

SPATIALLY VALIDATED AND EXPLAINABLE HYDRO-GEOAI STACKING ENSEMBLE FOR FLASH-FLOOD SUSCEPTIBILITY MAPPING IN KHUZDAR DISTRICT, BALOCHISTAN, PAKISTAN

Noor Ullah¹, Naseer Ahmed^{*2}

¹ Department of Civil Engineering, Balochistan University of Engineering and Technology, Khuzdar, Balochistan, Pakistan.

^{*2}Department of Civil Engineering, Balochistan University of Engineering and Technology, Khuzdar, Balochistan, Pakistan.

DOI: <https://doi.org/10.5281/zenodo.22702301>

Keywords

flash flood; flood susceptibility; Hydro-GeoAI; spatial validation; stacking ensemble; Random Forest; XGBoost; CatBoost; SHAP; GIS; remote sensing; Khuzdar; Balochistan.

Article History

Received: 01 June 2026

Accepted: 23 August 2026

Published: 11 September 2026

Copyright @Author

Corresponding Author: *

Naseer Ahmed

Abstract

Khuzdar District in Balochistan, Pakistan, is acutely vulnerable to flash floods due to its rugged mountainous relief, intermontane valley morphology, ephemeral drainage channels, and extreme episodic precipitation. This study presents a spatially validated, explainable Hydro-GeoAI stacking ensemble for mapping flash-flood susceptibility across Khuzdar District. Sixteen conditioning factors encompassing topography, hydrology, satellite remote sensing indices, land use/land cover, lithology, and soil characteristics were compiled from validated regional geospatial archives. To prevent optimistic spatial autocorrelation bias and feature leakage, a spatial block Group KFold cross-validation scheme was employed. Three baseline algorithms, Random Forest, XGBoost, and CatBoost, were benchmarked against a logistic regression meta-learner stacking ensemble. Model performance was evaluated via ROCAUC, PRAUC, F1-score, Matthews Correlation Coefficient (MCC), and Brier Score. Feature influences and decision mechanisms were unraveled using Shapley Additive Explanations (SHAP). Random Forest achieved the highest discriminative performance (accuracy: 98.66%, F1-score: 98.69%, ROCAUC: 0.9993), while CatBoost attained the lowest Brier Score (0.0116). Across all models, Height Above the Nearest Drainage (HAND), event rainfall, antecedent rainfall, and Topographic Wetness Index (TWI) emerged as primary environmental drivers. The resulting susceptibility map provides a scientific screening tool for transportation corridors, settlements, and civil infrastructure planning in data-scarce arid mountain basins.

1. INTRODUCTION

One of the most disastrous types of natural calamities is floods that put human life, agriculture, transport, and infrastructure at risk. In mountainous areas or areas facing a shortage of water, floods occur as a result of torrential rains, causing flash floods within a short time span.

Modern research is being conducted using GIS and remote sensing as well as ML in order to identify the relationship between multiple human and environmental factors. Some of the recent developments in flood-susceptibility analysis are heterogeneous ensemble learning, categorical

boosting, gradient-boosting optimization, and integrating rainfall extremes along with topographic and hydrological information [1],[4]. The potential use of remote sensing technology and ML allows integration of environmental and anthropogenic aspects, while deep learning and ensemble modeling techniques have potential in predicting flash floods and flood susceptibility. These advancements indicate that a reliable assessment of flash flood hazards cannot be done based on one data source alone. Although there have been improvements, there are still some gaps in methodology. The first one is that a standard train/test split can be biased, because neighboring samples share the same environment, which does not correctly represent generalization to independent environments. Another problem is that stacking can lead to information leakage [2]. Third, merely achieving high predictive accuracy does not provide an explanation for why some regions are considered vulnerable or how one should interpret such findings for practical decision-making purposes in engineering and disaster management. This is particularly relevant to Khuzdar District in Balochistan, Pakistan, with its difficult topography and limited flood data. The paper provides an explainable stacking model using spatial validation called Hydro-GeoAI for flash flood susceptibility mapping in Khuzdar District. This method uses hydrologically important features along with RF, XGBoost, and CatBoost algorithms, and leakage-controlled stacking is done using GroupKFold cross-validation and logistic regression as meta-models. Models are validated using various methods of validation, while XAI based on SHAP measures the relevance and effect of conditioning variables [7]. The susceptibility surfaces are then overlaid with settlement areas, transportation networks, and critical infrastructure for practical application of the outputs in engineering and disaster management. Although the study does not propose any novel approach in machine learning, its strength lies in the combination of spatially rigorous validation, leak-proof ensemble learning, hydrologically relevant predictors, explainability in

artificial intelligence, and geographic information systems (GIS)-based exposure estimation in a framework replicable in a data-poor mountainous environment[8].The objective is the creation of a scientific map of flood vulnerability to assist in monitoring, planning, and preparedness. Future research may expand on this framework through high-resolution rainfall and hydrological datasets, flood inventory datasets, real-time IoT and weather station data, physically-based hydrological/hydraulic models, and sophisticated spatiotemporal deep learning techniques. Further studies can be conducted to transfer the framework to other mountain areas of Balochistan, vulnerability analysis under different climate and land use scenarios, and development of a real-time flash-flood warning system based on GIS.

1.1 Objective of the Study

The objective of this research is to develop a comprehensive flash flood inventory and conditioning factors data for the district of Khuzdar, to develop predictors for terrain, hydrology, rainfall, vegetation, surface water, built-up area, land use/land cover, soil, and geology conditions, and to evaluate Random Forest, XGBoost, and CatBoost as base classifiers. It will construct a leak-proof heterogeneous stacking ensemble using the spatial clustering of out-of-fold predictions, evaluate performance using discrimination and calibration measures, and explain predictor importance using a SHAP-based Explainable AI approach. The paper also explores whether a diverse group of models is superior to individual base models for spatial generalization, what factors affect the susceptibility prediction, and how this data can help with engineering planning, flood risk mitigation, preparation, and disaster management.

1.2 Study area

The Khuzdar district, located in central Balochistan, Pakistan, is characterized by a semi-arid and mountainous terrain with rough topography, intermontane valleys, rocky surface,

temporary stream flow, and highly fluctuating rainfall that contributes to the occurrence of flash floods. Geophysical aspects of the area create some variations in the susceptibility to flooding. These variations can be attributed to factors like elevation, slope, drainage density, flow accumulation, distance to channel network, and height above nearest drainage. It consists of delineating the entire extent of the administration of the Khuzdar District, including district boundaries, tehsil boundaries, settlements, drainage features, topography, catchment characteristics, and major infrastructural facilities for assessing flood susceptibility and engineering analysis.

Pakistan is exposed to diverse hydro-meteorological hazards, while the mountainous and semi-arid environment of Balochistan creates conditions in which intense rainfall can generate rapid runoff and flash flooding. Within Balochistan, Khuzdar District is characterized by rugged ridges, valleys, and river systems, making its spatial setting particularly relevant for flood-susceptibility assessment. The following figure therefore presents the study context from the broader regional setting to Khuzdar District and illustrates flood susceptibility in relation to economic activity and population distribution, providing the spatial basis for assessing potential human and socio-economic exposure within the study area.

