EXPLORING THE ECONOMIC DAMAGES THROUGH FLOOD PREDICTION AND SPATIAL ANALYSIS: AN APPLICATION OF HYBRID BAGGING-BOOSTING DECISION TREES ENSEMBLE

By

Javeria Sarwar
PIDE2018FPHDETS-05

Supervisor
Dr. Saud Ahmed Khan

Co-Supervisor
Dr. Muhammad Azmat

School of Economics
Pakistan Institute of Development Economics
Islamabad

Author's Declaration

I Ms. Javeria Sarwar hereby state that my PhD thesis titled **"Exploring the Economic Damages through Flood Prediction and Spatial Analysis: An Application of Hybrid Bagging-Boosting Decision Trees Ensemble"** is my own work and has not been submitted previously by me for taking any degree from **Pakistan Institute of Development Economics, Islamabad'** or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduation the university has the right to withdraw my PhD degree.

Ms. Javeria Sarwar

PIDE2018FPHDETS05

Plagiarism Undertaking

I solemnly declare that research work presented in the thesis titled **"Exploring the Economic Damages through Flood Prediction and Spatial Analysis: An Application of Hybrid Bagging-Boosting Decision Trees Ensemble"** is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the **HEC** and **Pakistan Institute of Development Economics, Islamabad** towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD degree, the University reserves the rights to withdraw/revoke my PhD degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted plagiarized thesis.

Students/Author Signature:_____

Ms. Javeria Sarwar

PIDE2018FPHDETS05

Certificate of Approval

This is to certify that the research work presented in this thesis, entitled: **"Exploring the Economic Damages through Flood Prediction and Spatial Analysis: An Application of Hybrid Bagging-Boosting Decision Trees Ensemble"** was conducted by **Ms. Javeria Sarwar** under the supervision of **Dr. Saud Ahmed Khan and Dr. Muhammad Azmat.** No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Econometrics from **Pakistan Institute of Development Economics, Islamabad.**

Student Name: Ms. Javeria Sarwar
PIDE2018FPHDETS05

Signature: _____

Examination Committee:

a) **External Examiner: Dr. Iftikhar Hussain Adil**
Associate Professor
NUST, Islamabad

Signature: _____

b) **Internal Examiner: Dr. Zahid Asghar**
Professor,
Quaid-e-Azam University, Islamabad

Signature: _____

Supervisor: Dr. Saud Ahmed Khan
Assistant Professor,
PIDE, Islamabad

Signature: _____

Co-Supervisor: Dr. Muhammad Azmat
Associate Professor,
NUST, Islamabad

Signature: _____

Dr. Iftikhar Ahmad
Head, PIDE School of Economics (PSE)
PIDE, Islamabad

Signature: _____

Acknowledgments

I would like to express my deepest gratitude to Almighty Allah, the most gracious and merciful, Who made me able to accomplish this dissertation.

First and foremost, I would like to express my profound gratitude to my supervisors, Dr. Saud Ahmed Khan and Dr. Muhammad Azmat for encouraging this research by suggesting this research gap. Their support and encouragement throughout the research process was outstanding. From the earliest stages of formulating my ideas to the final stages of writing this thesis, their feedback, willingness to share knowledge, patience, mentorship and insightful discussions have pushed me to grow as a scholar. I deeply appreciate their availability, even during busy periods, and their assistance when I encountered obstacles.

I am deeply grateful for the unwavering love, support and prayers of my parents, who are a constant source of strength throughout my life. Their faith in my abilities and unconditional love have been the foundation of my journey. Their wisdom, encouragement, and sacrifices have guided me throughout the challenging and exciting journey. Their sacrifices, often made silently and without expectation of acknowledgment, have not gone unnoticed. I am eternally grateful for the countless ways you have supported me – emotionally, intellectually, and financially.

Abstract

A huge national budget is required for flood damage reduction projects; thus, it must be ensured that the public money utilized therein is spent effectively and efficiently. In this context, reliable flood damage assessment is pertinent to analyze the economic aspects of projects related to flood damages. Riverine floods cause significant damage to assets and adversely affect the economies. The research aims to propose a feature selection model for hydraulic analysis as such a model has not been proposed previously. For this purpose, hybrids of three metaheuristic algorithms, Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Genetic Algorithm (GA) with two machine learning models which are Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are employed. The dataset considered was hydraulic having an association with flood and possessed topographic, geo-environmental, and human-induced variables. The dataset considered had multicollinearity heteroscedasticity and autocorrelation problems. The metaheuristic algorithms were evaluated by varying the number of population size. Among them, PSO performed better by providing an appropriate number of features with a lower number of iterations. We have analyzed the performance of SVM with different kernels; linear, radial basis function (RBF), sigmoid, and polynomial, as the original SVM is designed only for linear datasets but the hydraulic dataset possesses non-linear characteristics as well. The performance of different kernels in terms of their accuracies is evaluated and recorded. This study showed that RBF performed the best and sigmoid showed the least accuracy for GA, PSO, and ACO algorithms. The performance of KNN is evaluated in terms of accuracies by varying the K-values. It was found that KNN shows low accuracy with a small K-value which then attained a maximum level by increasing K-values and it finally started decreasing, explicitly, by further enhancing K-values. Further, the research proposes a novel ensemble machine-learning model for flood susceptibility mapping. The ensemble model integrates four independent machine learning models namely, Random Forest (RF), Logistic Model Tree (LMT), Naïve Bayes Tree (NBT), and Reduced Error Pruning Tree (REPT). For susceptibility mapping, a spatial database is prepared by considering 5500 flood points and 5500 non-flood points. The 14 flood conditioning factors (selected by PSO algorithm) considered for the research possess topographic, geo-environmental, and human-induced variables. The dataset has been randomly divided into sample sizes of 70% and 30% for training and validation the models, respectively. The performance of the ensemble model is evaluated by utilizing various statistical techniques and is compared with the stand-alone models. The results revealed that the hybrid bag-boost ensemble model (RF-LMT-NBT-

REPT) performed the best with a 99.5% accuracy level for the training sample and 98.9% for the validation sample. The inundation maps hence acquired by utilizing the hybrid bag-boost ensemble model for the years 2010 and 2022 and the predicted flood of 2032 are used to provide district-level flood damages in the lower Indus basin. Moreover, the present research illustrates a complete land use land cover (LULC) transition analysis for the study area between the time period of 2010 to 2022 and illustrates the spatial association between flood and LULC transition by employing geographical weighted regression analysis. In this context, the regional heterogeneities have been considered and a complete district-level analysis for LULC change has been provided by considering each land cover transition from 2010 to 2022. Furthermore, the proposed simulated hybrid bag-boost ensemble model is employed for the calculation of flood depth and extent in the lower Indus basin for the assessment of associated economic damages. The research provides a district-level loss due to the 2022 flood and the forecasted 2032 flood. For this purpose, the study considers Land Use Land Cover and administrative boundary maps of the study area. The monetary values of the assets have been obtained from the concerned administrative departments and are utilized for the damage assessment of the floods at district level. The proposed ensemble model and the flood conditioning factors can be utilized for flood potential assessment in future studies. Moreover, this technique will assist in decision-making while evaluating the economic feasibility of flood damage reduction projects.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

Climatic changes have enhanced the frequency of floods on global level. Floods are destructive in nature and can lead towards immense human and monetary losses (Hardy et al., 2016; De Walque et al., 2017). Flood disaster can alter human settlements, agriculture, infrastructure and economic assets and natural resources, causing spatial alterations (Azareh et al., 2019). They account for almost 31% of the total international economic damages caused due to natural calamities (Dano et al., 20191Q

Floods are caused either due to Climatic alterations or are human induced (Chang and Chen, 2016). Thus, spatial analysis of flood damages is pertinent to reduce the associated destructions (Shehata and Mizunaga, 2018). Knowledge regarding flood locations is essential for devising strategies for reducing the flood associated risks (Yaseen et al., 2022). Mitigation of flood risk is a long-term challenge and is also a key issue when devising policies for different spatial locations (Naulin et al., 2013). Hence, flood risk management is a basic mitigation step that deals with mapping flood risks and disasters (Pravalie and Costache, 2013).

Geographic Information System (GIS), Remote Sensing (RS), and Geoinformation provide helpful tools to evaluate the flood potential and extent (Khosravi et al. 2016). The subsequent improvement in remote sensing has assisted the researchers by providing them satellite controlled images with various spatial resolutions, for all parts of the globe. Landsat sensors possess enough spatial resolutions which are ample to characterize the factors that may affect the land usage categories such as agriculture, residential, infrastructure, etc. The task of GIS is to process the remote sensing dataset and data attained from flood inventories and flood predictors, to create vector or raster inputs to be utilized in econometric and statistical modelling.

Contemporarily, more complex data-driven methods are being utilized for flash-flood susceptibility and vulnerability assessments as they are comparatively more robust and can assess and evaluate complex relationships between the input variables. In this regard, various Machine Learning tools such as Artificial Neural Networks (ANN) (Youssef et al., 2011), Adaptive Neuro-Fuzzy Inference Systems (ANFIS) (Termeh al., 2018), Decision Trees (DT) (Lee et al., 2017), or Support Vector Machines (SVM) (Tehrany et al., 2015). Moreover,

recently, hybrid or ensemble machine learning (ML) models are being used for obtaining accurate flood vulnerability mapping (Ahmadlou et al. 2019).

Various models have been proposed in literature, but no unified consent is yet established on any flood susceptibility model. Moreover, every model contains some limitations. Therefore, it is pertinent to formulate hybrid Machine Learning models. Hybrid Machine Learning models or ensembles are comparatively accurate and perform better than conventional models in literature. Hybrid methods are utilized in literature for modelling floods. Razavi Termeh et al. (2018) utilized bagging strategy for Adaptive Neuro-Fuzzy Interference System and its optimization algorithm. Chapi et al. (2017) employed bagging ensemble strategy on Logistic Model Tree, which worked better in comparison to the models compared. Likewise, Ngo et al. (2019) formulated novel hybrid approach, FA-LM-ANN, by integration of Firefly Algorithm (FA), Levenberg Marquardt (LM) and Artificial Neural Network (ANN) for flood susceptibility mapping, which appeared to be good fit model than the benchmark models.

In the context of Pakistan, flood susceptibility analysis has been conducted by a few recent studies but none of them focused on the Indus basin. Moreover, they have utilized conventional tools to assess susceptibility except one recent study which has utilized machine learning models. The research conducted by Hassan et al. (2021), assessed the susceptibility mapping in Chitral by comparing the efficiency of Analytic Hierarchy Process (AHP) model and Frequency Ratio (FR) model. Likewise, Awais et al. (2022) assessed susceptibility in Northern Sindh and Southern Punjab regions through FR model. Similarly, Munazza et al. (2022) and M. Farhan et al. (2018) have conducted studies to assess susceptibility in Panjkhora Valley and Charsadda, respectively, through FR approach. M. Farhan et al. (2020) conducted another study in flood plain of river Swat by employing Bivariate Model and highlighted the flood prone areas. Yaseen et al. (2022), proposed a novel ensemble of LR, MLP and SVM to assess the susceptibility in Karachi by utilizing the flood points of 2022 floods. It is pertinent to mention here that to the best of author's knowledge, flood susceptibility maps and flood prone areas have not been utilized in Pakistan to quantify the economic damages by overlaying official datasets over the flood prone areas. In this regard, the current study will contribute to literature.

The National Disaster Management Authority (NDMA) is currently working on the post-disaster evacuation and valuation of the damages after the occurrence of flood. Currently, to

the best of our knowledge, there is no mechanism developed by any of the existent bodies at national or provincial level which deals with the pre-flood prediction and the areas which may be most prone to the flood. In this context, Pakistan Metrological Department (PMD), deals with the rainfall predictions but there is no mechanism in place to evaluate the factors which are specifically responsible for occurrence of floods.

The current research, aims to predict the floods based on the geographical, hydrological, and environmental factors, obtained from remote sensing data and official documents, which will be overlayed on the geo-referenced maps of the study area. It is pertinent to mention here that, the disaster management authorities only analyze the post flood data to estimate damages using the standard formula[1] developed by international organizations. In this regard, Ministry of Planning, Development and Special Initiatives' conducted an assessment[2] on 2022 flood. The analysis was conducted using geospatial data layers, along with crowd-sourced and open-source mapping platforms. To ensure data accuracy and align the humanitarian response with the recovery process, the government and international partners organized Stakeholder Engagement Meetings for data validation. The Post-Disaster Needs Assessment (PDNA) follows the methodology jointly developed by the European Union, the World Bank Group (WBG), and the United Nations. The PDNA specifically addresses the impacts and the associated recovery and reconstruction needs resulting from the floods. *Damage*[3], loss, and needs assessments were based on a *pre-flood baseline*[4]. Data collection took place from September to mid-October 2022. Thus, in connection to the quoted material, the existent national agencies use remote sensing data either to assess the post flood scenarios or to gather information on the pre-flood baseline conditions (without any reference to the geographical factors contributing to floods).

Flood damages assessment can provide valuable contribution in the long-term land use planning (Neubert et al., 2016). Flood based predictions and associated economic plannings

[1] An asset is considered Partially Damaged if less than 40% of it is affected, the structural integrity remains intact, and the repair cost is under 40% of the asset's total value. Conversely, it is classified as Completely Destroyed if more than 40% of the asset is damaged or if the cost of replacement exceeds 40% of its total value.
[2] https://www.undp.org/pakistan/publications/pakistan-floods-2022-post-disaster-needs-assessment-pdna
[3] In the referenced report, damage refers to the direct costs associated with the destruction or impairment of physical assets. These costs are expressed in monetary terms and are estimated based on the expense of repairing or replacing the affected assets and infrastructure, using the pre-crisis replacement values as a basis.
[4] The number of lost assets is determined by calculating the difference between pre-flood and post-flood asset inventories.

are pertinent for economic hubs like the Indus River basin. The Indus River flows by the largest economic centers of Pakistan. Thus, there is a need to integrate present and future scenarios to identify the potential flood damages so that the present level of flood risk induced by climate change can be reduced. Therefore, the current research aims to assess and evaluate flood susceptibility, spatial analysis and quantification of economic flood damages.

1.1. Problem Identification and Background

In Pakistan, flooding is a common phenomenon as there does not exist any mechanism to predict floods and their damage assessment using advanced statistical and economic methods. This methodological gap makes urban and rural planning difficult. In this context, various proposals had been made, previously, in Pakistan by federal and provincial agencies for prediction of floods and assessment of the associated damages. Some of the major proposals are; in 2016, the National Disaster Management Authority (NDMA) issued some national policy guidelines for executing Multi-Hazard Vulnerability and Risk Assessment[5] (MHVRA). The report sets out procedures to be followed for development of National Risk Picture. In this regard, up-till now, MHVRA has been executed for five regions namely; Bahawalpur, Khushab, Rahim Yar Khan, Jhang and Multan. It assesses the risk by using a risk index. The Risk Index consists of indicators to cover physical, economic, demographic, social, environmental and economic dimensions of risk. Specific weights have been assigned to each indicator in order to calculate its impact on risk. The risk formula used by MHVRA is; Risk = Hazard*Vulnerability*Exposure/Capacity. Moreover, flood vulnerability is assessed by using Return Period formula. In 2017, NDMA issued guidelines for Multi-Sector Initial Rapid Assessment[6] (MIRA) to assess the areas affected by disasters within 24-72 hours of onset disaster for conducting relief and recovery measures.

