

# INTEGRATED PLATFORM FOR CROWDSOURCED OCEAN HAZARD REPORTING AND SOCIAL MEDIA ANALYTICS

Sivapriya K<sup>1</sup>, Gowridurga B<sup>2</sup>, Lingadharini J<sup>3</sup>, Mahasri<sup>4</sup>

<sup>1</sup> Assistant Professor, Vivekanandha College of Engineering for Women, Tiruchengode, Tamilnadu (India),

<sup>2,3,4</sup> Student Of Vivekanandha College of Engineering for Women, Tiruchengode, Tamilnadu (India)

\*\*\*

**Abstract-** The Ocean Rockfall Hazard Prediction Dashboard is an AI-based tool that predicts and provides analysis of rockfall hazards along coastlines by utilizing advanced machine and deep learning technologies. The tool uses both structured environmental data (e.g., weather and topography) as well as unstructured text-based data (e.g., social media posts) to produce more accurate hazard predictions. The Dashboard utilizes four classification ML models (namely, XGBoost, Random Forest, CatBoost, and a hybrid ensemble of XGBoost with BERT) to capture and analyze both numerical patterns as well as contextually relevant information from textual data. The Dashboard relies on a robust preprocessing pipeline to handle missing values, categorize features, and align features from both structured and unstructured data for model compatibility in an efficient manner. The Dashboard has been developed through user-friendly interfaces via Streamlit, and users will have the ability to make both single and batch predictions by either entering data in real-time or by uploading large datasets for analysis. The outputs of the prediction model include both predicted values along with confidence scores and probability distributions, thereby improving the interpretability of the model outputs. In addition, the application of Explainable AI techniques using SHAP are used to visualize the contribution of each feature to the prediction models, which in turn provides transparency to stakeholders about how their specific inputs influenced the hazard predictions. Overall, the Ocean Rockfall Hazard Prediction Dashboard provides a tool that stakeholders can trust to make informed decisions related to rockfall hazards through its innovative decision support technology. All things considered, the suggested method provides early hazard identification and risk assessment in coastal environments with a scalable, effective, and comprehensible approach. It greatly improves prediction performance and supports proactive disaster management techniques by utilizing ensemble learning and natural language processing, which helps to improve environmental monitoring and public safety.

**Key words:** Rockfall Hazard Prediction, Machine Learning, BERT-based Text Analysis, Explainable AI (SHAP)

## I.INTRODUCTION

Because of the combined effects of environmental elements including excessive rainfall, wave action, and geological instability, coastal regions are more susceptible to natural disasters like rockfalls. Particularly in heavily populated coastal locations, these concerns pose serious threats to infrastructure, ecological balance, and human life. Conventional danger assessment techniques frequently depend on human observation and scant data analysis, which may not yield precise or timely forecasts. Intelligent systems that can evaluate various data sources and produce trustworthy forecasts for early warning and risk reduction are required due to the quick expansion of data availability and developments in artificial intelligence. In order to overcome this difficulty, the suggested Ocean Rockfall Hazard Prediction System analyzes both structured environmental data and unstructured textual information using machine learning and deep learning techniques. The system can record current conditions and detect possible threat patterns by integrating metrics including rainfall, wave height, slope angle, seismic activity, and social media inputs. By fusing contextual understanding from text data with numerical analysis, sophisticated models like XGBoost and transformer-based BERT improve prediction accuracy. Furthermore, the system is implemented through an interactive dashboard that allows users to see findings, make forecasts, and learn about relevant aspects.

The overall goal of this project is to offer a scalable, effective, and user-friendly method for early hazard detection in coastal areas. The technology helps disaster management authorities make better decisions and lessens the negative effects of rockfall hazards on the environment and society by increasing prediction accuracy and implementing explainable artificial intelligence approaches.