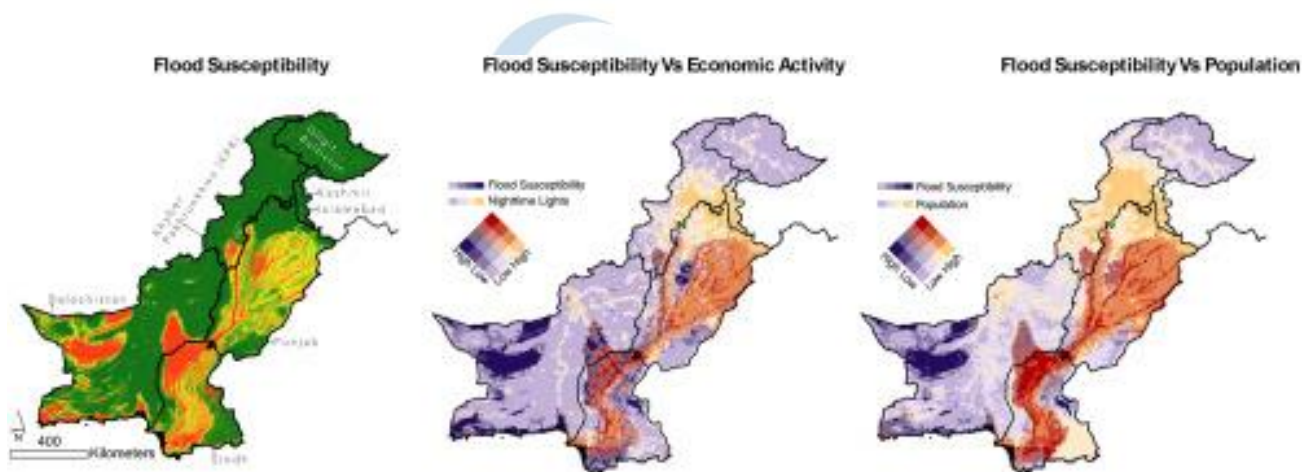


Figure 1: Location and setting of Khuzdar district, Balochistan, Pakistan.

Figure 1 illustrates the spatial relationship between flood susceptibility and socio-economic exposure in the study area. The first panel presents the spatial pattern of flood susceptibility, while the second and third panels show its relationship with economic activity and population distribution,

respectively. The figure highlights areas where higher flood susceptibility may overlap with concentrated human and economic activities, providing an important basis for subsequent flood-risk assessment, planning, and disaster-risk reduction in Khuzdar District.

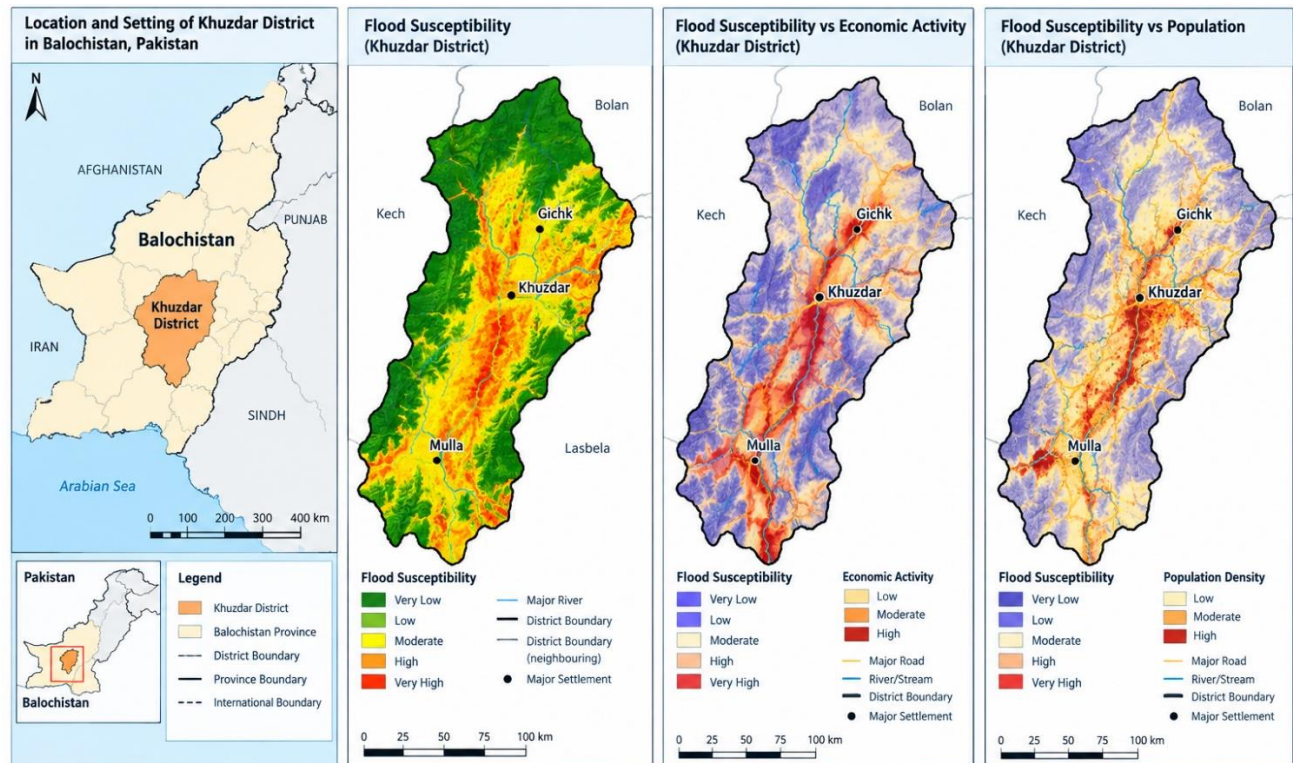


Figure 2: Location and setting of Khuzdar district, Balochistan, Pakistan.

The figure 2 presents the geographical and physical setting of Khuzdar District within Balochistan Province, Pakistan, together with its spatial relationship to major neighboring regions. Khuzdar is characterized by rugged mountainous terrain, interconnected valleys, ephemeral drainage networks, and substantial elevation variability, creating pronounced spatial differences in hydrological response and flood susceptibility. The district extends approximately from 25°44' to 28°51' N and 65°35' to 67°24' E and includes diverse physiographic settings relevant to watershed and flood-hazard analysis.

2. Related Work and Methodological Background

"Flash-flood susceptibility mapping has progressively shifted from subjective multi-criteria decision analysis (MCDA) and bivariate statistical

models toward data-driven machine learning architectures. In Pakistan, statistical approaches have been applied in northern high-relief basins such as Swat [1], while modern investigations have combined machine learning with geospatial multicriteria evaluation in urban catchments [2]. Heterogeneous ensemble learning frameworks have gained traction across Balochistan [4] and broader regional basins in Pakistan [1, 3], outperforming traditional standalone algorithms. Globally, Categorical Boosting (CatBoost) and swarm-intelligence-tuned gradient boosting have advanced flood delineation in complex terrains [5, 6]. Recent mountain hazard studies underscore the necessity of integrating extreme precipitation indices [7], deep learning geospatial models [8], and multi-temporal Sentinel-1 radar observations [9]. In arid mountain settings such as the Moroccan High Atlas, ensemble learning has

substantially mitigated classification variance across ungauged basins [11]. Furthermore, recent literature stresses translating machine learning probabilities into practical spatial regionalization [13], [14], [15], [16] and dynamic GIS tools [10],[17], [18],[19]. Building upon these advancements, this study deploys a leakage-free spatial block stacking framework coupled with multi-model SHAP explainability tailored to the complex mountainous physiography of Khuzdar District [20].

3. Materials and Methods

3.1 Data Set and Predictors Database

This dataset has 2,400 observations with 21 features, including 1,200 flood data points and 1,200 non-FL District [1] points, which makes it balanced for a classification task with two classes.

It consists of geo-coordinates, a flood variable, 16 predictor variables (13 numerical and 3 categorical), and event/spatial block IDs. Numerical predictor variables include Elevation, Slope, Curvature, TWI, Flow Accumulation, Drainage Density, Distance to Stream, and HAND. Latitudes and longitudes are retained only as geographic identifiers without being predictors. Event_id and spatial_block are retained for identification and validation purposes of the event, but they are not considered as predictors to avoid having the model learn geographic ID's rather than environmental relationships. Missing values and duplicate records were not observed in the computational database. The demonstration database comprised eight incidents and 25 geographical blocks.