The post disaster report[7] on 2022 floods, issued by Ministry of Planning Development and Special Initiatives, highlighted the limitation that their research team had to rely on self-reported net asset valuation before and after flood occurance. It depicts that there does not exist

[5] https://cms.ndma.gov.pk/storage/app/public/publications/October2020/G0wGoQX2IvEqb4Om54kG.pdf

[6] https://cms.ndma.gov.pk/storage/app/public/publications/October2020/zmd6m0dcelvarC2ZtgMX.pdf

[7] https://www.undp.org/pakistan/publications/pakistan-floods-2022-post-disaster-needs-assessment-pdna

any flood prediction and economic damages[8] assessment mechanism based on future simulations.

The NDMA issued a report (Sim Ex 1)[9] on March, 2023 which is based on floods and rains emergencies. It highlights the lessons learnt from 2022 floods and emphasized the urgency and need for forecasting future flood disasters for an effective and efficient response in future emergencies. The National Monsoon Contingency Plan[10], proposed by NDMA on 23rd June, 2023, has a future plan to forecast floods and their likely impacts on different regions through simulation based studies. The mechanism to forecast floods and quantification of economic damages by using simulation-based study will contribute greatly to the current procedures followed by NDMA as this practice is currently not being done by any agency in Pakistan. This depicts that, in Pakistan, there does not exist any methodological framework for flood prediction and associated damages evaluation. The current study aims to fill this gap by providing a model that can be used in any region for selecting the most important contributing factors in floods and has provided the most pertinent factors that can cause floods in the lower Indus basin in future. These flood contributing factors can be utilized for predicting floods, hazard and susceptibility assessment and calculation of future economic damages.

Hence, the 2022 floods highlight that there is an adaptation challenge to reduce the risk of future floods in Pakistan through simulation-based predictions of flood prone areas and quantification of future economic damages.

1.2. Statement of the Problem

In Pakistan, there is a need for conducting predictive studies for the areas prone to floods and its associated damages that may be caused due to future flood, consequently huge loss of life and properties such as agriculture, houses, livestock etc. To the best of my knowledge, the mechanism of flood prediction, identification of flood susceptible areas through simulation using geographic, environmental and economic factors, the spatial correlation between flood and land use land cover (LULC) to assess vulnerability and the quantification of asset damage

[8] In the report, damage is described as the direct cost associated with destroyed or damaged physical assets. These costs are expressed in monetary terms and are calculated based on the expense of repairing or replacing the assets and infrastructure, using the replacement prices that existed prior to the crisis.
[9] http://cms.ndma.gov.pk/storage/app/public/publications/May2023/4yqNNX7XRsnltHgZcOH0.pdf
[10] https://cms.ndma.gov.pk/storage/app/public/plans/July2023/Ad4JllgX4nLsyOGb9zjH.pdf

caused by flood to different land covers through temporal analysis (simulation based), do not exist in Pakistan.

An already dwindling economy like Pakistan cannot bear the unplanned losses due to natural disasters, it hampers the production capacity and hence increases the unemployment in the affected areas. Such frequent incidences may result in a slow down of economic growth. A disaster can hinder the economy in two ways; the development expenditure is spent on rehabilitation expenses instead of development and the local economic activities are also adversely affected and may not contribute in national income until full recovery has been obtained.

Exploitation of area specific knowledge of factors causing recurrent episodes of disasters may help predict the intensity and possible damage level before next episode. This analysis can be utilized to mitigate the consequences of unforeseen disaster. Subsequently, there is a need to have some prior information to mitigate the economic damages caused by natural disasters like riverine floods. Two types of methodologies are commonly used for riverine flood analysis; one is traditional methodologies such as inferential and regression analysis and the other one is advanced set of tools like machine learning and deep learning tools. As the underlying dataset for flood analysis is so complicated and contains a huge set of covariates, thus the machine learning algorithms, have been used in the thesis, for handling such data efficiently. It is fact that advanced tools application leads to more robust conclusion and policy implications in contrast to traditional methodologies. Referring to the econometric techniques, different machine learning models will be utilized and compared which can be used for geospatial dataset and spatial analysis. The spatial dataset includes topographic, human induced and geo-environmental factors specific to each region in the Sindh province. Moreover, the dataset is heterogenous in nature and is dependent on the geo-referenced locations. Furthermore, spatial analysis is pertinent as the current research takes flood as an endogenous factor causing economic damages rather than exogenous factor.

Since different models perform variably under specific conditions and each has its own limitations, there is still no universally accepted model for flood modelling. In Pakistan, previous studies conducted for assessing flood susceptibility mainly focused on conventional methods such as AHP, FR and Bivariate methods (Hassan et al., 2021; Awais et al., 2021; Munazza et al., 2022; Farhan et al., 2020). Hence, it is essential to tackle these challenges by developing hybrid or ensemble models that combine various machine learning approaches.

Numerous previous studies have demonstrated that ensemble or hybrid models often outperform traditional methods, offering higher accuracy and improved performance (Pham et al., 2017; Saha et al., 2021). A single algorithm, nevertheless, has some econometric limitations that may affect true predictions of floods (Yaseen et al., 2022). Moreover, the input data often lacks suitable representation, which may lead a model to miss its best-fit function. A single model either resolves variance or bias problem. A model resolving both problems is pertinent in this regard. Hence, ensemble modeling using bagging-boosting technique is an efficient way to resolve these issues simultaneously. To bridge this research gap, the research aims to present novel ensemble using machine learning models for flood predictions.

As far as the quantification of damages is concerned, the national agencies working in Pakistan quantify damages by calculating a difference between post-disaster and pre-disaster lost assets and properties as depicted in the report of Ministry of Planning Development and Special Initiatives, 2022. The theoretical problem involved in estimation of the economic outcomes of a disaster is the extreme data limitation which makes the actual estimates far different from the hypothetical true disaster costs. The data available on disaster impacts, generally, undertakes those factors which are easily observable ex-post. Hence, most disaster loss data, usually, underestimates the full economic impact of disaster. Hence, various studies use macroeconomic variables as proxy for the direct and indirect impacts of disasters. Another challenge is that it is easy to double-count losses (Cochrane, 2004). Thus, there is a need for defining a mechanism to quantify economic damages for the purpose of accuracy, cost effectiveness and efficiency. Thus, the research aims to fill this gap by providing a mechanism to quantify future flood damages by utilizing georeferenced area maps of flood susceptible locations and overlaying the LULC official data on these maps to assess losses. This will ascertain the damages which can be incurred in future based on today's scenarios. For the purpose of cost analysis, the study will utilize the cost of assets data available at local municipalities and aims to provide cost damages analysis at union council level.

1.3. Research Gap

The study aims to bridge the following research gaps;

1) To the best of my knowledge, none of the studies have utilized hybrid metaheuristic algorithms and machine learning models for feature selection by utilizing the geospatial dataset.

2) To the best of my knowledge, the chosen machine learning models and their ensemble have not been utilized yet for extraction of flood distribution, spatial susceptibility, and prediction under different temporal scenarios (past, present, future).
3) To the best of my knowledge, none of the studies have ascertained the spatial correlation between LULC and floods in the areas which are more susceptible to floods using machine learning algorithms and Geographically Weighted Regression (GWR), simultaneously.
4) To the best of my knowledge, none of the researches conducted over the study area, have utilized geo-informatics for quantification of economic flood damages.

1.4.Research Questions

The study aims to answer the following research questions;

1) As the geo-spatial variables are heterogenous in nature, so, they may violate the Independent and Identically Distributed (IID) assumption. Thus, what is the best subset of features, that is free from econometric issues of multicollinearity, heteroscedasticity and autocorrelation, among the geo-environmental, topographic and human-induced factors that can contribute to floods?

2) As the spatial variables are heterogenous in nature, so, they may encounter the problem of high variance and bias which may adversely affect the model performance and model's prediction efficiency. Do hybrid and ensemble models improve the model performance? What is the district-level flood risk delineation in the lower Indus basin?

3) Transition in land use land cover (LULC) of a region over a span of years leads to riverine floods. To what extent this correlation exists among LULC and flood in the lower Indus basin at district level? What is the spatial magnitude of flood due to LULC transformation at district level?

4) Quantitative flood damages include the monetary and physical evaluation of all the tangible assets that are directly affected by the occurrence of floods. In this regard, what is the impact of 2022 flood and the predicted 2032 flood on various physical assets in the lower Indus basin?

1.5. Objectives of the Research

The research has following objectives.

1) To predict the spatial susceptibility of floods in the study region under variant temporal scenarios through integration of geoinformatics and machine learning models. This objective

will also specify the contribution/sensitivity of each variable to the floods, spatially. It is further divided into two parts;

i. Hybrid simulations of metaheuristic algorithms (PSO, GA, ACO) and classifier models (SVM, KNN) simultaneously to find an optimal subset of data set free from spatial econometric problems.

ii. To map the floods distribution, spatial susceptibility and prediction using different temporal scenarios by utilizing ensemble base classifier (decision trees) simulations (bagging-boosting method) and stand-alone base classifiers (for comparison).

2) To integrate machine learning ensemble and geo-informatics to assess the correlation (by using Geographical Weighted Regression) between LULC and floods in the areas which are more susceptible to floods.

3) To quantify asset damages which may be caused by flood extent and intensity in the flood susceptible areas for past and future. The flood inundation map acquired by utilizing the ensemble model is employed for the calculation of flood depth and extent in the lower Indus basin for the assessment of associated economic damages to the assets.

1.6. Significance of the Research

The current procedures followed by national organizations are based on conventional methods of risk evaluation and vulnerability calculation. Recently, the Sim Ex 1 and Contingency Plan conducted by NDMA highlights the need of simulation based future predictions of floods. To the best of my knowledge, no prior research has been conducted on the selected study area for flood forecasting using machine learning models. Additionally, there has been no evaluation of the conditioning or influencing factors contributing to flooding in this region, even though it experiences floods nearly every year or every other year.

The majority of literature based on multi-country or within country analysis regresses macroeconomic variables on measures of disasters, their occurrence, associated damages, or number of fatalities. In this way, the economic impact of disaster is measured. A major chunk of papers do not discuss the physical aspects of the disaster and take disaster as an exogenous variable. For instance, Raddatz (2007), utilized panel vector autoregression model and assumed flood occurrence as exogenous, Noy (2009) used a panel of 109 countries and regressed GDP growth on standardized measures of disaster and a set of controls variables. He also examined endogeneity by examining disaster occurrence by utilizing physical measures such as Richter scale of a disaster. Likewise, Hochrainer (2009) developed a counterfactual projection of the

GDP and compared it to the actual post-disaster GDP. He employed an Autoregressive Integrated Moving Average Model to forecast GDP in the hypothetical no-disaster scenario, Similarly, Cuñado and Ferreira (2011), also analyzed the economic impacts of flood by assuming its occurrence as exogenous and used Vector Autoregressions with country fixed effects. Furthermore, Noy and Vu (2010) took output growth as response variable and included its lag as independent variable. They employed generalized method of moments estimator for dynamic panels. Anttila-Hughes and Hsiang (2011) analyzed disaster by using household survey data and took disaster as exogenous. They used a difference-in-differences approach with provinces and year as fixed effects. Strobl (2011) discussed the impacts of hurricanes on county growth rates in the USA and developed a hurricane destruction index from monetary loss, wind speed, and exposure variables to use as an explanatory variable in county Fixed-effects Model with a Spatial Autoregressive error term. Likewise, Deryugina (2013) also analyzed the impacts of hurricanes and used a propensity score matching to highlight the control group of counties with equivalent risk of hurricane and used a difference-indifferences approach. Loayza et al. (2012) examined the trend in post-disaster GDP growth with data of 84 countries by dynamic panel data model. Jaramillo (2009) investigated the long-term impacts of disasters. He estimated a Solow-style structural model, with cumulative measures of impacts of disasters as a variable to capture the influence of disasters on a country's steady-state growth rate. Furthermore, there may be an omitted variable bias problem as some measures that are not included in the model may, therefore, influence the economic outcomes and may be correlated with the disaster measures (Carolyn Kousky, 2014).

Therefore, this thesis has the potential to present effective models for predicting and identifying flood-prone areas and associated asset damages by treating floods as an endogenous factor that could occur in the future. Currently, such an approach is not implemented at the national or provincial levels. Since field-based surveys are both expensive and time-consuming, this research aims to generate high-resolution maps that can serve as valuable tools for management, policymakers, government agencies, and other relevant stakeholders. Hence, the current research is novel as;

1) It will analyze why some areas are currently more prone to floods than others. The reasoning will be provided on the basis of the specific geographic conditions. It will predict the specific areas which are more prone to floods in future. It will highlight, the specific and the

most important factors which may contribute in future floods with respect to the geographic locations.

2) It will assess the correlation between LULC and floods in the areas which are more susceptible to floods. This will highlight the land use category which may be damaged due to future floods in each flood prone area.

3) The assessment of the extent and intensity of economic damages based on proposed floods depths in each district of Sindh and calculation of respective costs.

1.7. Organization of the Thesis

This thesis is comprised of ten chapters. Chapter 1 provides a brief introduction to the research along-with, research gap, problem statement, research questions, research objectives, and significance of the study. Chapter 2 illustrates in detail about the related researches and the prevalent literature on floods and geospatial analysis and the utilized models in this research. Chapter 3 describes the data utilized and the study area along-with the related literature. Chapter 4 demonstrates the methodology employed. Chapter 5 is based on the analysis of the results of the feature selection modelling using meta-heuristics and machine learning models, followed by chapter 6 which describes the comparison of decision trees models and their hybrid bag-boost ensemble using various statistical techniques. The chapter 7 is about regional flood susceptibility and hazard analysis for the flood of 2022 and predicted flood of 2032. Chapter 8 is about modelling analysis for the assessment of correlation between flood and LULC transitions in the lower Indus basin, Chapter 9 illustrates the economic damages caused due to floods for the model based simulated flood of 2022 and predicted flood of 2032 by employing geoinformatics. Chapter 10 discusses conclusion, key findings, policy recommendations, limitations and future research suggestions.

Chapter 2
Literature Review

2.1. Introduction

Orun (2019) asserted that, Spatial analysis refers to a collection of methods used to extract fresh insights and understanding from spatial data. These approaches encompass all the methods of sampling, visualization, modification, and analysis that may be utilized for spatial data. Locations are crucial in spatial analysis as the outcomes of this analysis rely on the specific positions of the items under examination. Spatial analysis is characterized by its capacity to explicitly manage locations and spatial interactions, setting it apart from other forms of analysis. This capability, when viewed from a geographical perspective, enables efficient analysis of entities, occurrences, and operations that take place on or in close proximity to the Earth's surface. Spatial analysis focuses on the inherent geographic characteristics of things and the impact of location and spatial relationships. Through the process of recording, the connection between characteristics and phenomena and their geographic locations is established.

This chapter provides a detailed description of the previous studies related to the flood damages assessment measures, flood and hazard modelling, the need for feature selection in spatial analysis, the concept and application of metaheuristic algorithms, the application of machine learning models for spatial analysis and the usability ensemble and hybrid models in flood analysis.