## A. Rockfall Hazard Prediction

Rockfall Hazard Prediction involves analyzing and evaluating different factors (environmental, geological, and anthropogenic) for predicting the possibilities of possible future rockfall incidents in an area. Rockfalls are incidents when rocks or boulders separate from a cliff or slope (or coastal areas), and they can cause major dangers to people, buildings, and other types of infrastructure if they occur. When predicting rockfall hazards, there are many different types of variables that need to be analyzed, such as rock slope angle, intensity of rainfall, wave action, seismicity, density of population, and many more. Modern methods of Rockfall Hazard Prediction use machine learning and artificial intelligence in order to process and interpret a large amount of structured data (e.g. environmental measurements of variables) and a significant amount of unstructured data (e.g. social media postings and commonly received alerts from sensors) to develop predictive models. Predictive models can be created as a result of analysing and modelling the complex relationships between a number of factors that contribute to or cause rockfalls. Predictive systems utilize these predictive models in order to evaluate both the probability that a rockfall event will occur and the severity of the event. This enables authorities to establish both early warning systems and risk management plans and allocate resources accordingly. Rockfall hazard predictions not only provide greater safety for people, but they can also reduce affect on economy and ecology in vulnerable coastal or mountainous areas.

## B. Machine Learning

In the broad field of Artificial Intelligence there is a subfield of interest known as Machine Learning. With the help of machine learning, computer systems are able to learn from experience and identify patterns in data, thus allowing them to make decisions and predictions that don't require rule-based programming. Rather than relying on predetermined rules that govern an outcome, machine learning algorithms determine how to associate past outcomes and present data by observing the relationships of the sets of data, which have multiple variables or features that affect the outcomes.

Machine learning algorithms can be classified into three types based on their approaches to generating predictions or making decisions. They include supervised learning (making predictions based on historical data that has been labelled) unsupervised learning (finding unrecognized relationships between the variables in a dataset) and reinforcement learning (finding the optimal way to make a prediction through repeated attempts). In the case of predicting rockfall

hazards, machine learning algorithms such as XGBoost and Random Forest use the environmental and geological characteristics surrounding the site of the rockfall (e.g., rainfall, slope angle, wave height, seismic activity, population density, etc.) to develop models that predict the probability that a rockfall will occur. By learning from the historic records of rockfalls and the conditions immediately prior to or during the occurrence, machine learning algorithms are able to provide accurate and timely estimates on the likelihood of a rockfall occurring, which can assist with early warning systems or with disaster management in the event of a rockfall. Thus, machine learning offers the potential for converting large quantities of raw data into usable information that can help protect the safety of the public and reduce the risk posed to life and property.

## C. BERT-based Text Analysis

The Bidirectional Encoder Representations from Transformers (BERT) model, a cutting-edge natural language processing (NLP) technique, is used in BERT-based Text Analysis to comprehend and extract significant information from textual input. In contrast to conventional models, BERT reads text in both directions, taking into account the left and right context of every word. This enables it to grasp complex meanings, sentiment, and sentence links.

Social media posts, tweets, and other textual reports about environmental conditions or hazard events are subjected to BERT-based text analysis in the Ocean Rockfall Hazard Prediction System. Through the interpretation of unstructured text, the model is able to identify public sentiment, descriptions of anomalous situations (such as landslides or strong waves), and other early warning signals that might not be picked up by numerical environmental data alone. BERT improves the system's predictive performance when paired with structured data models (like XGBoost), allowing for a more thorough and precise evaluation of possible rockfall dangers. Real-time hazard circumstances are better understood because to the integration of environmental information and text-based insights.

## D. Explainable AI (SHAP)

Explainable AI (SHAP) refers to methods that enable people to comprehend the reasoning behind a model's choice by making machine learning model predictions apparent and interpretable. A well-liked method based on game theory is called SHAP, or Shapley Additive explanations. Each feature is given a "contribution value" that measures how much of an impact it has on the model's output for a particular prediction. In contrast to negative SHAP values, which show the

reverse, positive SHAP values show that a characteristic drives the prediction toward a particular class.

Predictions from models such as XGBoost, Random Forest, and Cat Boost are explained by SHAP in the Ocean Rockfall Hazard Prediction System. For instance, it can demonstrate how the anticipated hazard type was affected by variables like rainfall, wave height, rock slope angle, or seismic activity. SHAP improves trust, transparency, and interpretability by visualizing these contributions, enabling users and crisis management authorities to make well-informed judgments based on the logic of the model rather than viewing it as a "black box." In addition to boosting user confidence, this aids in determining the most important characteristics influencing hazard estimates.