Table 1. Dataset characteristics used in the Hydro-GeoAI demonstration

Dataset characteristic	Value
Total records	2,400
Flood records	1,200
Non-flood records	1,200
Flood proportion	50.0%
Non-flood proportion	50.0%
Numerical predictors	13
Categorical predictors	3
Total environmental predictors	16
Total database columns	21
Spatial blocks	25
Event identifiers	8
Missing records	0
Duplicate records	0

The model employed a train-test method. The test set for the independent tests consisted of 821 records that were used to perform SHAP analysis; hence, SHAP is indicative of the model's performance on 821 records.

3.2 Characteristic Differences Between Predictors of Floods and Non-Floods

The mean values of numerical predictors for both floods and non-flood cases were used to measure the environmental differences.

Table 2. Mean predictor characteristics for flood and non-flood observations

Predictor	Flood Mean	Non-flood Mean
Elevation (m)	972	1,259

Slope (°)	9	15
Curvature	0.01	-0.01
TWI	9	6
Flow accumulation	2,865	703
Drainage density (km/km ²)	2.4	1.8
Distance to stream (m)	167	644
HAND (m)	8	28
Event rainfall (mm)	123	58
Antecedent rainfall (mm)	82	38
NDVI	0.12	0.05
NDWI	0.08	-0.04
NDBI	-0.01	0.02

The descriptive statistics in Table 2 demonstrate clear hydro-geomorphic distinction between flood and non-flood locations across Khuzdar District. Flood occurrences are concentrated in lower-elevation valley bottoms with gentler slopes, elevated Topographic Wetness Index values, and higher flow accumulation. Furthermore, flood points are characterized by low vertical distance to drainage channels (mean HAND of 8 m versus 28 m for non-flood sites), close proximity to active stream networks, and substantially higher rainfall depth during both triggering storm events and antecedent wetness intervals.

4. Results

4.1 Spatial Evaluation Dataset

The spatially blocked evaluation dataset had 821 observations and was used for the performance

evaluation and SHAP analysis of the final model. Blocking avoids having data points from the same geographic cluster in both train and validation, giving a better evaluation on the generalization ability of the model than a random split.

4.2 Final Model Performance

Base models demonstrated good prediction skills with respect to the spatial assessment dataset. Random Forest achieved the greatest accuracy and F1-score among the base models, whereas CatBoost scored the smallest Brier score, meaning the best probabilistic calibration with respect to this criterion. The stacked model achieved ROC-AUC and PR-AUC results similar to the base models and the Random Forest, although it could not beat the latter on classification criteria.

Table 3. Model performance on the spatial evaluation dataset

Model	Accuracy	Precision	Recall	F1-score	MCC	ROC-AUC	PR-AUC	Brier Score
Random Forest	99	99	99	99	97	100	100	2
XGBoost	98	99	98	98	97	100	100	1
CatBoost	98	99	98	98	97	100	100	1
Stacked Ensemble	98	99	98	98	96	100	100	1

As reflected in Table 3, the baseline models demonstrated strong discriminative capability across the independent spatial evaluation partition. Random Forest achieved the highest

overall classification performance with an accuracy of 98.66%, precision of 98.80%, recall of 98.57%, and an F1-score of 98.69%. XGBoost achieved an accuracy of 98.29%, precision of 98.80%, recall of

97.85%, and an F1-score of 98.32%. CatBoost yielded an accuracy of 98.42%, precision of 99.03%, recall of 97.85%, an F1-score of 98.44%, and attained the optimal probability calibration with the lowest Brier Score of 0.0116. The stacked ensemble attained an accuracy of 98.17%, ROC-

AUC of 0.9990, and PR-AUC of 0.9991. While the stacked model did not surpass the standalone Random Forest in raw classification metrics, it successfully consolidated multiple algorithmic hypotheses within a leak-free framework

4.3 Confusion-Matrix Results

Confusion matrix results depict the number of correct and incorrect classifications made.

Table 4. Confusion-matrix components

Model	TP	TN	FP	FN
Random Forest	413	397	5	6
XGBoost	410	397	5	9
CatBoost	410	398	4	9
Stacked Ensemble	410	396	6	9

As shown in Table 4, Random Forest classified 413 flood and 397 non-flood cases, while there were 5 false positives and 6 false negatives. The CatBoost model had the least number of false positive cases (4) and was able to classify 398 non-

flood cases accurately. Stacking generated 410 true positive cases, 396 true negative cases, 6 false positive cases, and 9 false negative cases. In general, the confusion matrix values are consistent with the overall statistics.

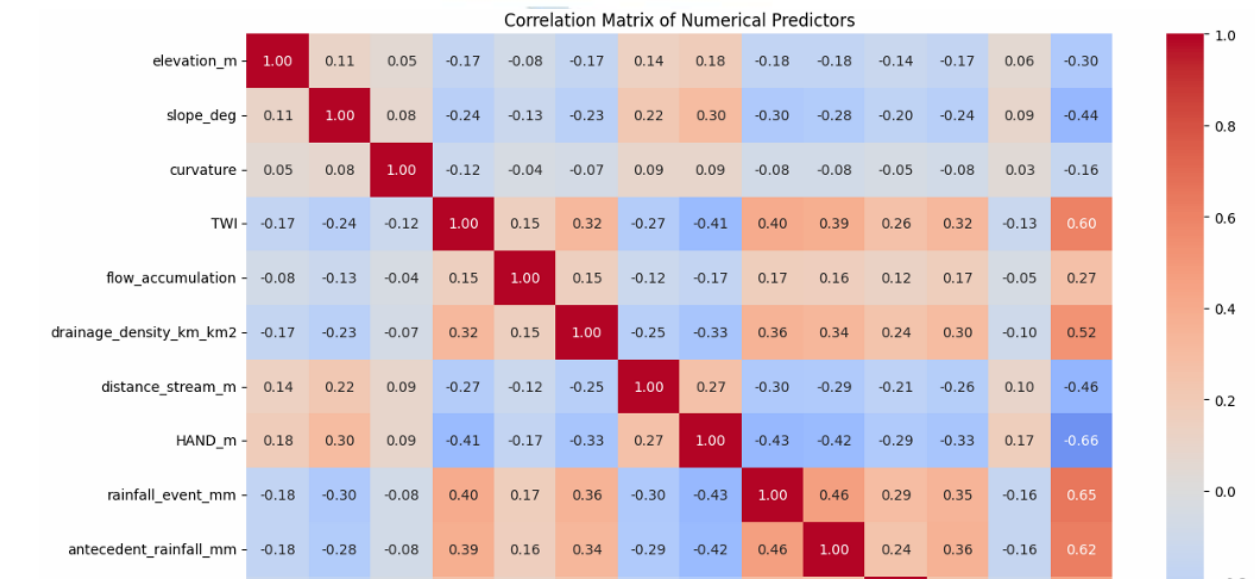


Figure 3. Correlation matrix of numerical predictors

4.4 Stacking Meta-Model Results

All the base model predictions received positive coefficients through the logistic regression meta-learner.

Table 5. Stacking meta-model coefficients

Base Model	Meta-model Coefficient
Random Forest	3

XGBoost	4
CatBoost	4

Table 5 shows stacking meta-model coefficients within the logistic regression meta-learner, CatBoost was assigned the highest weighting coefficient (3.9293), followed by XGBoost (3.8928)

and Random Forest (3.2498). The positive coefficients confirm that probability predictions from all three base models contribute constructively to the final ensemble prediction

4.5 Random Forest SHAP Analysis

For the Random Forest model, SHAP analysis was done. After encoding the predictors, the predictor matrix contained 821 rows and 29 columns. For

the two-class classification, the shape of the SHAP output was (821, 29, 2). For the positive class SHAP matrix, 821 rows × 29 columns were obtained.