2.2. Flood Damages Assessment

Depending upon the nature and intensity of damages, previous studies provide various methods for the evaluation of flood associated damages. The flood damage evaluations can be categorized into two primary components: tangible and intangible harm (James and Lee, 1971).

2.2.1. Types of Flood Damages

Adhering to Yi et al. (2010), damages from floods typically entail both qualitative and quantitative losses. Qualitative damages include social, intangible, and indirect effects, such as "emotional distress" from home instability, that cannot be measured. Quantitative damages are concrete economic losses, including direct and indirect damages, represented in monetary terms. Such damages are known as the economic damages.

Kuiper (1971) explained two primary categories of flood damages, namely tangible and intangible damages. Tangible damage refers to the type of damage that can be easily quantified in terms of monetary worth. On the other hand, intangible damage refers to the type of damage that has no direct monetary value (Kuiper, 1971). Furthermore, tangible damages can be categorized into two distinct sub-categories: direct damage and indirect damage. A direct damage refers to the harm inflicted on objects, such as structures and inventory goods, Indirect damage refers to the harm that arises from the interruption of economic and physical networks, like suspension of traffic flow and the loss of individual income. It also includes the implications of company cut-off (Lekuthai & Vongvisessomjai, 2001). From a different perspective, Merz et al. (2010) describe direct damages as the harm caused by the actual physical contact between flood water and humans, property, or any other object. On the other hand, indirect damages are the consequences that result from the direct impacts and occur in different locations and time periods, separate from the flood event itself. Both forms of damages are categorized as either tangible or intangible, based on their ability to be quantified in monetary terms (Parker, et al., 1987; Smith and Ward, 1998).

Table 2. 1: Flood Damage Categories and Losses

Category			**Losses**
Tangible	Direct	Primary	Agriculture, structures and contents
		Secondary	Environment and land recovery
	Indirect	Primary	Interruption of business activities
		Secondary	Impacts of regional and national level economy
Intangible			Health, traumatic and psychological damages

Tangible damages refer to the harm caused to human-made properties or resource flows that can be easily quantified in terms of money. On the other hand, intangible damage refers to the harm caused to assets that are not bought or sold in a market and are challenging to assign a monetary value to (Merz et al., 2010). Dutta et al. (2003) have examined various distinct types of flood damage. Dutta et al. (2003) classified the direct and indirect tangible damages, , into two categories i.e. primary and secondary, which is presented in Table 1. Merz et al. (2010) provided a summary of examples in Table 2.

2.2.2. Flood Parameters

The extent of flood damage is contingent upon various flood characteristics, including the depth and velocity of floodwater, the year and duration of the flooding, the presence of sediment and effluent, the region affected by the flood, and the effectiveness of the flood warning system (Merz et al., 2010). James and Hall (1986) and McBean et al. (1988) further

support these findings, asserting that flood induced damage is influenced by water depth and several factors that contribute to the rise in expenses resulting from flood episodes. Dutta et al. (2003) found that various factors can contribute to flood damages, however prior studies on flood damage assessment have primarily focused on water depth as the variable of interest.

Table 2. 2: Examples of Various Flood Associated Losses

Category	**Tangible**	**Intangible**
Direct	Damages to buildings and contents, interruption of infrastructure like paved roads and railways, agriculture damages such as land erosion and lack of crop harvesting, damages to livestock, rescue and evacuation measures, business interruption and flood water evacuation and cleaning costs	Life loss, injuries, psychological distress, damages to cultural heritage, adverse impacts of ecosystems
Indirect	Disruptions caused outside the flooded area such as delayed public services, induced losses to suppliers of the production companies in the flooded area, traffic disruptions, loss of taxes revenue	Trauma and lack of faith in administrative personnel.

2.2.3. Approaches of Flood Damages Assessment

The commonly used methods for estimating flood damage are unit loss model and model application. The former approach involves evaluating the impact of floods on individual properties, either in terms of their current value or their potential value. On the other hand, model applications analyze the broader effects of floods on the economy, specifically the interconnections between different sectors. This includes considering the ripple effects and links between sectors. The unit loss approach is supported by Dutta et al. (2003), while the model applications are supported by Parker (1992) and Islam (2000). The literature indicates that the majority of published information on damage collecting and analysis originates from countries such as the United States, the United Kingdom, Japan, and Australia, where a unit loss method has been employed. The United Kingdom and Australia have developed comprehensive procedures for estimating tangible losses (Smith, 1981). However, in the case of the United States, Japan, and other countries, the approach for estimating damages is limited to metropolitan areas only (USACE Manual, 1988; MOC Japan, 1996). These countries have been observed to use the same method for estimating damage, specifically the unit loss methodology (Smith, 1994).

According to Dutta et al. (2003), developing a reliable model for assessing flood losses is challenging due to the complex and unique nature of flood-related damages. Accurate estimation requires detailed information on flood parameters like flow velocity, depth, and duration at specific locations. Equally important is the precise grouping of damage types and the development of associations between flood factors and the resulting losses across various groups. The link between flood characteristics and damage levels is typically depicted using a stage-damage function. This function is derived from a combination of historical flood damage records, laboratory experiments, survey data and other relevant sources (Islam, 2000).

2.2.4. Quantification of Economic Damages of Flood

In 2004, the European Community launched the "DAMAGE" initiative to address the requirement of European civil protection services for a standardized approach to assess the damage caused by a disaster (Union, 2014). The project's objective is to furnish public administrations, such as municipal, provincial, and regional governments, with a management tool for damage information. A GIS-based model was developed to simulate flood events and assess resulting economic damages. The conventional method for damage assessment involves conducting a thorough on-site investigation to determine the exact extent of the loss. Efforts have been undertaken in recent years to establish procedures for swiftly evaluating the extent of damage following a catastrophe.

Our study aims to develop a mechanism that would allow for real-time evaluation of probable economic direct loss resulting from a catastrophe. The approach relies on a comprehensive understanding of the surrounding region combined with a depiction of certain physical characteristics of the natural occurrence. When a severe natural disaster occurs in a specific area, the economic losses resulting from direct damage to goods can be determined by considering the quantity and economic value of each element in the area, as well as the extent of damage to each unit (ranging from 0, indicating no damage, to 1, indicating complete destruction). The assessment of economic value of the losses to the elements that are exposed to flood is known as the valuation of the economic damage. The economic value of loss can be defined as:

$$Direct\ Economic\ Loss = \sum_i (Unit\ Value_i \times Number\ of\ Units\ Exposed_i \times Damage\ Degree_i(\%)) \quad (2.1)$$

Exposed elements refer to the population, property, economic activity, public services, environmental goods, and cultural heritage items that are located in a vulnerable area. The

damage degree refers to the level of vulnerability, which represents the extent of loss of one or more items at risk caused by a hazardous event of a specific intensity and length. The function is determined by the intensity and type of the element at danger.

Fifteen investigations focused exclusively on food modelling and did not assess the extent of losses incurred. These studies were conducted by Acosta et al. in 2017 and Faghih et al. in 2017. Approximately 66% of the remaining research focused exclusively on quantifiable damages that may be directly measured and expressed in monetary value. Multiple studies have effectively incorporated direct tangible damages along with indirect and intangible flood losses (Arrighi et al., 2018; Nga et al., 2018). Rather than quantifying the actual damages caused by flooding, some studies assessed flood severity using ordinal damage classes (Ronco et al., 2014; Ettinger et al., 2016). Economic damages can be evaluated on a per-event basis (disaggregated) and average annual losses (AAL) (Foudi et al., 2015; Lawrence et al., 2019). The food features that are most frequently considered when assessing damages are the size of the food, the depth of the food, the speed of the flow, and the duration of the food. A study conducted by Vozinaki et al. (2015) highlighted the need of considering the timing and seasonality of food incidents in agriculture. The food attributes most frequently utilized were flood extent and food depth. A single study conducted by Tarigan et al. in 2017 is the only one that measured the extent of food damages without considering the depth of the food. Although numerous articles emphasized the theoretic importance of water flow velocity and food interval, the majority of models only included food extent and food depth when evaluating food harms. This phenomenon is based on the fact that it is very easy to precisely evaluate the degree and depth of food damages (Mohammadi et al. 2014; Komolafe et al. 2019).

2.2.6. Prediction of Flood Inundation

A 'flood inundation area' refers to a low-lying region that is intentionally immersed in water for a specific duration. This results in the destruction of many assets within the area, including land, buildings, personal property, public facilities, infrastructure, and crops. The key components of a flood inundation map typically encompass the extent, magnitude, and duration of flooding, while considering the terrain and infrastructure for disaster mitigation in the region. Precise forecasting of flood inundation is crucial since it is the primary determinant of accuracy in evaluating flood damage. However, the inclusion of depth and period of inundation, as well as the accuracy of the outcome, will vary depending on the aim and technical limitations of predicting flood inundation. For instance, while doing a flood damage assessment for a pre-

feasibility study, it is inefficient and challenging to achieve a level of accuracy and precision that exceeds what is required due to constraints in terms of restricted resources such as time and money. Because of the aforementioned reasons, other technologies that expand the flood level to the safeguarded lowland are frequently employed. Nevertheless, employing such techniques may lead to an overestimation of the extent of flood damage based on the topographical features of the region. Lately, there has been significant research on flood inundation analysis methodologies that utilize two-dimensional diffusion wave and dynamic wave shallow water equations. However, studies mostly concentrate on analyzing flood waves caused by overtopping, making them unsuitable for predicting the flooding of protected lowland areas, which account for the majority of actual flood damage. Moreover, the economic analysis is characterized by a high level of complexity. The reference is from Yi et al., 2010.

2.2.7. Hazard Modelling

Hydrologic and hydraulic modelling software, often integrated with Geographic Information Systems (GIS), is widely used to simulate the characteristics of both historical and hypothetical flood events (Mahmood et al., 2019). The reviewed studies reveal that HEC-RAS is the most frequently utilized commonly used alongside HEC-GeoRAS (Zúñiga and Novelo-Casanova, 2019). Other software used in various studies include MIKE FLOOD and MIKE 11, SWAT and Flo-2D. These tools were employed in studies by Cham and Mitani (2015), Komolafe et al. (2018b), and Qiao et al. (2018, 2019). Many simulations focused on estimating flood extent and water depth in affected regions. In some cases, studies also estimated velocity of flood water flow and flood period (Bormudoi et al., 2013; Gergel'ová et al., 2013). The literature illustrates one-dimensional (1D), two-dimensional (2D), and combined 1D-2D hydraulic modelling approaches. 1D models assume unidirectional flow, typically from upstream to downstream. In contrast, 2D models provide a more accurate representation of water movement in areas with complex terrain, as they consider spatial variability in two directions. However, 2D modelling requires more detailed input data. Coupled 1D-2D models aim to combine the strengths of both methods, offering a balanced approach to simulate floods effectively. All three modelling approaches are generally suitable for flood damage calculations (Ahmadisharaf et al., 2015; Kobayashi et al., 2016; Nga et al., 2018).

Some researchers employed statistical surveys to analyze flood characteristics (Waghwala and Agnihotri, 2019). There are also some studies which focused specifically on individual historical flood events. A limited number also compared baseline conditions—such

as past flood scenarios—with future projections influenced by adaptation strategies (Brown et al., 2017).

Many studies utilize return periods to construct food scenarios. Examples of these studies include Tarigan et al. (2018), and Mahmood et al. (2019). A minimum of three distinct return periods is often used, as stated by Morita (2014) and Karamouz et al. (2015). This is a necessary requirement while calculating average annual damages, as mentioned by Nga et al. (2018). The study investigated return-periods ranging from 2 years to 1000 years, with 10-year, 50-year, and 100-year periods being the most commonly used (Gusyev et al. 2015; Pathak et al. 2016). To ascertain the various intervals at which different types of food are consumed, it is necessary to evaluate the frequencies of occurrence and the related magnitudes of food consumption (Waghwala and Agnihotri, 2019). The statistical methods of choice for frequency analysis were Gumbel distributions and Weibull distributions (Soliman et al., 2015; Gusyev et al., 2015).

Various researchers employed Pearson type III, generalized extreme value (GEV) or lognormal distributions in conjunction with Gumbel or Weibull distribution to assess flood frequencies (Eslamian 2014; Faghih et al. 2017). Instead of utilizing statistical approaches stated above, many studies relied on expert judgement to determine dietary frequencies (Ahmadisharaf et al. 2015; Aksoy et al. 2016; Schmid-Breton et al. 2018). If there is a lack of frequency data, it is advised to utilize a combination of the statistical extreme value distribution approaches mentioned above (Eslamian 2014; Faghih et al. 2017).

Many studies have highlighted the importance of validating modelled foods. A prominent method of validation was comparing modelled foods with historical food events based on their size and depth (Yu et al., 2013; Mahmood et al., 2019). Food models can be standardized and improved by comparing them to historical food events, as demonstrated by Kar amouz et al. (2015) and Zúñiga and Novelo-Casanova (2019). When complete data on the quantity and quality of historical foods were lacking, certain food models were verified using historical records and information from news sources. The research region underwent frequent validation by the surveys and interviews. A recurrent issue that has been identified is the lack of adequate validation data, as highlighted by studies such as Saini et al. (2016) and Komolafe et al. (2019).

2.2.8. Flood Susceptibility

The biggest challenge in flood susceptibility assessment is the accessibility to relevant and complete spatio-temporal dataset (Pravalie et al. 2013). Various methods are being used for flood susceptibility mapping, including; multicriteria analysis (Seekao and Pharino, 2016; Mahmoud and Gan, 2018; Tang et al., 2018), weight of evidence (Pourghasemi and Zeinivand, 2016), frequency ratio (FR)1(Tehrany et al., 2018; Youssef et al., 2015), maximum entropy (Siahkamari et al., 2018), the fuzzy weight of evidence (Hong et al., 2018), analytical hierarchy process (Ghosh and Kar, 2018), decision tree1(Khosravi et al., 2018), Shannon entropy (Khosravi et al., 2016b), multicriteria decision-making (Arabameri et al., 2019), principal component analysis (Nandi et al., 2016), artificial neural network (Zhao et al., 2018), multiple logistic regression (Marconi et al., 2016), classification and regression trees (CART) (Choubin et al., 2019), deep learning methods (Wang et al., 2020 and Fang et al., 2020) and support vector machines (Tehrany et al., 2018). Recently, hybrid methods are also being used for this purpose, some popular models used for flood susceptibility mapping are; Adaptive Neuro-Fuzzy Inference systems (ANFIS) (Wang et al., 2019), and bagging hybrid models with Logistic Model Trees (LMT) (Chapi et al., 2017). the Multivariate Adaptive Regression Spline, Generalized Linear Model, Random Forest and Flexible Discriminate Analysis (Mosavi et al., 2020).

One algorithm, however, has low predictive ability for flood susceptibility. The data for flood susceptibility is often less representative, which makes the model unfit in hypothesis space and true sample distribution. Hence, ensemble modelling is an appropriate method for solving such issues.

2.3. Statistical Applications of Machine Learning in Spatial Analysis

Since its inception in the 1980s, machine learning (ML) has attracted considerable attention across various fields that utilize quantitative methods. ML relies on automated algorithms to detect patterns in data and generate accurate predictions, even when the relationships between input variables are not explicitly understood. Unlike traditional statistical and econometric approaches—which emphasize formal equations, hypothesis testing, and drawing inferences about populations from sample data—machine learning places greater emphasis on predictive performance than on explanatory modeling.