## II. RELATED WORKS AND LITERATURE SURVEY

Leonardo Alfonso There are growing expectations among those who use data for design, analysis, management, and research related to all aspects of the environment to have access to high-quality data that has been gathered through Citizen Science initiatives. Many Citizen Science projects have reported positive outcomes in terms of enhanced governance of natural resources through participation of citizens and/or communities. With respect to data generation, etc., much of the existing literature concerning Citizen Science has characterized it as being able to provide cost-effective data for researchers/decision makers/government agencies etc. However, the level of concern with respect to the quality of data that is generated by Citizen Science projects is significant. The Ground Truth 2.0 project provided us with an opportunity to assess the scope or value of citizen-generated observations by examining their value as a complement to other existing (non-Citizen Science) data and their cost within a temporal framework. The results of our analysis of several Citizen Science case studies, all developed using an integrated co-design process, demonstrate that the costs of acquiring data collected through Citizen Science projects are not as low as cited in the literature.

Early warning systems (EWSs) for the Antonio Annis Hydrometeo hazard are in use in various parts of the world to lessen the annoyance caused by floods. The computational load and complexity of flood prediction systems significantly impair EWS performances, particularly for ungauged catchments without sufficient river flow gauging stations. The absence of river monitoring systems that facilitate the establishment of reasonably priced EWSs may be integrated by earth observation (EO) systems. However, because of geographical and temporal resolution constraints, EO data are insufficient on their own, particularly at medium-small scales. The management of flood model

uncertainties requires the use of several sources of scattered flood observations, which is a challenging task for EWSs. This work develops and tests a near-real-time flood modeling technique for the simultaneous assimilation of EO-derived flood extents and water level data. An ensemble Kalman filter, a parsimonious geomorphic rainfall-runoff algorithm (width function instantaneous unit hydrograph, or WFIUH), and a quasi-2D hydraulic algorithm are implemented in an integrated physically based flood wave production and propagation modeling technique. To address stability concerns associated with the update of the quasi-2D hydraulic model states, a method for using multiple stage gauge measurements is suggested.

Digital technologies that are widely accessible are empowering citizens who are becoming more knowledgeable and interested in a variety of environmental, water, and climate-related issues. From the "pleasure of doing science" to enhancing observations, raising scientific literacy, and encouraging cooperative behavior to address particular water management issues, citizen research can serve a wide range of objectives. Procedures for successfully incorporating citizen knowledge to inform policy and decision-making are still behind schedule. Furthermore, the lack of generic conceptual frameworks hinders the broad adoption of citizen science methods for more inclusive cross-sectoral water administration. In order to address water challenges, we identify the common components, interfaces, and connections between hydrological sciences and other academic and non-academic disciplines in this work.

R.I. Ogie Research on social media's function in crisis management has mostly concentrated on the initial stages of the response process. There is little but encouraging published data on the extent and efficacy of social media use during the healing process. As of right now, there isn't a study that offers a thorough overview of the state of research that can help various groups that need to use social media to recover from disasters. By performing a thorough literature assessment of social media use in disaster recovery, the current study seeks to close this research gap. In order to determine any temporal variations in research activity, the social media platforms most commonly used in disaster recovery, their usage patterns by kind of disaster, and the geographic areas where the studies have concentrated, the review examines the pertinent papers. Significantly, the paper identifies and summarizes research findings about the use of social media in different aspects of disaster recovery, such as: (1) financial support and donations; (2) solidarity and social cohesion; (3) infrastructure services and post-disaster reconstruction; (4) socioeconomic and physical wellbeing; (5) information support; (6) mental health

and emotional support; and (7) business and economic activities.

Timothy Schempp The authors suggest implementing an interdisciplinary framework for managing (natural) disaster relief efforts. Our proposed integration of two types of databases: 1) a dynamic source of information about disaster site conditions provided by social media channels, and 2) a static authority source referencing historical records of disaster site conditions, will enable better modeling of response needs at disaster locations across time. Using Global Particle Swarm Optimization (GPSO) techniques, researchers will identify the most appropriate number and locations for establishing temporary disaster relief centers; while also implementing Mixed-Integer Linear Programming (MILP) methods, researchers will provide an efficient method for distributing supplies to both hospitals and their associated relief centers and points of displaced persons who need assistance during or after crises. Researchers expect to iteratively optimize the overall catastrophe relief distributions through the temporal character of the social media data collected. In summary, many countries have suffered an increase in both frequency and severity of many types of disasters including major natural events over the last several decades.