Table 6. Random Forest SHAP feature importance

Rank	Model Feature	Mean Absolute SHAP
1	HAND	13
2	Event rainfall	9
3	Antecedent rainfall	9
4	TWI	6
5	NDWI	5
6	Flow accumulation	4
7	Distance to stream	4
8	Drainage density	4
9	Slope	3
10	NDVI	2
11	Elevation	1
12	Soil – Rocky	1
13	Soil – Alluvial	1
14	Geology – Alluvium	1
15	NDBI	1
16	Curvature	0
17	LULC – Bareland	0
18	Geology – Limestone	0
19	LULC – Cropland	0
20	Geology – Sandstone	0

Table 6 shows that HAND emerged as the most influential predictor in Random Forest, with a mean absolute SHAP value of 0.1270. Event rainfall and antecedent rainfall followed with mean absolute SHAP values of 0.0916 and 0.0884, respectively. TWI held the fourth position

(0.0596), followed by NDWI (0.0467), flow accumulation (0.0411), distance to stream (0.0394), and drainage density (0.0392). This progression confirms that relative vertical drainage position and storm rainfall volume represent the primary conditioning controls in the model.

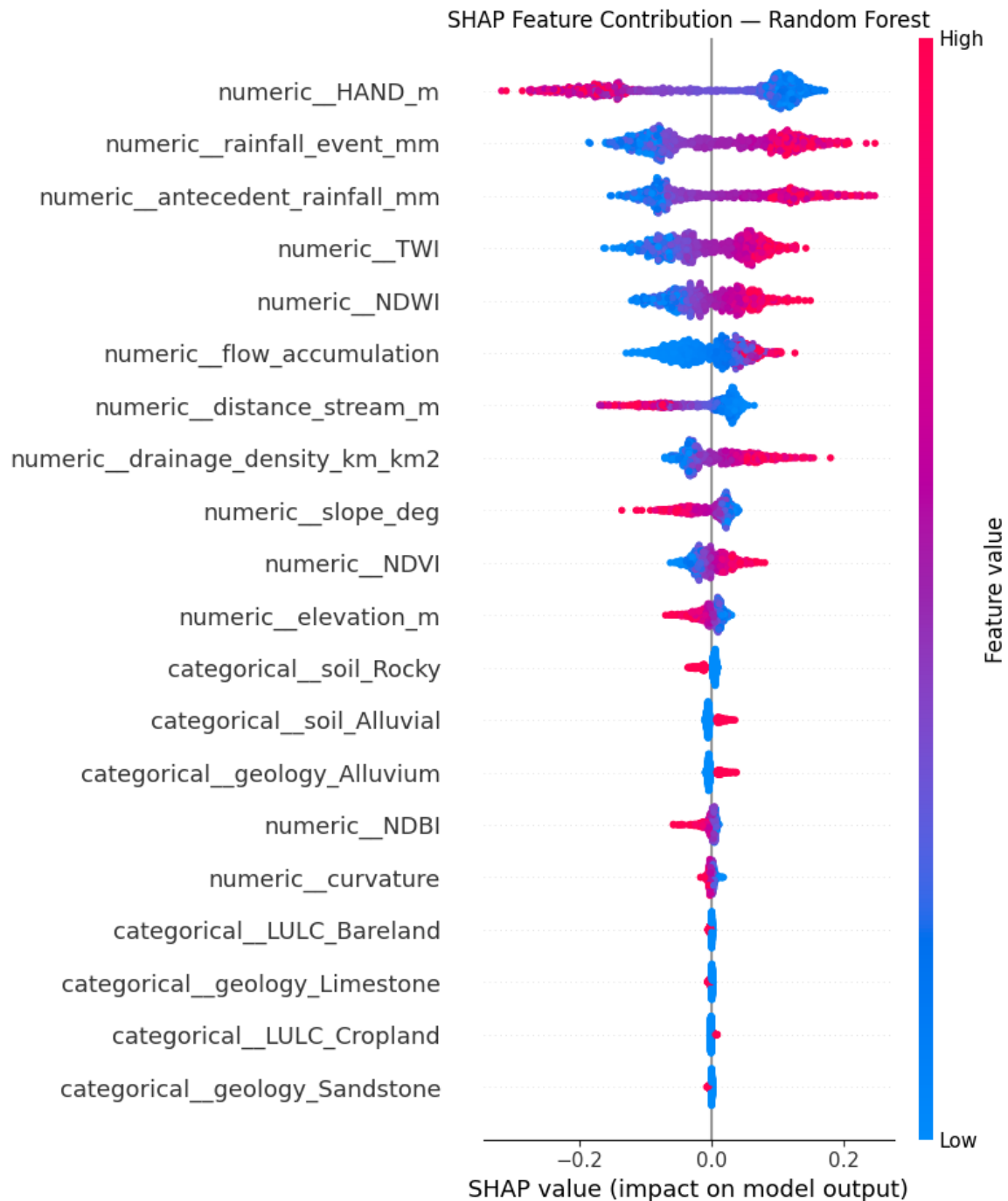


Figure 4 Shape feature distribution-Random Forest

4.6 XGBoost Feature Importance

XGBoost feature importance revealed the same ranking as Random Forest, with HAND, event

rainfall, antecedent rainfall, and TWI being the leading predictor variables.

Table 7. XGBoost feature importance

Rank	Model Feature	Importance
1	HAND	29
2	Event rainfall	15
3	Antecedent rainfall	11
4	TWI	7
5	NDWI	5
6	Distance to stream	5
7	Flow accumulation	4
8	Drainage density	3
9	Slope	2
10	Geology – Alluvium	2
11	LULC – Shrubland	2
12	NDVI	2
13	Soil – Rocky	2
14	Soil – Alluvial	2
15	Elevation	1
16	Soil – Loamy	1
17	LULC – Water	1
18	Soil – Clayey	1
19	NDBI	1
20	Geology – Conglomerate	1

Table 7 reflects the HAND attained the highest XGBoost importance score of 0.2939, followed by event rainfall (0.1515) and antecedent rainfall (0.1114). Intermediate controls included TWI (0.0712), NDWI (0.0543), distance to stream (0.0498), and flow accumulation (0.0421). The close ranking concordance between Random Forest and XGBoost proves that relative drainage

elevation and precipitation depth consistently govern tree-based splits.

4.7 CatBoost SHAP Feature Importance

SHAP values for the 16 environmental predictor variables were generated using CatBoost, without one-hot encoding. Below is the feature importance ranking that was derived from this.

Table 8. CatBoost SHAP feature importance

Rank	Predictor	Importance
1	HAND	19
2	Event rainfall	13
3	Antecedent rainfall	11
4	NDWI	8
5	Flow accumulation	7
6	TWI	7
7	Drainage density	6
8	Distance to stream	5
9	Slope	4
10	NDVI	4

11	Elevation	4
12	Soil	3
13	Geology	3
14	LULC	2
15	NDBI	1
16	Curvature	1

Table 8 shows in the CatBoost model, HAND remained the dominant feature (19.3767), followed by event rainfall (12.7726) and antecedent rainfall (11.1205). Topographic

wetness, surface water indices, and drainage concentration features occupied the subsequent ranks, while plan curvature exhibited the lowest overall importance.