Machine learning often functions as a black-box approach, meaning the internal workings of its models are not always transparent, in contrast to the more interpretable models used in classical statistics and econometrics. ML is typically used for three primary purposes: clustering data into naturally occurring groups, classifying data into predefined categories using trained algorithms, and making predictions based on input features.

Spatial approaches require the integration of spatial data. Recent assessments suggest that approximately 80% of all data contains a geographic component, much of which can be geographically referenced (VoPham et al., 2018). Spatial data can be sourced from traditional outlets such as regional databases maintained by statistical agencies, gridded datasets, and geographically referenced point data.

Modern tools like OpenStreetMap and Google Maps provide easy access to various spatial elements, including base maps, points of interest (POIs), road networks, and traffic information. Additional sources include geo-referenced imagery such as satellite photos (Rolf et al., 2021), night-time light imagery, drone footage, and geo-tagged content from social media platforms like Twitter, as well as data from climate sensors. Given the vast volume, complexity, and variety of spatial data, advanced computational methods are essential for effective analysis and interpretation.

Machine learning (ML) refers to a system's ability to improve its performance by learning from previous outcomes. It is a key component within the broader field of artificial intelligence (AI). ML techniques are typically grouped into three main categories: supervised learning, unsupervised learning, and semi-supervised learning—each defined by the type of algorithms they use. Deep learning, a specialized branch of ML, relies on neural networks for model training.

In the context of big data, machine learning plays a crucial role in data mining and knowledge extraction from large datasets. ML algorithms can be applied to existing datasets to generate independent predictions or integrated into data processing pipelines and decision-making systems within AI frameworks.

Despite its potential, the insular nature of many scientific disciplines can make machine learning seem inaccessible to a wider audience. However, ML holds considerable promise beyond big data applications—it can effectively complement spatial statistics and econometric methods, offering valuable insights even in smaller-scale or traditional data analyses.

2.3.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning technique that falls under the category of supervised learning. It is built upon the principles of risk minimization and mathematical theory, as described by Tehrany et al. in 2015. The SVM model is renowned for its rapid layer recognition and analysis, as indicated by (Micheletti et al., 2011). This model is commonly used to tackle classification and regression problems while reducing overfitting of the algorithm (Gayen et al., 2019). SVM is a widely acclaimed machine learning model known for its ability to create a hyperplane that separates training data, particularly when transitioning from actual SVM datasets to higher-dimensional feature spaces. The model's functional performance hinges largely on the selection of an appropriate kernel.

Similar to other neural networks, SVMs are prone to the challenges of overfitting and underfitting (Samantaray et al., 2023). Several researchers have documented the theory behind SVM, emphasizing that in high-dimensional feature spaces, the SVM's hypothetical space is essentially constrained to linear functions. These hypotheses are then trained using learning algorithms rooted in optimization theory, which leverages statistically driven learning principles. In this manner, the process of fine-tuning the learning machine plays a pivotal role in optimizing the SVM for generalization (Kecman, 2001).

Adhering to Wu et al. (2019), in hydrology, variables often involve nonlinear relationships, posing a challenge for linear SVM due to its limited capacity for handling nonlinearity. Solving nonlinear relationships with a linear SVM is challenging due to its limited capability to handle nonlinearity. Consequently, in this study, we have utilized linear and non-linear SVM for best feature selection among the flood-related features used in the literature. The SVM model's decision function can be expressed as follows:

$$f(x) = \omega \,.\, \Phi\,(x) + b \tag{2.2}$$

In this context, the SVM algorithm aims to find the optimal values for the parameter vectors ω and b, as well as certain internal parameters of the function $(\Phi)(x)$, where x represents the input variable vector. The following optimization problem is solved to achieve the best possible solution in SVM;

$$min_{\omega,b,\xi,\xi^*} \frac{1}{2}\omega^2 + C \sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.3}$$

$$\text{Subject to} \begin{cases} y_i - (\omega . \varphi(x_i) + b) \leq \varepsilon + \xi_i \\ (\omega . \varphi(x_i) + b) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i , \xi_i^* \geq 0, i = 1, \dots \dots, n \end{cases} \quad (2.4)$$

In equations 2.3 and 2.4, ξ and ξ_i represent those variables that are used to quantify training errors according to a specific threshold. *C* represents a positive constant penalty coefficient, determining the extent of training error. The goal of SVM is to minimize both these values simultaneously.

The kernel, a critical component in SVM, is a function responsible for calculating the inner product directly from input data points. The choice of kernel function has significant importance as it enables the transformation of a non-separable dataset in the original input space into a separable dataset in the feature space

The performance of Support Vector Machines (SVM) is significantly influenced by the choice of hyperparameters and kernel functions (Nieto et al., 2014). The commonly used RBF kernel is prevalent in SVM models (Deka, 2014). However, SVM suffers from challenges related to dimensionality and substantial time consumption during parameter determination, as mentioned. To address these issues, scientists have developed various nature-inspired evolutionary optimization algorithms (Samantaray et al., 2023).

2.3.2. K-Nearest Neighbors (KNN)

There are numerous methods for carrying out hydrological modeling. For performing this task, Rajkumar and Subrahmanyam (2021) compared KNN, SVM, and tree-based classifiers and asserted that KNN performed better. Moreover, they highlighted the challenge of selecting the appropriate parameter *k* in KNN, where a small *k* can make the algorithm sensitive to outliers, while a large *k* may introduce too many points from different classes. Therefore, the proper choice of *k* is crucial. Despite the simplicity of KNN and its effectiveness as a nonparametric technique for dataset subset selection, it faces issues related to accuracy and sensitivity to *k*.

K-nearest neighbors (KNN) is a learning strategy that uses distance measurements to estimate the future response of a given location by evaluating the major class of the k-closest points. This approach was introduced by Cover and Hart in 1967. The KNN algorithm is a straightforward and intuitive learning technique commonly used in many applications (Cheng et al., 2014).

The K-nearest neighbor (KNN) classifier is a fundamental and straightforward algorithm that operates by calculating similarity measures. It does this by storing all available examples and classifying new instances based on their similarity to the examples. It employs the lazy learning approach. This method identifies the k nearest neighbors of the training data set for each test image that needs to be forecasted. The Euclidean Distance measure is used to calculate the proximity of each member of the training set to the test class. The class labels of a test image can be determined by employing the K closest neighbor algorithm, followed by conducting majority voting to obtain the final class label. The value of k is dependent on the data used. Increasing the value of k reduces the effect of noise in the classification, but it also leads to less variation in the boundaries formed by the classes. Euclidean distance is quantified using the formula:

$$d_x(x, w_k) = \sqrt{(x - w_k)^T (x - w_k)} \tag{2.5}$$

To enhance the prediction accuracy of KNN and overcome its known limitations, researchers have turned to data reduction techniques using optimization algorithms. Methods such as Particle Swarm Optimization (PSO) and Correlation-based Feature Selection (CFS) are instrumental in feature selection, particularly for improving KNN's accuracy in predicting occurrences like malignant tumors, such as sarcoma in skeletal bones (Baskaran et al., 2018). For instance, the PSO-KNN method achieved 85% accuracy compared to the CFS-KNN model's 81% accuracy in sarcoma prediction.

In the research conducted by Li et al. (2019) for forecasting water quality in Poyang Lake, China, the PSO-KNN model outperformed the CART model with 86.68% prediction accuracy. Similarly, the Genetic Algorithm GA-KNN model achieved 83.12% accuracy, making it optimal for predicting health conditions like diabetes. Additionally, the GA-KNN model demonstrated an impressive 92.68% forecasting accuracy for the prostate disease dataset (Gunavathi and Premalatha (2018).

2.3.3. Decision Tree (DT)

A decision tree (DT) creates a hierarchical structure that starts at the root and extends to the leaf nodes (Breiman et al., 2017). The data is divided into distinct subsets using a prediction rule. This means that population subgroups are defined in a hierarchical manner using a series of binary partitions of the model's expected data (Ashwini Venkatasubramaniam et al., 2017). Decision trees are widely recognized as one of the most effective methods for representing

classifiers. This technique utilizes the given data to construct a decision tree. The structure resembles a tree that generates a prediction model. Each internal node represents a test on an attribute, and each outgoing branch indicates the test outcome. Similarly, each leaf node is labeled with the class of the image.

The decision tree algorithm is often employed for classification and prediction tasks. This strategy appears straightforward, yet it is the most impactful approach for representing knowledge. Decision tree models are typically represented as a structure like a tree. Decision tree learning involves determining the optimal location for splitting at each node and accurately estimating the depth of the tree structure through careful analysis. The leaf node denotes the class of the data. Classifying the instances of the tree can be achieved by sorting the decision tree from the root node to the leaf node. Decision trees are mostly recognized for their noise resistance, which is achieved by pruning procedures that effectively eliminate overfitting in both common and Gaussian noisy data. Decision trees are visualized as a hierarchical structure resembling a flowchart. Each internal node represents a test on an attribute, each branch reflects the conclusion of the test, and each leaf node or terminal node represents a class label. The qualities of a given tuple X are evaluated against the decision tree. A path is traced from the root to the leaf node to handle the class prediction for the tuple. Converting a decision tree into categorization rules is a simpler process. Learning a decision tree involves utilizing the decision tree as a predictive model that maps the observations of items to their corresponding target value conclusions.

It is utilized across several disciplines such as statistics, data mining, and machine learning, and is regarded as one of the methods for predictive modelling. The creation of a decision tree is rather rapid when compared to other categorization approaches. The decision tree assists in the creation of SQL statements to efficiently access the database. The accuracy of decision trees is shown to be comparable or, in some situations, superior in performance when compared to other categorization approaches. Using a decision tree inducer algorithm, decision trees can be automatically constructed from a given dataset. The primary objective is to identify the optimal decision tree by reducing the generalization error. However, the remaining aim functions can also be elucidated, such as reducing the number of nodes or diminishing the average depth (Rajkumar and Subrahmanyam, 2021). Decision trees can be categorized into two distinct categories.

a) **Classification decision trees:** are a specific type of decision tree that is used when the decision variable is categorical.
b) **Regression decision trees:** are a specific type of decision tree that is used when the decision variable is continuous.

2.3.3.1. Base Classifiers

In this research, four tree based classifiers will be considered as base classifiers in the ensemble modelling. Random Forest (RF) (L. Breiman, 2001), Naive-bayes Tree (NBT) (R. Kohavi, 1996), Logistic Model Trees (LMT) (Landwehr at al. 2005) and Reduces Error Pruning Tree (REPT) (Quinlan, 1987) will be used since they are computationally feasible and possess better predictive accuracy in various applications (Tama and Rhee., 2015). The base classifiers are briefly discussed as follows.

2.3.3.2. Random Forest (RF)

Random Forest (RF) is a bagging ensemble learning method that demonstrates strong performance in both classification and regression domains, as stated by Breiman (2001). It is founded on the concept of integrated learning and the delineation of many autonomous trees. This generates several trees. These trees are generated with no pre or post pruning. At each node end, the feature going to be split is selected from a random split of the original feature. Classification accuracy is attained because of diverse nature of the trees. There are only two parameters in Random Forest, which are the number of trees and number of variables at each split (L. Breiman, 2001). The predictions are obtained by aggregating the outcomes from randomized and de-correlated decision trees (Svetnik et al., 2003).

Random Forest (RF) is generated by integrating decision trees with random sampling techniques. The algorithm trains decision trees by randomly selecting a set of data attributes and generates several decision trees. The ultimate outcome is the amalgamation of the outcomes from numerous decision trees, typically accomplished through a straightforward process of majority voting or averaging. An advantage of RF is its ability to effectively manage the correlation between features without requiring any further handling. Furthermore, it exhibits a high degree of adaptability to diverse data distributions. Moreover, the RF method possesses a high degree of interpretability. It has the capability to compute the significance of each attribute for each decision tree, enabling the determination of the importance of that

attribute to the data. The process of tweaking the hyperparameters of the RF algorithm in real applications is straightforward and user-friendly.

The algorithm enhances the variety of classification trees by systematically alternating data and making arbitrary adjustments to the collection of explanatory elements during the different stages of tree induction (Arabameri et al., 2020). The essential hyperparameters for tree growth are the number of trees (k) and the number of predictive factors (m) utilized for node splitting. The OOB error, also known as the out of bag error, is a measure of the percentage of misclassified items out of the total number of objects. It provides a reliable approximation of the generalization error. The out-of-bag (OOB) error is calculated throughout the model building process. The paper by Breiman (2001) states that the random forest algorithm establishes an upper bound for the generalization error. This inaccuracy frequently decreases as the quantity of trees increases. Consequently, k must possess a sufficient magnitude to facilitate such convergence. This method determines the value of the predictive variable by analyzing the decrease in error when the data is rearranged for that variable, while keeping the other variables constant. The increase in error is directly proportional to the value of the explanatory variable (Breiman, 2001). One of the primary benefits of the random forest algorithm is its robustness against overfitting and its ability to generate several trees without the risk of overtraining. Hence, there is no necessity to resize, modify, or alter the algorithm. Regarding the predictors, the random forest algorithm is quite robust to outliers and has the ability to handle missing values automatically (Crippen, 1990).

Random Forest (RF) is a robust and efficient machine-learning technique designed to enhance skill prediction by expanding on the principles of regression and classification trees (Razavi-Termeh et al., 2019). This technique exhibits reduced computing complexity, superior performance in feature spaces with high dimensions, and enables the quantification of input variable significance, hence enhancing comprehension of their impact on total classification accuracy (Rodriguez-Galiano et al., 2012).

Random Forest (RF) is a widely used and reliable classifier in the field of remote sensing. It is favored because it does not rely on assumptions of normal distribution and variance homogeneity. Additionally, RF has a good classification performance, as noted by Belgiu and Drăgut (2016). RF, a classifier, has demonstrated its effectiveness in various applications such as land use studies using satellite imagery (Gislason et al., 2006; Jin et al., 2018; Eisavi et al., 2015), urban studies based on aerial photography (Schlosser et al., 2020), and identification of

geomorphological objects using DTMs (Phinzi et al., 2020; Zhu and Pierskalla, 2016). The crucial hyperparameters of the Random Forest Regression (RFR) are the number of trees and the randomly picked features. Zhou and Kang (2023) and Katipoğlu and Sarıgöl (2023) employed the RFR method for flood routing and conducted a comparative analysis of its performance against several other machine learning models. Zhou and Kang (2023) observed that the RFR model exhibited overfitting when used to estimate the inflow hydrograph of the Three Georges Reservoir.

2.3.3.3. Logistic Model Trees (LMT)

Adhering to Quinlan (1993), LMT is classification technique that integrates logistic regression machine learning approaches and decision trees (using the C4.5 algorithm). These approaches are derived from a model composed of hierarchical structure with inner nodes and terminal nodes. C4.5 technique is utilized at nodes, whereas the function of logistic regression is employed at the leaves. The C4.5 decision tree is capable of categorizing flood influencing factors into flood and non-flood classes by considering their probability (Chen et al., 2018, 2019a, 2019b). The posterior probability of a leaf node in derived from the linear logistic regression.

$$P(N|x) = \frac{\exp(L_i(x))}{\sum_{i=1}^{n} \exp\,(L_i(x))} \tag{2.6}$$

The expression $P(N|x)$represents the posterior probability in a leaf node, where "N" is the number of classes and "x" represents the input vector. On the other hand, $L_i(x)$ refers to the least-square fits, which may be attained by employing the following equation.

$$L_i(x) = \sum_{i=1}^{n} \alpha_i x_i + \alpha_0 = 0 \tag{2.7}$$

The variables α_0 and α_i reflect the coefficients of the influencing factors in the vector x = x_i, and n is the number of flood factors.