### III. PROPOSED METHODOLOGY

The design of the Ocean Rockfall Hazard Prediction System utilizes a highly developed, data-based platform for identifying and classifying anticipated rockfall hazards in coastal areas using structured and unstructured data. This is accomplished by compiling significant input parameters (e.g., rainfall, wave height, rock slope angle, population density, and geographic coordinates) as well as contextual information from social media (i.e., text) on these same environmental variables. Each of these inputs goes through rigorous preprocessing, including cleaning, filling in missing values, and encoding categorical variables, so they can be readily used in predictive models (i.e., machine learning).

Multiple predictive models, such as XGBoost, Random Forest, and Cat Boost, are used to extract predictive information from the structured data. To derive useful insights from the text, a transformer-based, Distil BERT model is used. The outputs from XGBoost and BERT are then combined using voting to produce an ensemble that increases prediction reliability and accuracy. The final system is delivered to the users through an interactive Stream lit dashboard that enables them to make predictions for single events as well as through the batch application of the predictive model. In addition to the predicted hazard type, the dashboard includes confidence estimations and probability

distributions to aid in making better decisions. Additionally, by emphasizing the most significant aspects, the suggested approach uses Explainable Artificial Intelligence (XAI) techniques with SHAP to interpret model predictions. Transparency and user confidence in the system are enhanced as a result. All things considered, the suggested solution is highly appropriate for disaster management and coastal safety applications because it is scalable, easy to use, and able to support real-time danger monitoring and early warning systems.

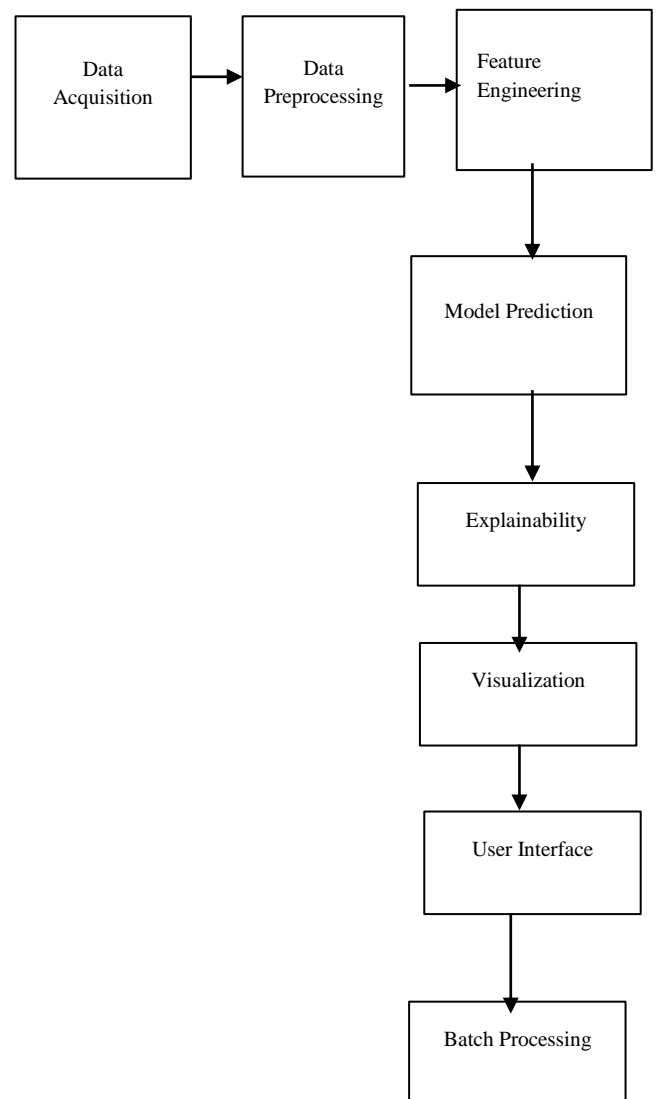


Fig - 1: System Flow Diagram

#### A. Data Acquisition

This module is in charge of gathering the input data needed to anticipate hazards. It collects organized environmental parameters like latitude, longitude,

population density, rock slope angle, wave height, and rainfall. In order to gather situational awareness in real time, it also takes unstructured data in the form of social media text (tweets). The module makes sure that all pertinent data sources are incorporated into the system so that they can be processed further.