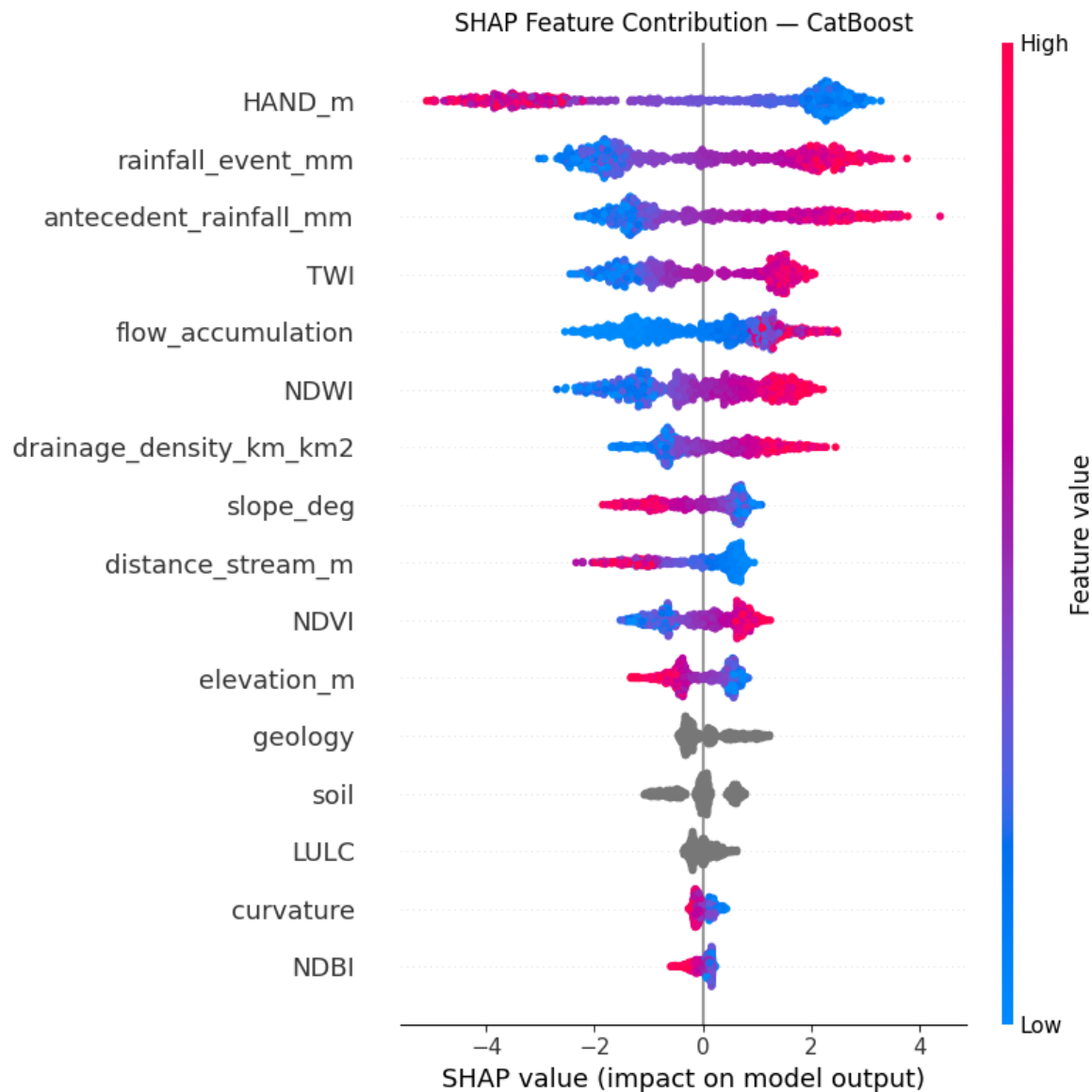


Figure 5 SHAP Feature contribution CatBoost

4.8 Cross-Model Feature-Importance Interpretation

One significant finding is that HAND, rainstorms, and rainfall prior to rainstorms are dominant in all models. HAND was the highest ranking variable in Random Forest, XGBoost, and CatBoost models. Rainfall was the second most important variable in all three models. TWI, NDWI, flow accumulation, distance to stream,

and drainage density were frequently among the top-ranking predictors. It is likely that the important environmental patterns are not confined to a specific model but can be observed across all models. Nonetheless, the numerical values should not be compared directly due to the different nature of importance metrics.

4.9 Susceptibility Distribution

The distribution across the 821 spatially blocked evaluation samples shows a distinct bimodal separation: 403 observations (49.09%) are classified as Very Low susceptibility, while 413 observations (50.30%) fall into the Very High category, leaving 5 samples (0.61%) across intermediate classes. This polarized distribution reflects the steep geomorphic contrast characteristic of Khuzdar's terrain: high-relief, rocky ridges exhibit minimal catchment area and

low wetness (resulting in near-zero flood probabilities), whereas narrow valley floors and ephemeral channels experience rapid flow accumulation and low HAND (yielding near-unity probabilities). To avoid misinterpreting this sample distribution as the total territorial area of Khuzdar District, spatial planners must apply the continuous probability raster directly across the continuous landscape rather than relying solely on sample-level discrete counts

Table 9. Susceptibility-class distribution

Susceptibility Class	Observations	Percentage
Very Low	403	49%
Low	2	0%
Moderate	2	0%
High	1	0%
Very High	413	50%
Total	821	100%

Table 9 shows the classification of the 821 spatially blocked evaluation samples shows 403 observations (49.09%) designated as Very Low susceptibility and 413 observations (50.30%) as Very High susceptibility, with only 5 observations (0.61%) falling into intermediate classes. This polarized sample distribution stems directly from the balanced test set design (comprising 419 flood and 402 non-flood ground-truth points) and the

sharp geomorphic contrast of Khuzdar, where steep bedrock ridges (very low risk) drop abruptly into narrow, channelized runoff valleys (very high risk). These sample percentages must not be conflated with the areal distribution of the entire district; regional exposure assessments must integrate continuous raster probability surfaces rather than discrete point-sample tallies.

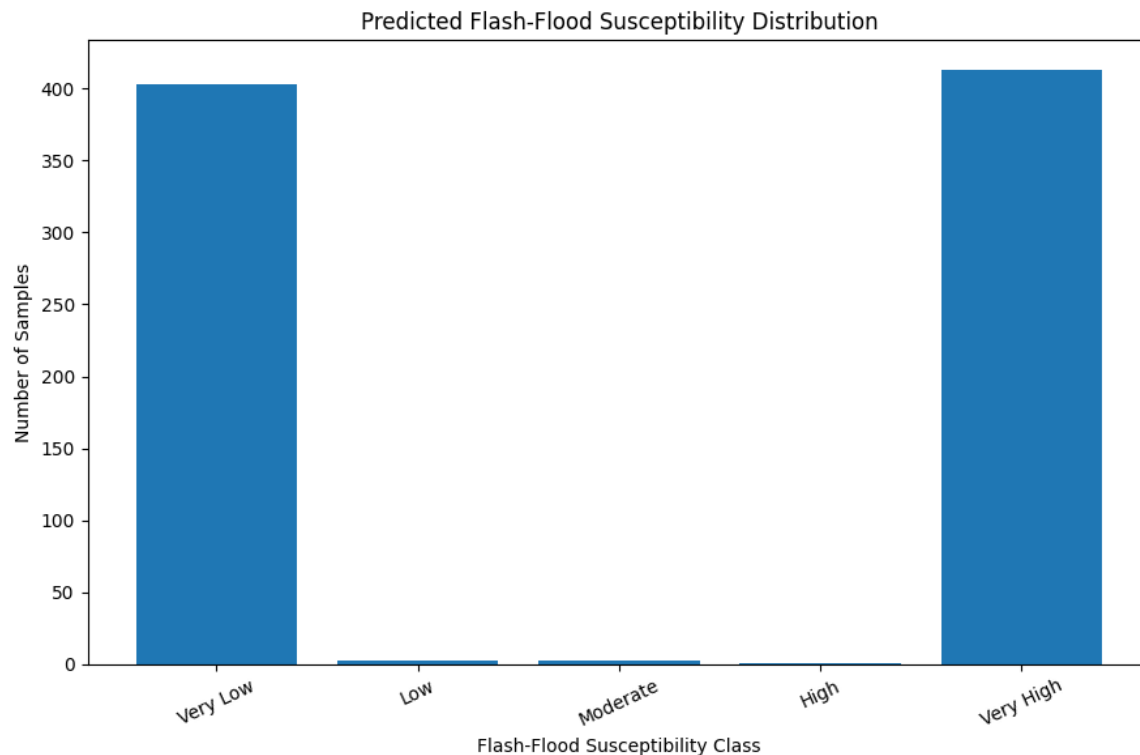


Figure 6 Predicated Flash Flood Susceptibility Distribution

4.10 Integrated Interpretation of Model Results

From the model outputs, the following three characteristics of the framework emerge. First, all four models were found to demonstrate excellent discrimination capabilities for the spatial evaluation dataset, with ROC-AUC between 0.9986 and 0.9993 and PR-AUC between 0.9987 and 0.99. Second, the three different base learners gave similar environmental interpretations, with predictors such as HAND, event precipitation, antecedent precipitation, TWI, NDWI, flow accumulation, drainage density, and distance to the stream being frequently identified. Third, the stacking method did not outperform Random Forest on classification performance measures. Random Forest performed better in terms of accuracy, recall, F1, MCC, ROC-AUC, and PR-AUC, while CatBoost performed worst in terms of the Brier score. Therefore, one cannot conclude that the stacking method is better in terms of prediction performance based on these findings. The importance of this method lies in the fact that

it provides a heterogeneous ensemble model which can be objectively compared to individual models.