2.3.3.4. Reduced Error Pruning Trees (REPT)

REPT is a combination of DT method and Reduced Error Pruning (REP) method. It constructs a regression tree using the information gain reduction technique proposed by Quinlan in 1987. The decision tree constructs the classification tree by identifying the input variable with the largest gain ratio, which is calculated according to the formula provided by Tien Bui et al. (2012).

$$Gain\ Ratio(x, Z) = \frac{Entropy(Z) - \sum_{i=1}^{n} |Z_i| Entropy(Z_i)}{-\sum_{i=1}^{n} \frac{|Z_i|}{|Z|} \log_2 \frac{|Z_i|}{|Z|}} \quad (2.8)$$

The attribute x is a part of the training dataset Z, which consists of subsets Z_i, where i ranges from 1 to n.

The Reduced Error Pruning (REP) technique is used in decision tree (DT) models to simplify the tree structure by removing branches and leaves that have minimal impact on classification accuracy (Mohamed et al., 2012; Galathiya et al., 2012). Originally introduced by Quinlan in 1987, REP is widely recognized for its straightforward implementation and effectiveness. It is often evaluated alongside methods such as feature selection and cross-validation (Galathiya et al., 2012). One of the key benefits of REP is its ability to streamline the decision tree, helping to prevent overfitting during the training process while still preserving a high level of predictive accuracy (Polo et al., 2008).

2.3.3.5. Naive Bayes Trees (NBT)

The NBT is a combination of Naïve Bayes (NB) and Decision Tree (DT) models that are built on the principles of Bayes' theorem (Kohavi, 1996). The last decision tree is formulated by a univariate split at each node, but with the NB classifiers at leaves. The DT divides the data and each part of the data, represented by a leaf, is depicted through a NB classifier. No pre-parameter setting is needed for this algorithm (R. Kohavi, 1996). This categorization approach is widely used since it is simple, efficient, performs well, and is easy to understand. This approach has a low computer memory requirement and may be quickly learned from a training set (Wang et al., 2015a).

During the construction of a decision tree, pre-pruning can be applied using one of two approaches: (1) by splitting the data at a given node, or (2) by creating a leaf node that incorporates a locally trained Naive Bayes (NB) model using the data associated with that node (Landwehr et al., 2005). As noted by Kohavi (1996), the Naive Bayes Tree (NBT) model demonstrates superior performance compared to standalone decision tree (DT) and Naive Bayes models. The NBT approach leverages the concept of entropy and uses attribute selection measures to guide the tree-building process. Let D represent a set of cases and |D| represent the total number of cases. These cases can be categorized into m classes, denoted as Di (where i = 1,2, …m), with $|D_i|$ representing the number of cases belonging to class D_i. The entropy values for classifying the set D can be approximated using the following method:

$$Entropy\ (D) = -\sum_{i=1}^{m}(\frac{|D_i|}{|D|})\log_2[\frac{|D_i|}{|D|}] \quad (2.9)$$

The NB algorithm assumes that the predictive qualities$a_1, a_2, \ldots\ldots, a_m$ are independent of each other. To calculate the joint distribution, we can use the class attribute C_j, which is a class attribute of the class set C.

$$P(a_1, a_2, \ldots\ldots, a_m|C_j = \prod_{i=1}^{m} P(a_i|C_j) \quad (2.10)$$

The NB statistical classification method relies on the principle of conditional probability, assuming that the qualities are independent of one another (Soni et al., 2011). The tool enables the user to predict the required parameters for classification using a small amount of training data (Bhargavi and Jyothi, 2009). Equation 2.10 is employed to ascertain the classification rule by considering the observation of k attributes. The flood conditioning factor, a_i, ranging from 1 to k, determines the output class, while C_j indicates whether the scenario is a flood or non-flood. Naïve Bayes computes the probability $P(a_i|C_j$ for each possible output class. The prediction is generated for the class that has the highest probability after considering all available information, whereas the prior probability $P(C_j)$ can be calculated by determining the proportion of observations with output class cj in the training dataset (Chen et al., 2018). As a result, the NBT removes the limitation of assuming attribution independence, which broadens the classifier's potential applications and enhances the accuracy of the classification results (Chen et al., 2020).

2.4. Ensemble Learning

In Machine Learning, every trained model gives different outputs and fail under some statistical conditions. Thus, models are modified for acquiring better performance and accuracy on a training data set. The modification is a complex procedure and there exists a possibility that a model giving best performance for training data set may not be accurate for the new data set. Ensembles propose that there always exists a model which can perform better than other comparative models on the new data set. (Sk Ajim Alia et al., 2020). Hence, by effectively integrating multiple trained models, performance can be enhanced comparatively than a single model. Therefore, the purpose of ensembles is to find a set of models that are different from each other but provide performance diversity and each single model is still accurate enough to give an overall performance improvement (A. Ethem, 2014).

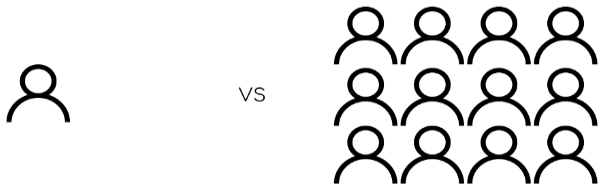


***Figure 2. 1*:** An Ensemble

Ensemble models are more diverse, accurate and tend to reduce overfitting. The ensemble methods include **bagging, boosting**, and **stacking-based** methods (Islam et al., 2008).

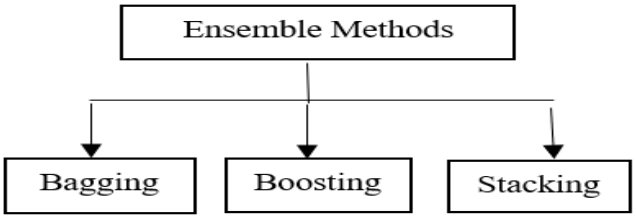


***Figure 2. 2*:** Ensemble Methods

2.4.1. Classifier Ensemble

Classifier ensemble has better performance compared to the single classifier (Tama and Rhee., 2015). It is developed by integrating different base classifiers for prediction of final output. Ensemble of SVM and DTs using weighted ensemble approach is suggested by Peddabachigari et al. (2007). Three different classification integration techniques, i.e. maximum probability, minimum probability and product rule are suggested by Giacinto et al. (2008). The authors utilized classification combination approach to improve the output results of base classifiers, i.e. υ-support vector and *k*-means classification. Govindarajan and Chandrasekaran (2011) conducted classifier fusion by employing bagging strategy and exploited multilayer perceptron and radial basis function (RBF) as base classifiers. Voting combiner is adopted in the study conducted by Sidhu et al. (2012) to fuse wo base classifiers i.e., neural network and decision tree.

2.4.2. Ensemble Learning Process

There are three ways for ensemble modelling, namely, bagging, boosting and stacking. The former two categories are considered for homogenous models and the later for heterogenous models. One more difference between these ensemble learning processes is that the bagging reduces variance and boosting and stacking reduces bias. The current research

considers an integrated bagging-boosting ensemble learner as the considered models are homogenous in nature, it will be advantageous for reducing variance and bias as well.

2.4.3. Bagging

Bagging is an ensemble method in which bootstrapped samples are drawn from training the dataset in parallel. The bootstrap sampling generates a different training dataset for each classifier, which enhances the diversity of an ensemble and in the last step the outputs are combined through majority voting. It is based on the concept that that a small change in the training dataset leads towards a large change in the classifier output (M.M. Islam et al., 2008). Bagging can reduce the variance of unstable models, leading to an improved prediction. Figure 2.3 depicts the bagging strategy.

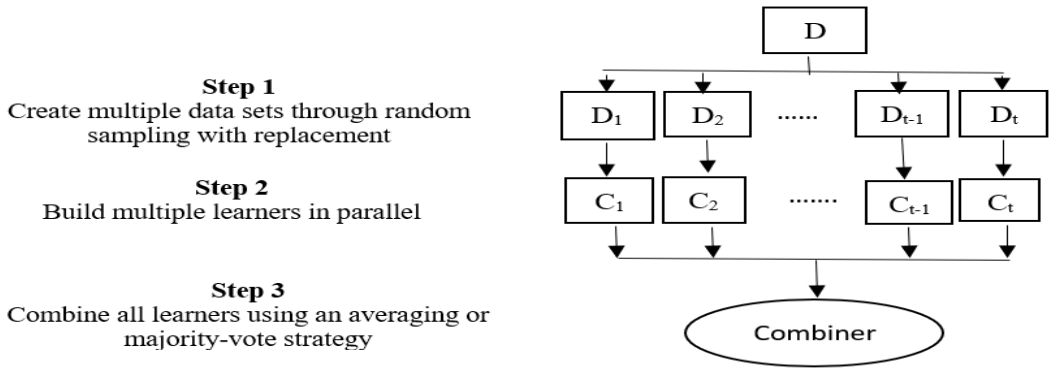


Figure 2. 3: Steps of Bagging Methodology

2.4.4. Boosting

Boosting is also a generation process. The aim of boosting technique is to sequentially use weak classification algorithms to repeatedly modified versions of the dataset. The predictions attained from all of them are then integrated by using weighted majority vote for acquiring the final prediction (Hastie et al., 2009). In boosting, training of a particular model is dependent upon the performance of the previous models. This can generate considerable improvement in the performance compared to the use of a single model (M.M. Islam et al., 2008). Figure 2.4 depicts the boosting strategy.

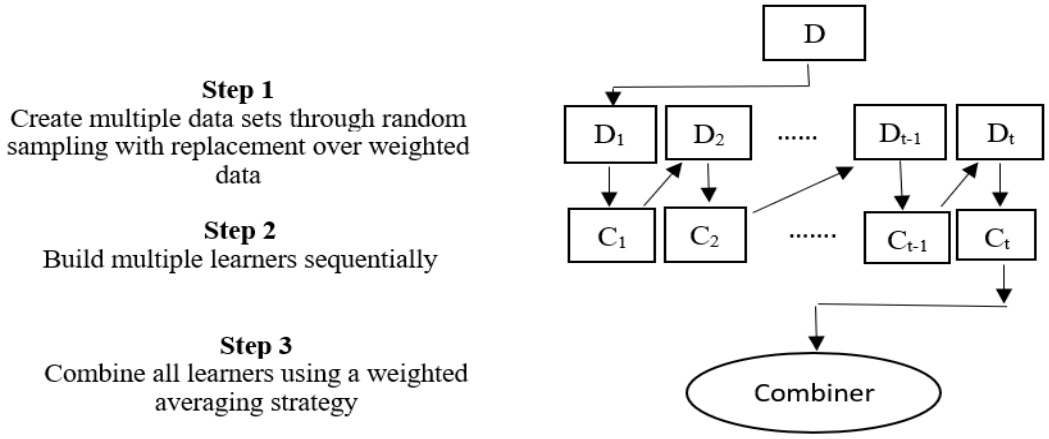


Figure 2. 4: Steps of Boosting Methodology

2.4.5. Joint Bagging Boosting

Within the field of machine learning, several trained models might yield disparate outcomes and exhibit vulnerabilities in distinct scenarios. While it is feasible to adjust a model to achieve optimal accuracy on a specific training data set, this modification process is intricate and there may still be cases where the model's performance on the training data set does not adequately translate to accuracy on a new data set. Ensemble learning operates under the premise that there is always a possibility of finding a model that outperforms another model when applied to a fresh dataset. Thus, by effectively amalgamating many trained models, the accuracy can be enhanced compared to using a single model. The objective of ensemble learning is twofold: (1) to identify a collection of models that exhibit variety in order to enhance performance, and (2) to ensure that each individual model is sufficiently correct to contribute to an overall improvement in performance (Ethem, 2014).

This research presents an ensemble learning strategy employing decision trees. An ensemble is comprised of two elements: a technique for generating individual decision trees model and a technique for merging the individual decision trees models together (Islam et al., 2008). Both strategies should be designed to achieve an overall enhancement in performance. Bagging generates Nbag decision tree models for the ensemble by training these decision trees models independently on Nbag distinct training datasets. These datasets are created by generating bootstrap replicas of the original training data. Boosting employs Nboost decision trees models that are taught in a sequential manner. Bagging and boosting have been separately utilized for forecasting concerns. Our proposal is to integrate bagging and boosting in order to leverage the variance and bias reduction capabilities of both methodologies.

2.5. Multicollinearity, Heteroscedasticity, and Spatial Autocorrelation

The presence of multicollinearity, heteroscedasticity, and autocorrelation problems violate the independent and identically distributed (IID) assumption of the Ordinary Least Square (OLS) model and therefore, causes the model estimates to be inefficient, inconsistent, and biased. Thus, the dataset ought to be free from these issues as they adversely affect the model performance and results. The causes and consequences of these econometric problems are discussed as follows. Multicollinearity occurs when variables are significantly correlated not only to each other but also with the dependent/response variable. Multicollinearity results in making some statistically significant variables insignificant and hence creates biased model outputs (Shrestha, 2020). Similarly, heteroscedasticity is also an econometric problem that is caused due to the presence of outliers and omitted variables in the dataset. It results in an inflated variance of the error term in the regression model. With the presence of heteroscedasticity the OLS estimators are still unbiased and consistent but are inefficient because the variance is biased and therefore outputs of the model do not remain accurate Klein et al., 2015). Spatial autocorrelation (SA) is the correlation among georeferenced data sets arising due to their relative locations in geographical space. Spatial autocorrelation arises due to; self-correlation, information content, map pattern, spatial spillover, missing variables, and an indicator of areal unit demarcation. It is defined as; the correlation arising due to observations' relative geographical and locational positions. Due to nearby location value concentrations, there exists a strong tendency for georeferenced observations of some variables to form data clusters (Goodchild, 2005). Hence, with the presence of SA, the estimates of the model become inconsistent and biased. In light of the above-mentioned causes and consequences of these econometric problems, there arises a need for feature selection as these issues negatively affect the model performance.

2.5.1. Spatial Autocorrelation

Spatial autocorrelation (SA) is the correlation among georeferenced data sets arising due to their relative locations in geographical space. In 1974, Paelinck coined the term Spatial Econometrics. Besag (1974) developed the term auto-model and described a wide range of probability models incorporating spatial autocorrelation.

Spatial autocorrelation arises due to; self-correlation, information content, map pattern, spatial spillover, missing variables, and an indicator of areal unit demarcation. It is defined as; the correlation arising due to observations' relative geographical and locational positions. Due

to nearby location values concentrations, there exists a strong tendency for georeferenced observations of some variables to form data clusters. Socio-economic or demographic data has a moderate SA propensity, but remotely sensed images have a strong SA propensity as it signifies duplicate or redundant information. SA indicates a spillover of information from one location to another, resulting in redundant information for georeferenced data, this duplicative information creates complications in the statistical analysis of georeferenced data. This problem does not exist in the statistical analysis of traditional data which is composed of independent and identically distributed observations (Goodchild, 1986).