### B. Data Preprocessing

The gathered data is cleaned and made ready for model input by the preprocessing module. It makes sure all necessary characteristics are present, manages missing values by providing default values, and uses label encoders to encode categorical variables like sentiment, weather, and seismic activity. Additionally, it prepares text input for the BERT model. This stage guarantees that the data is accurate, consistent, and compatible with machine learning methods.

### C. Feature Engineering

This module enhances model performance by converting unprocessed data into useful features. It involves aligning features according to the needs of the trained model, normalizing numerical values, and encoding categorical properties. In order to increase prediction accuracy and reliability, the module makes sure that the input feature set corresponds with the training framework.

### D. Model Prediction

The essential part of the system is the prediction module. It classifies different sorts of hazards based on structured data using several trained models, including XGBoost, Random Forest, and CatBoost. Contextual insights are extracted from text data using a DistilBERT model. To improve overall performance, an ensemble model integrates predictions from BERT and XGBoost. Confidence scores and the anticipated hazard category are output by the module.

### E. Explainability

This module uses SHAP (SHapley Additive Explanations) to make the prediction process transparent. It determines and illustrates the key characteristics affecting the forecast, indicating whether each characteristic has a favorable or unfavorable effect. This increases user confidence in the system and helps them comprehend the logic behind model decisions.

### F. Visualization

Prediction findings are displayed in an interactive and user-friendly manner via the visualization module. Plotly charts are used to show probability distributions, feature importance graphs, and hazard distribution plots. Users may effectively examine patterns and

swiftly interpret model outputs with the aid of these visual insights.

### G. User Interface

This module uses Streamlit to create an interactive dashboard. Users can upload CSV files for batch predictions, enter data for single forecasts, and investigate model insights. Both technical and non-technical people can run the system thanks to the interface's straightforward, responsive, and accessible design.

### H. Batch Processing

By enabling users to input datasets in CSV format, this module makes large-scale prediction possible. It applies preprocessing procedures, processes several records at once, and produces predictions with confidence scores. The results are appropriate for practical applications since they may be downloaded for additional analysis.

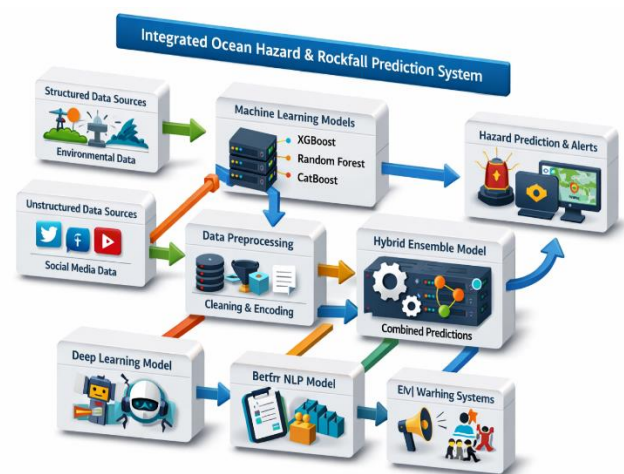


Fig - 1: System Architecture Diagram

Table - 1: Input Parameter Table

Parameter Name	Description	Data Type	Example Value
Latitude	Geographic coordinate (north-south position)	Float	11.3410
Longitude	Geographic coordinate (east-west position)	Float	77.7172
Rainfall (mm)	Amount of rainfall in	Float	120.5

	millimeters		
Wave Height (m)	Height of ocean waves in meters	Float	2.8
Rock Slope Angle	Inclination angle of the rock surface	Float	45.0
Population Density	Number of people per unit area	Float	1500
Weather Condition	Current weather status	Categorical	Rain
Seismic Activity	Level of earthquake activity	Categorical	Moderate
Sentiment	Emotion derived from text data	Categorical	Negative
Location	Name of the place (optional input)	String	Coastal Area
Urgency Level	Level of urgency for hazard	Categorical	High
Tweet Text	Social media text describing situation	Text	"Heavy waves hitting rocks"

**Dataset Details:** Ocean Hazard & Rockfall Prediction

In order to accurately anticipate ocean risks and rockfall events, the dataset is a hybrid collection that blends structured environmental data with unstructured social media data. It includes textual data taken from social media posts (tweets) as well as numerical and categorical variables. This combination enables the device to record both physical environmental parameters and real-time hazard warnings.