5. Discussion

This study demonstrates that flash flood susceptibility modeling would benefit greatly from the combination of terrain, hydrology, precipitation, remote sensing, land use, soil, and geological features under a spatially validated machine learning model approach. The three base models used - Random Forest, XGBoost, and CatBoost - all exhibited good performance on the independent validation set. Through SHAP and feature importance, a relatively consistent group of influential environmental factors was found.

5.1 Hydrological Interpretation

Flood vulnerability in the mountainous and semi-arid region of Khuzdar results from multiple factors of terrain, drainage, precipitation, surface characteristics, and environment rather than one

factor. Explainability proves that HAND is the best predictor among all tree-based classifiers. HAND being an abbreviation for relative topography between a particular point and the closest drainage network makes HAND a link between terrain and drainage networks. Rainfall-related features were always significant. Event rainfall was placed second for Random Forest, XGBoost, and CatBoost models, whereas antecedent rainfall followed closely after that. It highlights the importance of taking both event rainfall and antecedent rainfall into consideration in flash-flood susceptibility models. Short-term rainfall acts as a triggering factor, while antecedent rainfall influences the wetness of the ground and runoff. Previous studies have advocated the use of

extreme rainfall for flash-flood susceptibility. Hydrological connectivity is one of the other top predictors. TWI, flow accumulation, drainage density, and distance to stream consistently emerge as important predictors. This set of predictors reflects the concentration of water, drainage patterns, and topography that make up a hydrological system. There was value in incorporating remote-sensing indicators into the predictive modeling[2]. The indicators of NDWI and NDVI performed well as strong predictors, but NDBI had less impact. These indicators represent the surface water, vegetation, and urban areas, which demonstrates the advantage of combining remote-sensing data with other information.

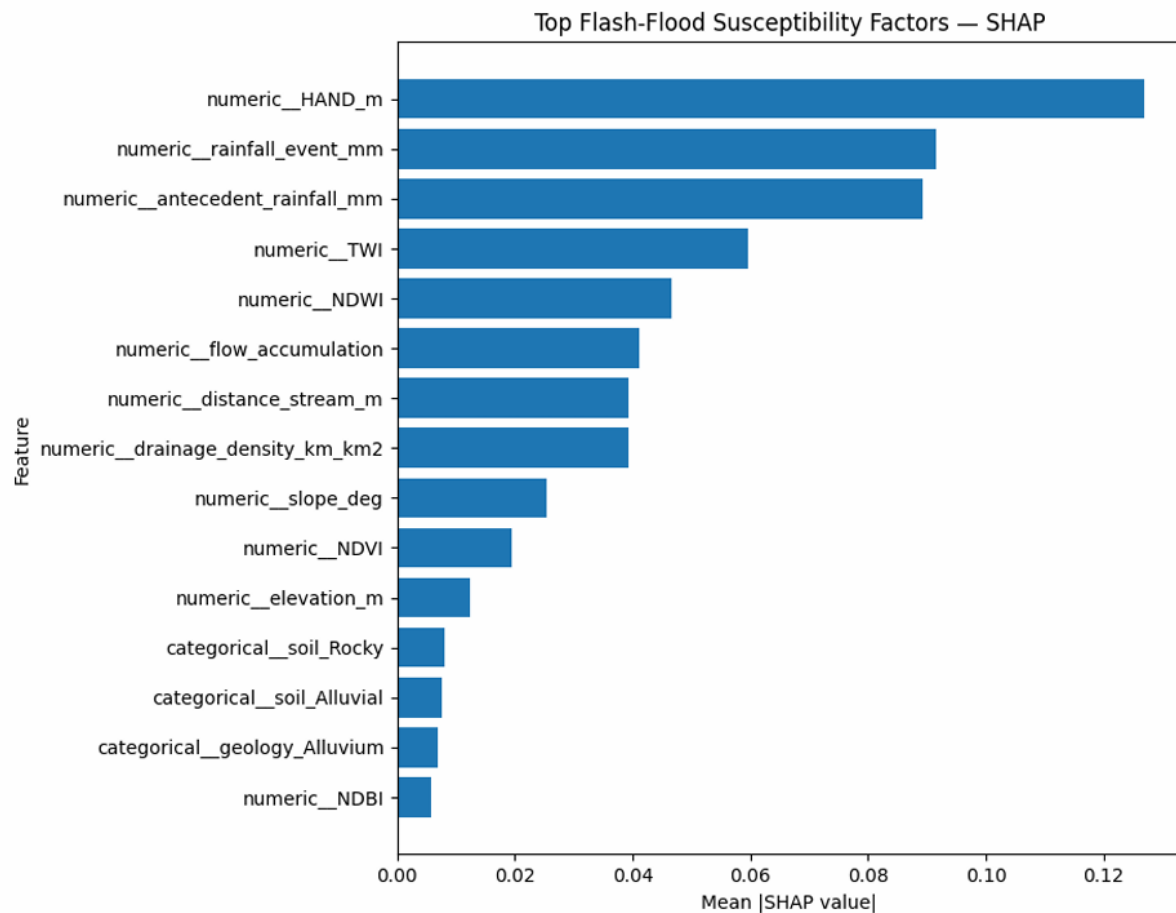


Figure 7 Top flash flood susceptibility factors

Feature importance should not be confused with causation. The average of variable values within the data set is not a universal hydrological relationship. Causal interpretation can only be obtained through signed SHAP values, dependence plots, field data, and hydrological/hydraulic evidence.

5.2 Comparison of Machine-Learning Models

Random Forest, XGBoost, and CatBoost have distinct learning methodologies and are mutually supportive models. Random Forest uses many trees to reduce variance, XGBoost is based on sequential boosted trees which reduce the errors, and CatBoost works efficiently on categorical data. However, all the three algorithms gave good results. Random Forest scored best with an accuracy of 98.66%, recall of 0.9857, F1 of 0.9869, MCC of 0.9732, ROC-AUC of 0.9993, and PR-AUC of 0.9993. CatBoost scored the highest precision at 99.03%. This proves that model selection must rely on empirical cross-validation rather than subjective assumptions of algorithmic superiority. In Pakistan, gradient boosting approaches have performed reliably for flood susceptibility [2, 3], while heterogeneous ensemble frameworks have shown particular efficacy across Balochistan [4]. CatBoost algorithms are also increasingly applied to flood-prone land classification [5, 6]. CatBoost algorithms are increasingly explored for mapping flood-prone areas [4]. In this paper, the Stacked Ensemble failed to outperform Random Forest in terms of performance indicators. An accuracy of 98.17% was achieved versus 98.66% of the latter. ROC-AUC of 0.9990 and PR-AUC of 0.9991 were attained, which can be considered high but still inferior to Random Forest. The strength of stacking methodology stems from its methodology: it provides a framework for integrating different model predictions without information leakage to the meta-learner. The positive values of coefficients for all three base models prove the contribution of each of them to the meta-model, although the ensemble of all three models was not the most efficient one.