2.5.2. Spatial Heteroscedasticity

Spatial econometrics (Anselin 1988) deals with the missing variables and diagnostic tools for spatial heteroscedasticity. Geographical Weighted Regression (GWR) addresses this data feature and specify smooth geographic variations in local relationships between a response variable and covariate variables across a geographical landscape. Therefore, heuristic methods are often utilized to quickly find optimal or near-optimal solutions. Meta-heuristic models provide a general framework to strategically design heuristics in order to achieve improved solutions and computational efficacy (Lin and Gen, 2009)

In spatially referenced data, as the number of neighborhood values change from point to point, so, even if the element of noise is independent and identically distributed (I.I.D.), the variance of error becomes heteroskedastic, thus stationarity of covariance is not satisfied (Anselin, 2001). Heteroskedasticity creates difficulty in estimation of discrete choice models (McMillen, 1992). Intrinsically, variance is not homoscedastic in spatial econometric models, hence, the IID assumption is usually not fulfilled in such models. (Hajime Seya et al., 2020)

2.5.3. Spatial Multicollinearity

In the global regression based models, like Ordinary Least Squares Regression (OLS), results are not reliable due to presence of multicollinearity. Geographical Weighted Regression (GWR) provides a local regression analysis for all features in the dataset. Whenever, explanatory variables cluster spatially, there exists the problem of local multicollinearity. When nominal or categorical data categories cluster spatially, there is a risk of encountering problem of local multicollinearity. Results in the existence of local multicollinearity are unstable. (Hajime Seya et al., 2020)

2.6. Feature Selection

Feature selection is a crucial process in machine learning, aimed at identifying and extracting the most relevant features from datasets, as emphasized by Liu and Yu (2005). This challenge is encountered across various fields, demonstrating the diverse applications of feature selection. Numerous feature selection approaches have been developed, broadly categorized into three types: filter, wrapper, and embedded methods, as highlighted by Hoque et al. (2014). Filter methods operate independently of any specific classification algorithm and concentrate on the general characteristics of a dataset (Xu et al., 2018). In contrast, wrapper methods incorporate the classification algorithm and interact directly with it, providing more accurate results but being computationally more intensive compared to filters (Hoque et al., 2014). Embedded methods combine elements of both filters and wrappers. These methods use learning algorithms within their process and are therefore categorized as wrapper approaches (Tang et al., 2014). Wrapper methods, although delivering superior results to filter methods, are slower due to their dependence on the modeling algorithm. They generate subsets based on different search strategies, as distinguished by Jovic et al. (2015).

Mathematically, the feature selection can be defined within a dataset *S* containing *d* features. The goal is to obtain the most pertinent features from dataset *S*, creating subsets like $D = \{f_1, f_2, f_3, \dots \dots f_n\}$, where $n < d$ and D denotes the features of any dataset. In the realm of machine learning, choosing the right number of predictors is vital, and one approach to accomplish this is through feature selection. Feature selection is a technique used to identify the most relevant set of features in a considered dataset (Dodangeh et al., 2020). These methods are broadly categorized into heuristics, exhaustive search, and non-deterministic search techniques. In the present study, heuristics are employed for feature selection as our dataset had econometric issues of high multicollinearity, heteroscedasticity and spatial autocorrelation.

2.7. Metaheuristic Algorithms for Feature Selection

Metaheuristic algorithms are optimization techniques designed to find optimal or near-optimal solutions for complex problems. These algorithms are unique because they do not rely on derivatives, offering simplicity, flexibility, and the ability to avoid getting stuck in local optima (Mirjalili et al., 2014). Metaheuristic algorithms initiate their optimization process by generating random solutions, eliminating the need for calculating derivatives of the search space, which is a requirement in gradient search methods.

These algorithms are flexible and can be conveniently adapted to specific problems, making them highly versatile. One of the key characteristics of metaheuristic algorithms is that they can prevent premature convergence. Their stochastic nature allows them to function like a black box, exploring the search space efficiently and effectively while avoiding local optima, as noted by Olorunda and Engelbrecht in 2008 and Lin and Gen in 2009. Moreover, they also mentioned that during the exploration phase, these algorithms thoroughly investigate promising areas within the search space. In the exploitation phase, they focus on local searches within these identified promising regions. This balance between exploration and exploitation is a fundamental aspect of metaheuristic algorithms.

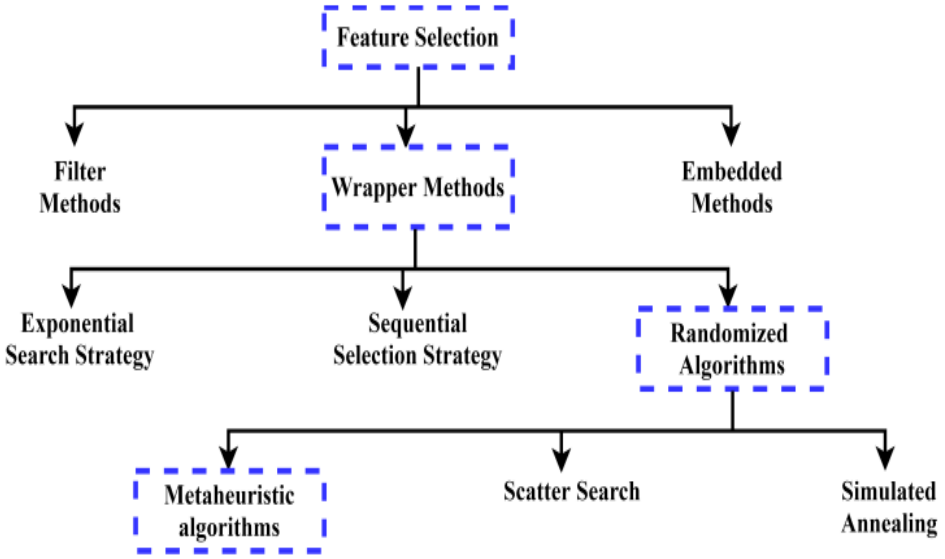


Figure 2. 5: Classification of Feature Selection Methods

2.7.1. Fitness Function

The fitness function is a key component in metaheuristic algorithms, as it evaluates the quality of each candidate subset. In the context of online classification tasks, the goal is to identify a feature set that is both compact and capable of delivering high classification accuracy. Therefore, the fitness evaluation aims to balance these two objectives by combining the classifier's accuracy with the size of the selected feature subset. The fitness function described in this study is given by Equation 2.11;

$$S(X) = \upsilon J(X) + \Psi|X| \qquad (2.11)$$

In the given context, X represents a subset, $J(X)$ represents the classification accuracy of subset X, |X| represents the number of features in subset X, and υ and ψ are variables used to

alter the relative relevance of accuracy and size. This study utilizes two classifiers, KNN and SVM, to assess the quality of the subgroup. Following meta-heuristics have been employed in this research based on their applicability in hydraulic analysis.

2.7.1.1. Particle Swarm Optimization for Feature Selection

PSO is classified as a population-based metaheuristic algorithm, which was introduced by Kennedy and Eberhart in 1995. Particle Swarm Optimization (PSO) navigates the search space by the movement of particles, each possessing a velocity. The position of each particle is updated by considering its previous best position and the current best position of all particles. The Particle Swarm Optimization (PSO) algorithm effectively achieves a balance between exploration and exploitation, making it an efficient optimization technique (Chen et al., 2012). The particle *i* is represented by the notation $\overset{r}{X_i} = (x_{i,1}, x_{i,2}, L, x_{i,d})$, and has the velocity $\overset{1}{V_i} = (v_{i,1}, v_{i,2}, L, v_{i,d})$. $\overset{1}{P_i} = (p_{i,1}, p_{i,2}, L, p_{i,d})$, signifies the previous optimal position of particle *i*. $\overset{r}{P_g} = (p_{g,1}, p_{g,2}, L, p_{g,d})$ denotes the optimal position of the best particle. Each particle bit exists in one of two states, specifically zero or one, and can transition between these states based on probabilities. The approach use the following sigmoid function to convert velocity from continuous space to probability space:

$$sig(v_{i,j}) = \frac{1}{1+\exp(-v_{i,j})}, \quad j = 1,2,L,d \tag{2.12}$$

The velocity has been recalculated as;

$$v_{i,j}^{t+1} = w \times v_{i,j}^{t} + c_1 \times r_1(p_{i,j}^{t} - x_{i,j}^{t}) + c_2 \times r_2(p_{g,j}^{t} - x_{i,j}^{t}) \tag{2.13}$$

The inertia weight, denoted as w, is changed per iteration t by using the following equation;

$$w_t = w_{min} + (w_{max} - w_{min})\frac{(t_{max}-t)}{t_{max}} \tag{2.14}$$

Where w_{max} represents the maximum value of the inertia weight, and "w_{min} represents the minimum value of the inertia weight. The t_{max} represents the maximum number of iterations. The variables c_1 and c_2 represent the coefficients that determine the rate of acceleration. The parameters r_1 and r_2 are randomly generated values that range from 0 to 1. $x_{i,j}$, $p_{i,j}$, and $p_{g,j}$ are elements that can have a value of either 0 or 1. v_{max} represents the upper limit of velocity. The revised position of the particle is determined using Equation 2.15:

$$x_{i,j}^{t+1} = \begin{Bmatrix} 1, \ if \ rnd < sig(v_{i,j}) \\ 0, \ if \ rnd \geq sig(v_{i,j}) \end{Bmatrix}, j = 1,2,L,d \Bigg\} \quad (2.15)$$

Where rnd is a random number ranging from 0 to 1, following a uniform distribution.

2.7.1.2. Ant Colony Optimization for Feature Selection

The Ant Colony Optimization (ACO) algorithm is inspired by the behavior of ants efficiently finding the shortest path between a food source and their nest, as originally introduced by Dorigo et al. in 1996. In ACO, the problem is modeled as a graph, where ants search for the path with the lowest cost. The favorable paths reflect the result of the ants' collective cooperation on a global scale. In each iteration, multiple ants construct potential solutions by leveraging both heuristic information and pheromone trails, guiding them toward optimal outcomes (Zhao et al., 2015). In the Ant Colony Optimization for Feature Selection (ACOFS) algorithm, each potential solution is transformed into an ant, which is represented by a binary vector. In this vector, a value of one indicates that the relevant feature is selected, while a value of zero indicates that the feature is not picked. Heuristic data: Heuristic information typically indicates the desirability of each characteristic. Without the use of heuristic information, the algorithm would exhibit a greedy behavior, resulting in the inability to find the optimal solution (Bolón-Canedo et al., 2013). In this study, the information gain is estimated to assess heuristic information (Forsati et al., 2014). Selection of features: During each iteration, the ant k decides the selection of feature i based on the transition probability p_i. The transition probability, denoted as pi, is defined as follows:

$$p_i^k(t) = \begin{Bmatrix} \frac{[\tau_i(t)]^\alpha [\eta_i(t)]^\beta}{\sum_{u=J^k} [\tau_u]^\alpha [\eta_u]^\beta} & if \ i \epsilon J^k \\ 0 & otherwise \end{Bmatrix} \quad (2.16)$$

The symbol J^k represents the set of feasible features, while η represents the heuristic desirability of feature *i*, and τ_i represents the pheromone value of feature *i*. The variables α and β are utilized to fine-tune the relative significance of the heuristic information and pheromone.

Pheromone Update

Once all ants have completed constructing their feature sets, the algorithm initiates the process of pheromone evaporation. According to Equation 2.17, each ant k deposits a specific amount of pheromone $\Delta\tau_i^k(t)$, which is computed using the following equation:

$$\Delta\tau_i^k(t) = \begin{cases} \gamma(s^k(t)) & if\ i \in s^k(t) \\ 0 & otherwise \end{cases} \quad (2.17)$$

The notation $s^k(t)$ represents the feature set found by ant k at iteration t. The notation γ represents the fitness function. The pheromone might be modified in accordance with the subsequent regulation.

$$\tau_i(t+1) = (1-\rho)\tau_i(t) + \sum_{k=1}^{m} \Delta\tau_i^k(t) + \Delta\tau_i^g(t) \quad (2.18)$$

The symbol ρ belongs to the interval (0,1) and represents the pheromone decay coefficient, which helps prevent stagnation. m represents the quantity of ants, while g symbolizes the most superior ant. The pheromone is updated for all ants according to Equation 2.18.

2.7.1.3. Genetic Algorithm Optimization for Feature Selection

The Genetic Algorithm (GA), introduced by Holland, is a metaheuristic that draws on the principles of genetics and natural selection. It has been effectively applied to feature selection (FS) tasks, as demonstrated by Pedergnana et al. (2013). In evolutionary terms, species adapt their genetic makeup to survive in complex and changing environments. Similarly, in GA, a chromosome represents a potential solution to a problem and is evaluated using a fitness function. Through processes such as crossover and mutation, GA generates new populations with improved chances of success. The typical steps involved in a genetic algorithm include encoding, selection, crossover, and mutation.

In the Genetic Algorithm Feature Selection (GAFS), a chromosome represents a potential set of features that consists of genes. Each gene is a feature that is encoded as a binary string. If the gene is encoded as '1', it indicates that the relevant characteristic is selected. Otherwise, if the gene is encoded as '0', it means that the feature is not selected. The term "bit i" in the chromosome represents the specific characteristic denoted by the feature i. The random initialization of each chromosome occurs. Selection: The process of selection involves choosing the most desirable chromosomes from the present population to be included in the following generation. A specific quantity of chromosomes is stochastically chosen for the succeeding generation, with the probability of selecting chromosome i being represented by equation 2.19.

$$\Pr(i) = \frac{S(i)}{\sum_{j=1}^{m} S(j)} \quad (2.19)$$

$S(i)$ represents the fitness value of chromosome *i*, while *m* indicates the total number of chromosomes.

a) Crossover

The crossover operator enables the exchange of genetic information between two chromosomes. Common types of crossover operators include single-point crossover, two-point (double-point) crossover, and multi-point crossover. The study conducted by Pedergnana et al. (2013) involves the execution of double-point crossover. Once some chromosomes have been chosen for the population of the following generation, more chromosomes are produced by the implementation of a crossover operation. Crossover is a process that selects two parent people from the present generation based on a probability function described in Equation 2.19. It then combines pieces of both parents to form two offspring.

b) Mutation

The mutation operator is used to ensure genetic diversity among chromosomes by introducing local modifications, which helps preserve variation within the population. This process supports the exploration of optimal solutions within the search space. The number of chromosomes undergoing mutation depends on the defined mutation rate. During mutation, a gene is randomly selected from the target chromosome and its value is flipped—changing from '1' to '0' or from '0' to '1'.

Chapter 3
Description of Data

3.1. Study Area and Dataset

This section provides a detailed description of the methodology that has been utilized in this research.

3.1.1. Study Area

In the study, the lower Indus basin is taken as a study area. The total length of the Indus River is about 2800 km of which 2682 km is Pakistan. It can be divided into two parts: alluvial plain and deltaic plain. The alluvial plain covers an area of about 207,200 km^2 while the deltaic plain encompasses 200,000 km^2. The study lies between the following geographical coordinates; 26°21′N 68°51′E. Figures 3.1 and 3.2 depict the district and area maps of the lower Indus basin, respectively.