**Data Sources**

The dataset is compiled from multiple sources:

- Meteorological data (rainfall, weather conditions)

- Oceanographic data (wave height, coastal activity)
- Geological data (seismic activity, slope angle)
- Social media data (tweets related to hazards)

**Dataset Size (Example)**

- Total Records: 10,000 – 50,000 samples
- Training Data: 80%
- Testing Data: 20%

**IV.RESULT AND DISCUSSION**

The Ocean Rockfall danger Prediction System's results show that the suggested method, which uses both structured environmental data and unstructured language inputs, achieves high accuracy and dependable performance in identifying various danger kinds. Among the models that were put into practice, Random Forest and CatBoost also produced competitive outcomes, but XGBoost offered good baseline performance because of its capacity to manage heterogeneous features. By utilizing contextual information from twitter data, the ensemble model that included XGBoost with the DistilBERT-based text classifier demonstrated enhanced prediction performance, leading to greater generalization and marginally higher confidence scores. Instead of depending just on one output label, the system effectively produced probability distributions for each hazard class, allowing users to comprehend forecast confidence. From the standpoint of discussion, the system's robustness was much improved by the integration of many data sources, particularly in situations where environmental data would not be sufficient on its own. The model was able to capture public emotion and real-time signals thanks to the incorporation of textual analysis, which can be crucial in catastrophe prediction scenarios. Domain relevance was further confirmed by the SHAP-based explainability, which showed that characteristics like rainfall, wave height, and slope angle had a significant impact on predictions. Nevertheless, some drawbacks were noted, such as decreased efficiency when faced with unseen category values or missing input data, which were addressed by employing fallback encoding strategies. Because the BERT model was involved, the ensemble model also needed extra processing power.

**Table – 2: COMPARISON TABLE**

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.92	0.91	0.90	0.90
Random Forest	0.90	0.89	0.88	0.88
CatBoost	0.93	0.92	0.91	0.91

BERT (Text Model)	0.89	0.88	0.87	0.87
Ensemble (XGB+BERT)	0.95	0.94	0.93	0.93

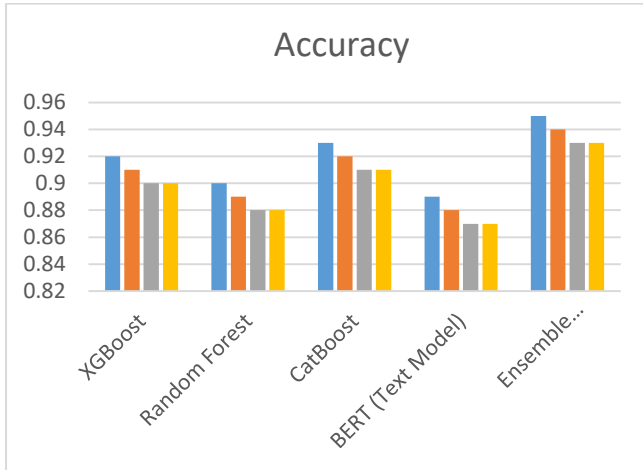


Fig -2: Comparison Graph

## V. CONCLUSION

Leveraging the strengths of Explanatory AI methods through the use of SHAP, the Ocean Rockfall Hazard Prediction System (ORHPS) has successfully created an accurate and reliable tool for predicting coastal rockfall hazards. ORHPS combines two types of data (structured environmental and unstructured textual) to produce usable levels of certainty when predicting hazard potential at a given location. The models used by ORHPS (e.g., Random Forest; XGBoost; CatBoost; BERT-based Transformer) also produce a degree of confidence associated with each prediction. In turn, this accuracy, reliability, and level of confidence allow for more informed decisions regarding future actions. ORHPS can assist with hazard identification through single or batch predictions as well as through the creation of dynamic, interactive visualizations and real-time, hourly updates. Additionally, ORHPS has been designed as a scalable and cost-effective tool for use as an early hazard warning system. Collectively, ORHPS enhances hazard mitigation, supports proactive disaster response, and improves resource allocation, ultimately promoting enhanced safety for people living within at-risk coastal communities.