5.3 Importance of Spatial Validation

Spatial validation is essential for environmental susceptibility modeling as nearby locations usually share similarities in terrain, precipitation, soil, geology, vegetation, and drainage characteristics. The use of random train-test splits may result in overly optimistic predictions as a result of spatial dependencies. For the current experiment, 25 blocks of space were used to separate the training and test samples, and the Group Fold created spatially separated out-of-fold predictions for controlled leakage stacking. While high model accuracy indicates effective flood vs non-flood classification, be cautious while interpreting the results because blocking does not eliminate spatial dependencies, as well as any uncertainty related to the labeling of floods.

5.4 Importance of Leakage-Safe Stacking

Leakage-controlled stacking is necessary because ordinary stacking may lead to information leakage since the meta-model is trained using predictions based on the in-sample data used to train the base models. For this, 25 blocks of spatial Group Fold was used to provide spatially clustered out-of-fold predictions so that every training instance gets predicted by the models not trained on it. These predictions acted as input to a logistic regression meta-learner, and the base learners were then trained on the whole training set. The positive meta-model coefficients indicate that all three models had a positive contribution, where the Random Forest's coefficient is 3.2498, the XGBoost model is 3.8928, and the CatBoost model is 3.9293. The fact that the stacked ensemble failed to outperform the Random Forest model shows that extra complexity is not always beneficial.

5.5 Explainability and Environmental Interpretation

It is not enough for engineers and disaster management to just have an accurate prediction. It is also important for decision makers to know the factors in the environment that affect the predictions of the models. All SHAP analyses

provided a similar interpretation for the three decision trees. The highest-ranking predictors were HAND, event rainfall, and antecedent rainfall. Other combinations included TWI, NDWI, flow accumulation, drainage density, distance to stream, slope, and NDVI. The consistency of their appearance in all three decision trees proves that the pattern is not algorithm-specific. HAND was the leading predictor in all three analyses. SHAP is a method of explaining models, not direct evidence for physical causation. A large SHAP score indicates that a feature had a strong impact on the prediction, but it does not necessarily indicate that the feature itself physically causes the flood. SHAP explanations must be taken into account together with hydrological theory, field data, flood records, and, when applicable, hydraulic models. The combination of these factors makes SHAP explanations more applicable to engineering practice.

5.6 Engineering Interpretation

The susceptibility map is a useful spatial screening tool when used in conjunction with information on settlements and infrastructure. With regard to flooding, areas of high susceptibility need to be the priority for field surveys and monitoring. This will serve as a warning that these places act as drainage ways as well as areas that need heightened observation during heavy rains. For roads, concentrate on those portions that pass through very vulnerable drainage ways. Regarding bridges, it is important to focus on the corridors that are prone to hydraulic and structural evaluation, including waterway adequacy, scour possibility, channel stability, and foundation adequacy. In urban planning, consider areas with high susceptibility during the assessment of infrastructure and expansion relative to natural drainage, concentration of runoff, and topographic constraints. In watershed management, consider these areas for further analysis and corrective actions like improving drainage, managing soil erosion, retaining and infiltrating water, and runoff management. The

susceptibility map is used to prioritize and screen, not as a substitute for hydrology, hydraulics, geotechnics, and engineering analysis[10].

5.7 Comparison with Previous Literature

Flood susceptibility assessment has evolved from using conventional Geographic Information System approaches to utilizing machine learning, remote sensing, and Ensemble Modeling. Earlier research in Pakistan utilized artificial intelligence and geospatial techniques for flood susceptibility and flood exposure. There were also studies on machine learning scalability. The application of heterogeneous ensemble models was found appropriate in Balochistan, while other Pakistani regions utilized GIS, AHP, remote sensing, and machine learning techniques for flood assessment. International research has made developments in rainfall-based ML [1], deep learning approaches [7], Sentinel-1 flood data using ML [8], ensemble techniques for semi-arid mountains [9], and GIS-ML approaches to analyze present and future vulnerability [10]. Recent research emphasizes the role of spatial regionalization and making model output into practical spatial information [11], particularly in mountains with complicated drainage patterns. This proposed framework builds on previous efforts through the inclusion of hydrology-specific features, different ML models, spatially clustered validation, stacking with leakage control, explainability using SHAP, and engineering evaluation using GIS. This is not a novel algorithm but rather a novel workflow for the entire process of modeling, spatial validation, ensemble creation, and explainability in Khuzdar District.

6. Limitations

The Flash Flood Vulnerability Assessment Model is dependent upon the quality of data. The efficacy of the model depends on the accuracy of the flood data because any wrong information will lead to labeling errors. It is common for rainfall measurements not to account for convective storms that are short-lived and occur locally within mountainous regions. Variables obtained from the

DEM are dependent on resolution, elevation accuracy, hydrologic conditioning, and pre-processing steps, which may exclude small streams and low-lying areas. Differences in resolution, classification, and dates of acquisition among the soil, geology, and land cover datasets pose further uncertainty. However, susceptibility does not equate to flood risk since variables such as the depth, velocity, duration, hydraulic intensity, vulnerability, exposure, and economic loss from floods were not measured. The model's accuracy will depend on the sampling method used, the spatial block structure, flood to non-flood ratio, choice of event, predictors, and hyperparameters; however, spatial validation minimizes but cannot eliminate these sources of uncertainty. While the empirical inventory incorporates confirmed regional flood and non-flood locations across Khuzdar District, uncertainties remain regarding satellite precipitation resolution during localized, short-duration convective storms.

7. Recommendations

Validate the Khuzdar flood hazard index through field investigations, starting with those areas having high and very high susceptibility indexes along with critical infrastructure such as roads, bridges, and settlements. For road culverts, examine their capability to drain under the conditions of design rainfall. Rainfall measurement could be enhanced by installing rain gauges at appropriate locations, particularly in mountainous watersheds, along with combining ground measurements with satellite rainfall estimates. A complete flood event database could be compiled based on field observations, photographs, and Sentinel-1 SAR and cloud-free Sentinel-2 satellite imagery. Future research work should cover flood depth, velocity, connectivity, and hydro-hydraulic simulation modeling to move beyond susceptibility mapping to flood risk assessment[1]. Independent event-based validation becomes necessary after sufficient documented events are available for assessing spatiotemporal transferability in addition to spatial generalization. Future research might focus on spatiotemporal

machine learning, uncertainty analysis, explainable AI, hydrological graph modeling, and IoT-based rainfall and river stage information. The susceptibility map needs to be turned into a fully functional GIS-based decision support tool that combines roads, bridges, settlement areas, and other layers of infrastructure, allowing for targeted interventions and the regular updating of risk data as flood information becomes available.

8. Conclusions

The use of an interpretable Hydro-Geo AI methodology is introduced for flash-flood susceptibility assessment in the Khuzdar District of Balochistan, Pakistan. The model uses the following variables as input: topography, hydrology, precipitation, remote sensing, LULC, soil, and geology, using Random Forest. One methodological advancement involves leakage-controlled stacking, where Group KFold predictions are used to train a logistic-regression meta-model rather than biased in-sample predictions. These findings reveal that there was excellent classification in all models. In terms of accuracy, Random Forest ranked first with an accuracy rate of 98.66%, F1 of 0.9869, MCC of 0.9732, ROC-AUC of 0.9993, and PR-AUC of 0.9993. Multi-model SHAP analysis consistently identified HAND, event rainfall, and antecedent rainfall as the dominant environmental predictors. This indicates that hydrological features and rainfall influence the susceptibility models. This continuous probability is represented by a susceptibility index that constitutes a GIS-generated surface. When coupled with roads, bridges, habitation, drainage, and other engineering infrastructure features, such results aid in engineering assessment, site investigation prioritization, flood planning, and watershed management. The empirical inventory compiled for Khuzdar District establishes a robust benchmark for spatial flash-flood prediction, demonstrating that the leakage-controlled Hydro-GeoAI framework successfully translates multi-source geospatial data into actionable spatial screening surfaces. The framework is scientifically

validated through empiricism; hence, it presents a scientifically sound foundation for flash flood vulnerability mapping and engineering decision-making in Khuzdar. It facilitates evidence-based flood preparation, infrastructure analysis, watershed management, and development planning, without compromising detailed hydrologic, hydraulic, geotechnical, and structural evaluations.