Floods are a commonly observed phenomenon in the Indus River basin. Between 1950 and 2022, around 21 floods have been generated in the Indus basin, producing a drastic direct loss of about $19 billion (in 2010 dollars) also killing 8887 people and destroying and damaging a total of 109,822 villages. Monsoon rainfall is the major source of flood generation in the Indus basin. Other factors also involve the size and shape of land use of catchments and the conveyance capacity of the streams. Monsoon rainfall is generally intense, widespread and continues from June to September (Ali, 2011). A shift in the monsoon precipitation has been observed in the country's central and southern parts during the last decades. Nanditha et al. (2023) asserted that the flood of 2022 was caused due to the depression that formed on the Arabian Sea, which led to heavy monsoon precipitation in the southern regions, particularly Baluchistan and Sindh, of Pakistan.

Hence, these changing environmental conditions emphasize the need to manage this area of the lower basin in the long run for better planning and relief from future flood disasters that may hit this area.

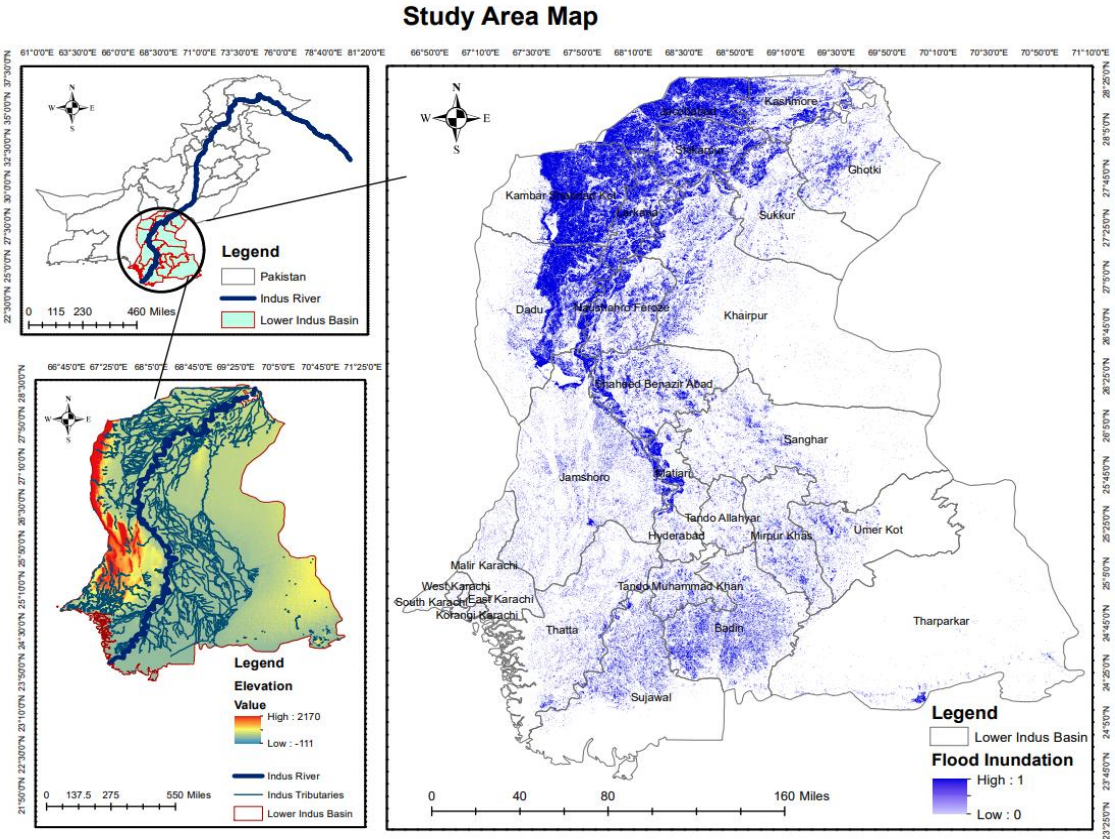


Figure 3. 1: Districts in Lower Indus Basin

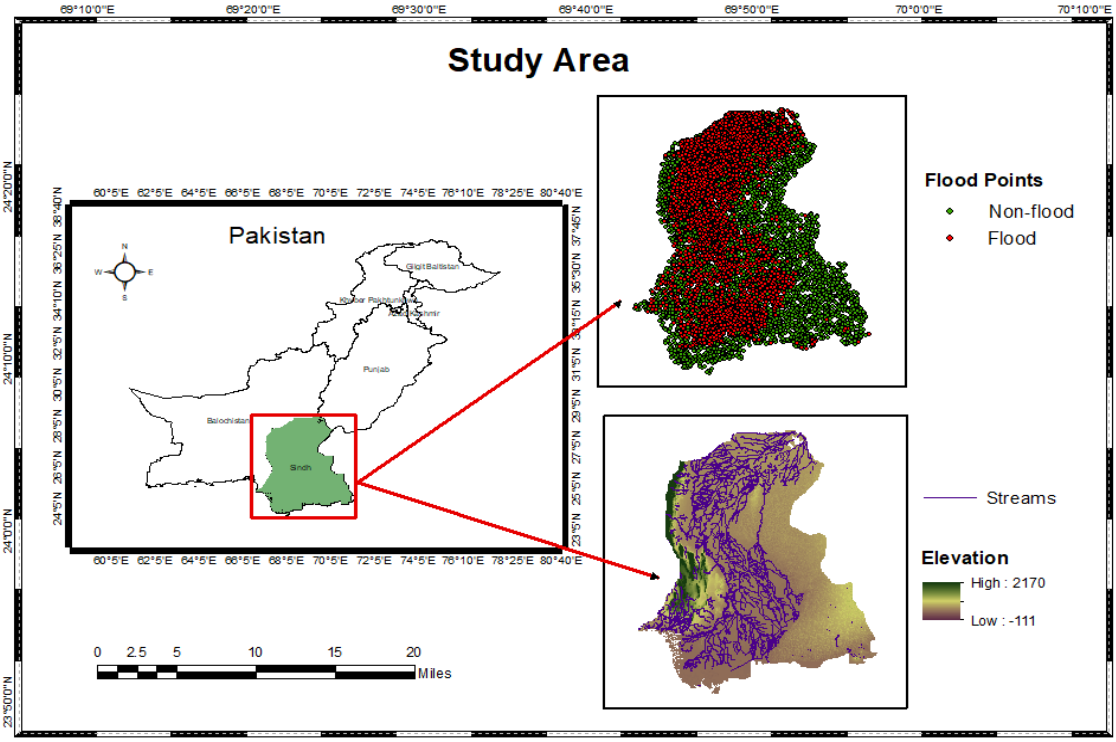


Figure 3. 2: Study Area Map

3.1.2. Flood Conditioning Factors

Preparing a spatial database and the selection of pertinent flood factors is a crucial stage in flood analysis (Papaioannou et al., 2015). Feature selection for floods is an intricate process that involves identifying various flood conditioning factors. The likelihood of floods is influenced by diverse factors, primarily dictated by the unique characteristics of each watershed (Bui et al., 2016a). Hence, accurate flood mapping in any area requires a thorough understanding of the specific influencing factors for that watershed (Chapi et al., 2017). This research relies predominantly on topographic, geo-environmental, and human-induced datasets. The details regarding the data along with their source are depicted in Table 3.1.

While different researchers have emphasized various factors but some common factors have been highlighted in flood analysis, as pointed out by studies such as Tehrany et al. (2015), Rahmati et al. (2015), Blistanova et al. (2016), and Ali et al. (2019). In hydrology and geospatial analysis, a consensus has not yet been made on the choice of variables (Costache et al., 2020). In the thesis, the most relevant flood-associated variables, that are lower Indus basin specific, are chosen by employing all the important factors considered in the recent literature for flood analysis. Moreover, the thesis is mainly concerned with hydraulic analysis of riverine floods and their prediction in Sindh province. For the purpose of flood susceptibility and hazard analysis, various researches have utilized different variables, some have focused solely on environmental factors such as Nanditha et al. (2023) and Pradhan (2015) and others have focused on topographic or anthropogenic factors such as Costache et al (2020), Bui et al. (2018), Khosravi et al. (2018), Mahmood et al (2019), Mukharjee and Singh (2020). All the variables are integral for flood analysis and their pertinence is explained in literature. Nevertheless, all the environmental, topographic and anthropogenic factors have not been utilized collectively in the previous literature. Thus, in the thesis, a dataset composed of 21 variables, including environmental, topographic and human-induced factors, is employed for flood analysis in connection with the major research goal i.e. susceptibility and hazard analysis of riverine flood. Furthermore, to the best of our knowledge, literature has not provided the best and the most representative dataset that can be employed for hydraulic analysis. Hence, the feature selection is employed, in the thesis, to select the most representative subset of data that can be used for conducting flood analysis in Sindh as this gap has not been explored yet. Moreover, the explanatory power of each variable, its relationship with flood occurrence along with its source and method of data generation is described in detail in table 3.1.

Table 3. 1: Flood Conditioning Factors

Conditioning Factors	Association between Conditioning Factors and Flood	Source of Conditioning Factors	Legend/Classification	Legend/ Classification Method
Topographic Factors				
Aspect	Aspect has a significant role in flood analysis as it assists in determining the direction of water flow (Tehrany et al. 2015).	Digital Elevation Model (DEM 30m) Advance Spaceborne Thermal Emission and Reflection (ASTER) as Global Elevation Model (GDEM)	(i) Flat (ii) North (iii) East (iv) South (v) West	Manual
Curvature	The curvature provides valuable insights into its geomorphological features (Paul et al., 2019). The curvature of the terrain significantly impacts the flow of floodwaters, making its study essential for accurate flood modeling.		(i) Concave (-1) (ii) Flat (0) (iii) Convex (+1)	Manual
Stream Power Index	The Stream Power Index (SPI) assesses erosive power and discharge concerning a specific area (Mukherjee and Singh, 2020). The SPI formula is given by; $$SPI = A \times tan(\beta)$$ where, A represents an area of the specific basin, and β signifies the local slope gradient measured in degrees.		(i)-13.81- -11.46(ii)-11.47- -10.46 (iii) -10.47- -8.11, (iv)-8.12- -2.59 (v) -2.60- -10.39	Quantile
Slope (Degree)	The slope map represents the effect of gravity on surface runoff formation and its speed (Mojaddadi et al., 2018). This factor holds significant importance in flood analysis.		(i) 0- 0.59 (ii) 0.60-0.65 (iii) 0.66 – 1.25 (iv) 1.26-7.74 (v) 7.75-78.14	Geometric Interval
Stream Density	Stream density is a vital factor in floods as it directly influences peak discharge rates (Ureta et al., 2020). Increased stream density corresponds to a higher likelihood of flooding, as indicated by Mojaddadi et al., 2018. For this research, the map was created from DEM data with a 30 m resolution. It was calculated by Rahman et al. (2021), as follows; $$SD = \frac{Stream\ Length}{Basin\ Area}$$		(i)0.001-3.37 (ii) 3.38-6.88 (iii) 6.89-10.66 (iv) 10.67-15.66 (v) 15.67-34.43	Quantile
Sediment Transport Index	STI value directly affects the potential for flow accumulation and thus plays a significant role in the flooding process (Rahman et al., 2021). $$STI = \left(\frac{\left(\frac{F_\alpha}{\delta_x}\right)^2}{\left(\frac{S_\alpha}{\delta_Y}\right)^2} \right)$$ Where, F_α represents the flow accumulation, $S_\alpha$is the slope raster, and δ_x and δ_Y are constants.		(i)0 (ii) 0-11.79 (iii) 11.80-23.59 (iv)23.60-47.18 (v)47.19-3007.97	Natural Breaks
Topographic Wetness Index (TWI)	TWI determines the wetness conditions across the stream basin area which may directly influence the occurrence of floods in the region. The calculation of TWI followed the method outlined by Beven and Kirkby in 1979; $$TWI = ln\left(\frac{\partial}{tan\beta}\right),$$ Here, ∂ denotes the cumulative upslope area drainage through the point per unit contour length, while $tan\beta$ represents the slope angle at that specific point.		(i)1.89-6.82 (ii)6.83-7.66 (iii)7.67-8.51 (iv)8.52-10.28 (v)10.29-21.53	Quantile

Topographic Ruggedness Index (TRI)	TRI quantifies the elevation variation between neighboring cells in a digital elevation grid. This method calculates the elevation variance from a central cell and the surrounding eight cells by squaring each of these variances to ensure they are all positive. The average of these squared values is then computed. A lower TRI value signifies a flatter terrain surface, while a higher value indicates an exceptionally rugged surface. Areas with lower TRI values and flat terrains are comparatively more susceptible to flooding than regions with higher TRI values.		(i)0.11-0.33 (ii)0.34-0.44 (iii)0.45-0.55 (iv)0.56-0.66 (v)0.67-0.88	Quantile
Valley depth	The depth of valleys plays a significant role in weathering processes, transportation, and water accumulation, thereby influencing the likelihood of floods (Tien Bui et al., 2016a).		(i)-144.69-112.72 (ii)112.73-210.23 (iii)210.24-315.55 (iv)315.56-448.16 (v)448.17-849.90	Natural Breaks
Multiresolution Ridge Top Flatness (MrRTF)	MrRTF is used in terrain analysis. It characterizes the flatness of ridge tops at multiple spatial scales. This measure is valuable in flood modeling, where the terrain's flatness can influence water flow and flood susceptibility (Zsuzsanna Csatáriné Szabó, et al., 2020).		(i) 0-1.21 (ii)1.22-3.14 (iii)3.15-4.67 (iv)4.48-6.13 (v)6.14-7.94	Natural Breaks
Multiresolution Valley Bottom Flatness (MrVBF)	It quantifies the flatness of valley bottoms across multiple spatial scales and characterizes the flatness of valley bottoms, which is essential for flood modeling. It is calculated by analyzing the flatness of valley bottoms at different resolutions or scales. It tells about the landscape's characteristics, particularly in terms of water flow, drainage patterns, and flood susceptibility (Zsuzsanna Csatáriné Szabó, et al., 2020).		(i)0-1.18 (ii)1.19-3.18 (iii)3.19-5.45 (iv)5.46-7.56 (v)7.57-9.45	Natural Breaks
Openness positive	Topographic openness is computed as the average of either zenith (φ) or nadir (ψ) angles across eight azimuths (0, 45, 90, 135, 180, 225, 270, and 315) within a radial distance L (Yokoyama et al., 2002). Openness is always represented with a positive sign and falls within the range of 0 to 180 degrees. The terminology "positive" and "negative" are consistent with how terrain-slope curvature has been expressed (Pike et al., 1988). Positive openness φL denotes a convex-upward calculation using zenith angles (Yokoyama et al., 2002). Openness positive φL of an area within a radial distance L can be calculated as: $$\varphi L = \frac{0\varphi L + 45\varphi L + \cdots . + 315\varphi L}{8}$$		(i)0.71-1.36 (ii)1.37-1.46 (iii)1.47-1.51 (iv)1.52-1.54 (v)1.55-1.66	Natural Breaks
Openness negative	The negative openness ψL signifies a concave-upward evaluation using nadir angles (Yokoyama et al., 2002). The negative openness ψL within a radial distance L can be calculated as; $$\psi L = \frac{0\psi L + 45\psi L + \cdots . + 315\psi L}{8}$$		(i)0.69-1.30 (ii)1.31-1.49 (iii)1.50-1.55 (iv)1.56-1.60 (v)1.61-1.65	Natural Breaks
Human Induced Factors				