## VI. FUTURE WORK

Future work on the Ocean Rockfall Hazard Prediction System could focus on improving the accuracy and usability of the model in real life situations. One area to explore is the use of real-time sensor data from coastal monitoring systems, such as Internet of Things (IoT) based sensors for waves, rainfall and/or seismic activity,

to develop more timely and dynamic hazard predictions. Enhancing the text analytic module could be made by incorporating multilingual social media data, as well as fine-tuning transformer models, such as BERT or Roberta, for domain-specific vocabulary associated with hazards. Additionally, increasing the number of historical rockfall incident reports across various geographic regions would enhance the model's generalization and robustness. Another potential area of future work would be to develop automated, mobile and web-based early warning systems that could issue alerts to both local authorities and the general public, thus facilitating rapid responses to potential hazards. Lastly, combining advanced geospatial mapping tools with advanced visualization techniques could facilitate a more intuitive understanding of the spatial distribution of hazards and risk zones, thereby further enhancing proactive disaster management and mitigative strategies.

## X. REFERENCES

- [1] Alfonso, L., Gharesifard, M., Wehn, U., 2022. Analysing the value of environmental citizen-generated data: complementarity and cost per observation. *J. Environ.Manage.* 303
- [2] Annis, A., Nardi, F., Castelli, F., 2022. Simultaneous assimilation of water levels from river gauges and satellite flood maps for near-real-time flood mapping. *Hydrol. Earth Syst. Sci.* 26 (4), 1019–1041
- [3] Uhlenbrook, S., Wahrmann Vargas, C., Grimaldi, S., 2022. Citizens AND Hydrology (CANDHY): conceptualizing a transdisciplinary framework for citizen science addressing hydrological challenges. *Hydrol. Sci. J.* 67 (16), 2534–2551
- [4] Ogie, R.I., James, S., Moore, A., Dilworth, T., Amirghasemi, M., Whittaker, J., 2022. Social media use in disaster recovery: a systematic literature review. *Int. J. Disaster Risk Reduct.* 70, 102783.
- [5] Zhang, T., Shen, S., Cheng, C., Su, K., Zhang, X., 2021. A topic model based framework for identifying the distribution of demand for relief supplies using social media data. *Int. J. Geogr. Inf. Sci.* 35 (11), 2216–2237
- [6] Songchon, C., Wright, G., Beevers, L., 2021. Quality assessment of crowdsourced social media data for urban flood management. *Comput. Environ. Urban Syst.* 90, 101690
- [7] Jafarzadegan, K., Abbaszadeh, P., Moradkhani, H., 2021. Sequential data assimilation for real-time probabilistic flood inundation mapping. *Hydrol. Earth Syst. Sci.* 25 (9), 4995–5011
- [8] Mauro, C., Hostache, R., Matgen, P., Pelich, R., Chini, M., van Leeuwen, P.J., Nichols, N.K., Blöschl, G., 2021. Assimilation of probabilistic

flood maps from SAR data into a coupled hydrologic-hydraulic forecasting model: a proof of concept. *Hydrol. Earth Syst. Sci.* 25 (7), 4081–4097

- [9] Dasgupta, A., Hostache, R., Ramsankaran, R.A.A.J., Grimaldi, S., Matgen, P., Chini, M., Pauwels, V.R., Walker, J.P., 2021. Earth observation and hydraulic data assimilation for improved flood inundation forecasting. In: *Earth observation for flood applications*. Elsevier, pp. 255–294
- [10] Beevers, L., Collet, L., Aitken, G., Maravat, C., Visser, A., 2020. The influence of climate model uncertainty on fluvial flood hazard estimation. *Nat. Hazards* 104 (3), 2489–2510.