–163. doi.org/10.1080/02571862.1999.10635003.

Sriwongsitanon, N., Gao, H., Savenije, H. H. G., Maekan, E., Saengsawang, S., Thianpopirug, S., 2016. Comparing the Normalized Difference Infrared Index (NDII) with root zone storage in a lumped conceptual model. *Hydrology and Earth System Sciences*, 20(8), 3361–3377. doi.org/10.5194/hess-20-3361-2016.

Svoray, T., 2022. *Soil Erosion: The General Problem*. Springer International Publishing, Cham, 1–38.

Vrieling, A., 2007. *Mapping Erosion from Space*. Tropical resource management papers, Wageningen University and Research Centre, Department of Environmental Sciences, Erosion and Soil & Water Conservation Group.

Vu Dinh, T., Hoang, N.-D., Tran, X.-L., 2021. Evaluation of Different Machine Learning Models for Predicting Soil Erosion in Tropical Sloping Lands of Northeast Vietnam. *Applied and Environmental Soil Science*, 2021, 6665485. doi.org/10.1155/2021/6665485.

Vujovic, Z., 2021. Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, Volume 12, 599-606. doi.org/10.14569/IJACSA.2021.0120670.

Wang, J., Zhen, J., Hu, W., Chen, S., Lizaga, I., Zeraatpisheh, M., Yang, X., 2023. Remote sensing of soil degradation: Progress and perspective. *International Soil and Water Conservation Research*, 11(3), 429-454. doi.org/10.1016/j.iswcr.2023.03.002.

Webster, R., 2005. Morgan, R.P.C. Soil Erosion and Conservation, 3rd edition. Blackwell Publishing, Oxford, 2005. x + 304pp. £29.95, paperback. ISBN 1-4051-1781-8. *European Journal of Soil Science*, 56(5), 686-686. doi.org/10.1111/j.1365-2389.2005.0756f.x.

Woodward, D., 1999. Method to predict cropland ephemeral gully erosion. *CATENA*, 37(3), 393-399. doi.org/10.1016/S0341-8162(99)00028-4.

Xu, H., Hu, X., Guan, H., Zhang, B., Wang, M., Chen, S., Chen, M., 2019. A Remote Sensing Based Method to Detect Soil Erosion in Forests. *Remote Sensing*, 11(5). doi.org/10.3390/rs11050513.

Xue, J., Su, B., 2017. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *Journal of Sensors*, 2017, 1353691. doi.org/10.1155/2017/1353691.

Zhang, X., Schaaf, C., Friedl, M., Strahler, A., Gao, F., Hodges, J., 2002. MODIS tasseled cap transformation and its utility. 2, 1063-1065 vol.2. doi.org/10.1109/IGARSS.2002.1025776.

Zhu, J., Jin, Y., Zhu, W., Lee, D.-K., 2024. High spatiotemporal-resolution mapping for a seasonal erosion flooding inundation using time-series Landsat and MODIS images. *Scientific Reports*, 14(1), 4203. doi.org/10.1038/s41598-024-53552-9.

6. Appendix

Link to code on GitHub: https://github.com/Oraegbuayomide10/Mapping-Soil-Erosion-Classes-using-Remote-Sensing-Data-and-Ensemble-Models

Link to Kaggle Competition: https://www.kaggle.com/competitions/esa-eo4soilprotection-2024-predicting-erosion-cat