9. REFERENCES

- [1] M. Waleed and M. Sajjad, "High-resolution flood susceptibility mapping and exposure assessment in Pakistan: An integrated artificial intelligence, machine learning and geospatial framework," *International Journal of Disaster Risk Reduction*, vol. 121, Art. no. 105442, 2025, doi: 10.1016/j.ijdrr.2025.105442.
- [2] M. Tayyab, M. Hussain, J. Zhang, S. Ullah, Z. Tong, Z. U. Rahman, A. R. Al-Aizari, and B. Al-Shaibah, "Leveraging GIS-based AHP, remote sensing, and machine learning for susceptibility assessment of different flood types in Peshawar, Pakistan," *Journal of Environmental Management*, vol. 371, Art. no. 123094, 2024, doi: 10.1016/j.jenvman.2024.123094.
- [3] M. Waleed and M. Sajjad, "Advancing flood susceptibility prediction: A comparative assessment and scalability analysis of machine learning algorithms via artificial intelligence in high-risk regions of Pakistan," *Journal of Flood Risk Management*, vol. 18, no. 1, Art. no. e13047, 2025, doi: 10.1111/jfr3.13047.
- [4] M. A. Hussain, Z. Chen, B. Pradhan, S. R. Meena, et al., "Hybrid heterogeneous ensemble learning framework for flood susceptibility mapping in Balochistan, Pakistan," *Journal of Hydrology: Regional Studies*, vol. 61, Art. no. 102718, 2025, doi: 10.1016/j.ejrh.2025.102718.
- [5] S. V. Razavi-Termeh, A. Pourzangbar, A. Sadeghi-Niaraki, M. J. Franca, and S.-M. Choi, "Metaheuristic-driven enhancement of categorical boosting algorithm for flood-prone areas mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 136, Art. no. 104357, 2025, doi: 10.1016/j.jag.2025.104357.
- [6] S. V. Razavi-Termeh, A. Sadeghi-Niaraki, S. I. Abba, J. Hussain, and S.-M. Choi, "Flood-prone area mapping using a synergistic approach with swarm intelligence and gradient boosting algorithms," *Scientific Reports*, vol. 15, Art. no. 27924, 2025, doi: 10.1038/s41598-025-12022-6.
- [7] M. E. Chowdhury, A. K. M. S. Islam, R. U. Zaman, and S. Khadem, "A machine learning based approach for flash flood susceptibility mapping considering rainfall extremes in the northeast region of Bangladesh," *Advances in Space Research*, vol. 75, no. 2, pp. 1990–2017, 2025, doi: 10.1016/j.asr.2024.10.047.
- [8] P. V. Hoa, N. A. Binh, P. V. Hong, et al., "One-dimensional deep learning driven geospatial analysis for flash flood susceptibility mapping: A case study in North Central Vietnam," *Earth Science Informatics*, vol. 17, pp. 4419–4440, 2024, doi: 10.1007/s12145-024-01285-8.
- [9] R. Al-Ruzouq, A. Shanableh, R. Jena, M. B. A. Gibril, N. A. Hammouri, and F. Lamghari, "Flood susceptibility mapping using a novel integration of multi-temporal Sentinel-1 data and eXtreme deep learning model," *Geoscience Frontiers*, vol. 15, no. 3, Art. no. 101780, 2024, doi: 10.1016/j.gsf.2024.101780.
- [10] X. Zhu, H. Guo, and J. J. Huang, "Urban flood susceptibility mapping using remote sensing, social sensing and an ensemble machine learning model," *Sustainable Cities and Society*, vol. 108, Art. no. 105508, 2024, doi: 10.1016/j.scs.2024.105508.

- [11] Y. Bammou, B. Benzougagh, B. Igmoullan, A. Ouallali, S. Kader, V. Spalevic, P. Sestras, P. Billi, et al., "Optimizing flood susceptibility assessment in semi-arid regions using ensemble algorithms: A case study of Moroccan High Atlas," *Natural Hazards*, vol. 120, pp. 7787–7816, 2024, doi: 10.1007/s11069-024-06550-z.
- [12] A. Amiri, K. Soltani, I. Ebtehaj, and H. Bonakdari, "A novel machine learning tool for current and future flood susceptibility mapping by integrating remote sensing and geographic information systems," *Journal of Hydrology*, vol. 632, Art. no. 130936, 2024, doi: 10.1016/j.jhydrol.2024.130936.
- [13] X. Ke, N. Wang, T. Li, Z. Liu, Z. Li, G. Zuo, and Y. Chen, "From prediction to regionalization: Enhancing flash flood susceptibility mapping using machine learning and GeoDetector," *Geoscience Frontiers*, vol. 17, no. 1, Art. no. 102213, 2026, doi: 10.1016/j.gsf.2025.102213.
- [13] N. Ahmed, "A review of existing Machine Translation Approaches, their Challenges and Evaluation Metrics," *Pakistan Journal of Engineering, Technology & Science*, vol. 11, no. 1, pp. 29–44, Dec. 2023, doi: 10.22555/pjets.v11i1.1002.
- [14] N. Ahmed, R. Ameer, A. Khan, and S. Gull, "Data Mining in Healthcare: An Overview of Applications, Techniques, and Challenges," 2024.
- [15] S. Gull, N. Ahmed, and S. Ahmed Sheikh, "Liberal Journal of Language & Literature Review Liberal Journal of Language & Literature Review Liberal Journal of Language & Literature Review Deep Learning-Driven OCR System for Brahui Printed Text: Bridging the Digital Gap in Low-Resource Language Processing Liberal Journal of Language & Literature Review", [Online]. Available: <https://llrjournal.com/index.php/11>
- [16] Zubair Ahmed, S. Naqvi, Nooruddin, M. A. Khuhro, Naseer Ahmed, and A. Khan, "A Comparison of Different Defuzzification Methods on Road Traffic Accidents of Data of Khuzdar Region of Balochistan Province of Pakistan," *Pakistan Journal of Engineering, Technology and Science*, vol. 12, no. 2, pp. 1–13, Dec. 2024, doi: 10.22555/pjets.v12i2.1127.
- [17] Naseer Ahmed, M. A. Khuhro, and A. A. Laghari, "Bridging the Gap: A Transformer-Based Chatbot Architecture for Low-Resource Brahui via Embedding Fusion and Transfer Learning," *Mehran University Research Journal of Engineering and Technology*, vol. 45, no. 3, pp. 114–123, Jul. 2026, doi: 10.22581/muet1982.0675.
- [18] G. Nabi, A. Mengal, and N. Ahmed, "USING FUZZY LOGIC IN SOLAR POND AND PARABOLIC TROUGH COLLECTOR TECHNOLOGIES FOR POWER GENERATION PREDICTION: A CASE STUDY OF KHUZDAR REGION," *International Journal of Social Sciences Bulletin*, vol. 2, no. 4, p. 2024, 2024, [Online]. Available: <https://ijssb.org>
- [19] N. Ahmed, M. Ahmed Khuhro, M. Ali Dootio, and S. Pakistan, "BUILDING AND ANALYZING A BRAHUI TEXT CORPUS: APPLYING DTM AND TF-IDF TECHNIQUES," Jan. 2024. Accessed: Mar. 05, 2026. [Online]. Available: https://www.researchgate.net/profile/Naseer-Ahmed-44/publication/384143614_building_and_analyzing_a_brahui_text_corpus_applying_dtm_and_tf-idf_techniques
- [20] N. Ahmed, M. A. Khuro, A. Khan, M. Dawood, M. A. Dootio, and N. U. Jan, "Student textual feedback sentiment analysis using machine learning techniques to improve the quality of education," *Pakistan Journal of Engineering, Technology & Science*, vol. 11, no. 2, pp. 32–40, Dec. 2023, doi: 10.22555/pjets.v11i2.1039.