Population density	The likelihood of flooding rises with an increase in the population (Li et al., 2019). The data has been taken from 1990-2022.	Humanitarian Data Exchange	(i) < 15000 (ii) 15000-30000 (iii) 30000-450000 (iv) 45000-60000 (v) 60000-80000	Manual
Distance from Road (m)	Impervious surfaces like roads intensify the rainfall-runoff process (Mukherjee and Singh, 2020). Consequently, the distance from roads is a significant factor influencing flood. Proximity with roads is calculated by utilizing the Euclidean distance method with digital roads order as described by Mukherjee and Singh (2020).		(i) 0-10000 (ii) > 10000-20000 (iii) > 20000-30000 (iv) >30000 – 40000 (v) >40000-51217	Manual
Land Use Land Cover (LULC)	Land cover plays a pivotal role in evaluating flood susceptibility, particularly in areas with sparse vegetation, which are more susceptible to flooding. Urban areas, characterized by impermeable surfaces and barren lands, exacerbate surface runoff flow (Mojaddadi et al., 2018).	US Geological Survey, MODIS 12 Q1	(i) Savanas (ii) Grasslands (iii) Wetlands (iv) Croplands (v) Urban buildup lands (v) Natural Vegetation (vi) Barren lands (vii) Water bodies (viii) Shrublands	MODIS classifier
Normalized Difference Vegetation Index (NDVI)	NDVI is a pertinent factor that may affect floods (Paul et al., 2019). The mathematical calculation formula is, $$NDVI = \frac{Near\ Infrared\ Band - Red\ Band}{Near\ Infrared\ Band + Red\ Band}$$ Where, NIB = band 2 and Red = band 1. The NDVI ranges between -1 to 1.		(i) 0-0.004 (ii) > 0.004 – 0.008 (iii) > 0.008 – 0.012 (iv) > 0.012 -0.016 (v) > 0.016 – 0.020	Manual
Geo-environmental Factors				
Distance from River (m)	The probability of flood increases in areas near main channels; therefore, the proximity to rivers is a crucial factor in flood susceptibility assessment (Chowdhuri et al., 2020). The map is generated using the Euclidean distance from streams, relying on digital stream and river order data (Mukherjee and Singh, 2020).	Humanitarian Data Exchange	(i) 0 - 20000 (ii) > 20000-40000 (iii) > 40000-60000 (iv) >60000 – 80000 (v) > 80000- 100000	Manual
Rainfall (mm)	Rainfall and flood are positively correlated. The rainfall map is prepared using annual rainfall data (2001-2022) which was extracted by using the inverse distance weighting interpolation method (Ureta et al., 2020). The data utilized for this research is for the period of 2000-2022.	Climatic Research Unit database	(i)500 -1000 (ii) > 1000 – 2000 (iii) >2000 – 3000 (iv) > 3000 – 4000 (v) > 4000 - 6000	Manual
Temperature	Temperature and floods have a high positive correlation. In this study, the temperature map is prepared to utilize annual temperature data (2001-2022) which was extracted by using the inverse distance weighting (IDW) interpolation method (Ureta et al., 2020).	Climatic Research Unit database	(i) 2.59-26.6 (ii) 26.7-27.0 (iii) 27.1-27.2 (iv) 27.3-27.6 (v) 27.7-28.5	Manual
Lithology	Soil texture plays a vital role in flood occurrence as it regulates the rate of infiltration and surface runoff. Regions with subsoils that have high permeability and rocks with high resistance tend to experience minimal drainage (Ureta et al., 2020).	FAO soil map	(i) Clay (ii) Loam (iii) Silt Loam (iv) Clay Loam	FAO soil Classifier

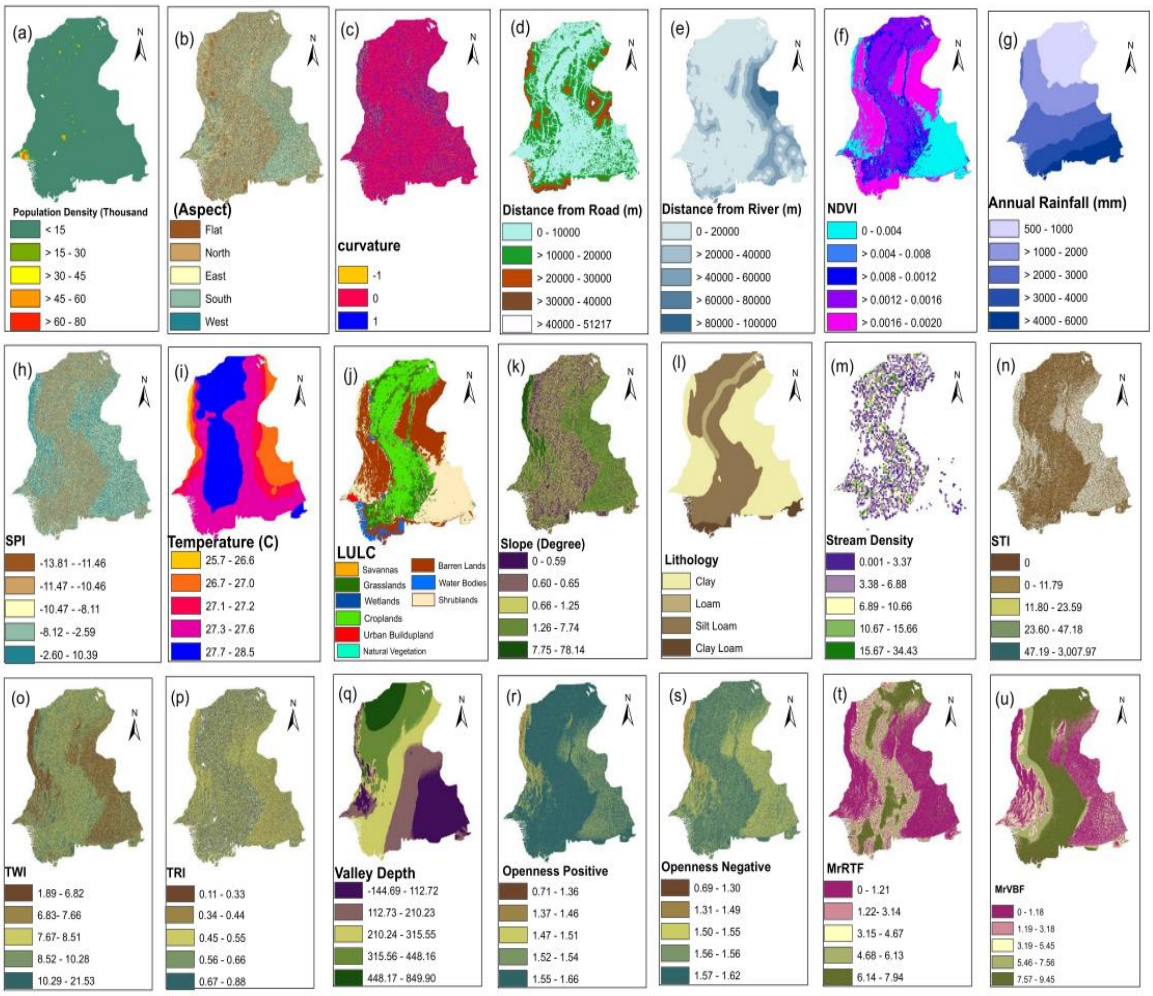


Figure 3. 3: Flood Conditioning Factors for 2022 Flood Analysis

For the analysis of flood 2010 and the predicted flood 2032, all the used independent variable data layers were same except for population density, rainfall, temperature, and LULC. These data layers were prepared by utilizing the same procedure and data sources as depicted in table 3.1. The temperature and rainfall datasets have been utilized for the time period of 1990-2010. The flood points for 2010 flood are also prepared by utilizing the same procedure. Again, for the purpose of analysis 5500 flood points and 5500 non-flood points are considered. The data layers are given in figure 3.4.

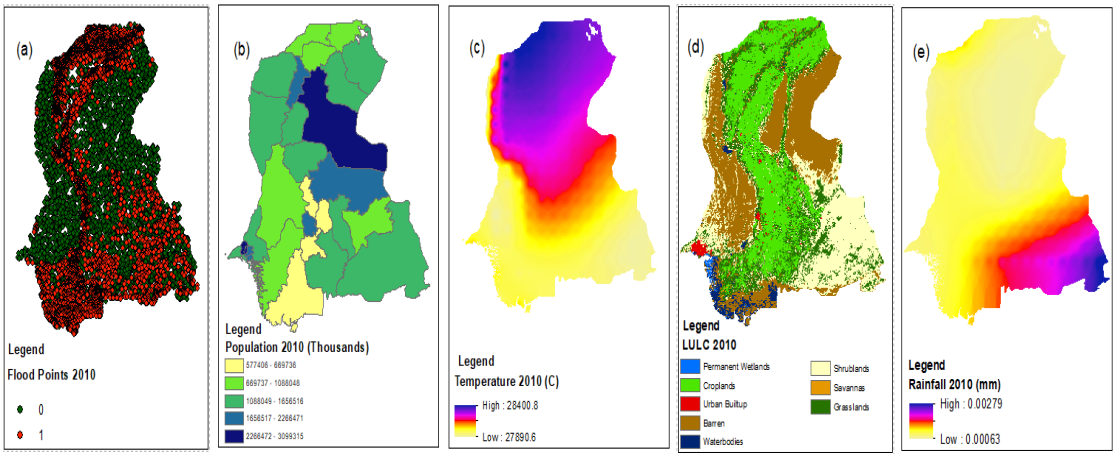


Figure 3. 4: Additional Data Layers Utilized for Analysis of Flood 2010

For making prediction for 2032 flood in the considered study area, again the same procedure has been followed. All the data layers utilized were same as that for flood 2022, the additional layers have been depicted in figure 3.5. The population density, rainfall, temperature and LULC datasets have been obtained from Pakistan Bureau of Statistics (PBS), CMIP6 database and Sentinel 1 database, respectively.

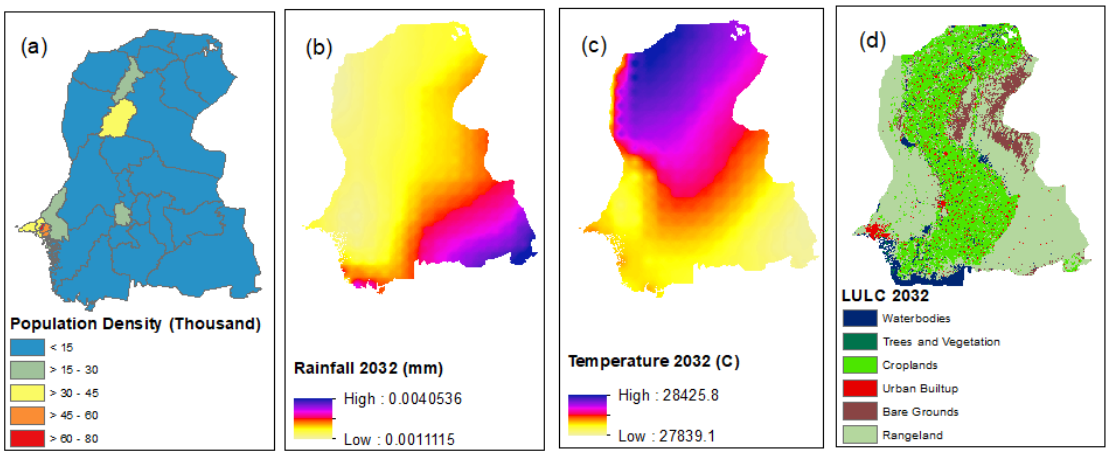


Figure 3. 5: Additional Data Layers Utilized for Analysis of Flood 2032

Chapter 4
Research Methodology

4.1. Introduction

Estimating flood damage is a crucial aspect of water resources planning, particularly for evaluating the advantages of flood prevention. Traditionally, flood control strategies prioritize the establishment of design standards and the implementation of structural solutions to mitigate floods. Traditionally, flood mitigation buildings were created with the purpose of managing a specific, predetermined design flood, which is determined by the frequency of the design rainfall. Recently, there has been a shift from the traditional structural flood control technique to a more advanced concept known as "flood risk management" (Merz et al., 2010). The level of protection is determined by factors that go beyond a pre-established design flood, with a greater emphasis on non-structural techniques to mitigate flooding. An important development observed in this context is a prevailing shift from flood danger to flood risk. Currently, flood policies are primarily concerned with regulating and minimizing flood hazards. This involves reducing the likelihood and severity of flood discharges and inundations (Merz et al., 2010). In the field of flood risk management, the term "risk" refers to the potential harm that may occur or above a certain threshold within a specific time frame, based on a specified probability. Therefore, it is crucial to thoroughly assess and consider the various factors that contribute to damage when managing flood risks.

This research is an endeavor to evaluate the economic damages caused by floods. For this purpose, a novel hydraulic modelling has been done in this research by using metaheuristic algorithms and machine learning models. For this purpose, a spatial and temporal analysis has been done to analyze district-level flood vulnerability by considering geospatial heterogeneities in the lower Indus basin. Two types of methodologies are commonly used for spatial data analysis: one is traditional methodologies and the other is advanced set of tools like machine learning and deep learning tools. As the underlying dataset is so complicated and contains a huge set of covariates, thus we have applied the machine learning algorithms for handling such data efficiently. It is fact that advanced tools application lead to more robust conclusion and policy implications in contrast to traditional methodologies. Referring to the econometric techniques, different machine learning models are utilized and compared which can be used for geospatial dataset and spatial analysis. The spatial dataset includes topographic, human induced and geo-environmental factors specific to each region in the Sindh province.

Moreover, the dataset is heterogenous in nature and is dependent on the geo-referenced locations. Furthermore, spatial analysis is pertinent as the current research takes flood as an endogenous factor causing economic damages rather than exogenous factor.

The chapter consists of following sections, section 1 discusses the pre-processing of the data layers, section 2 illustrates the feature selection method used for this research, section three provides the methodology utilized for classification modelling, section five is about the sensitivity analysis of the flood conditioning factors, section 6 depicts the gain ratio analysis, section seven demonstrates the frequency ratio model for flood factors, section seven is about flood susceptibility mapping and forecasting, section eight discusses the method used to assess correlation between flood and LULC, and the last section discusses the research method used to calculate the monetary damages of floods.

4.2. Data Layers Pre-processing

4.2.1. Preparation of Spatial Database

The data processing occurred in ArcGIS 10.3.1, resulting in a database structured as a matrix with 15391 columns and 17762 rows, representing a spatial grid of cell size 30×30 meters. The study area experienced a disastrous flood in 2022 and 2010. Thus, in this research, flood and non-flood points from the year 2022 and 2010 are utilized, separately. To enhance accuracy, for both flood episodes, 5500 flood points and an equal number of non-flood points, to prevent data biases, were digitized within the study area. Non-flood locations, predominantly elevated areas were digitized using detailed topographic maps. The flood and non-flood samples were split into training (70%) and validation (30%) sets. Therefore, the training set comprised 3850 flood and 3850 non-flood points, while the validation set included 1650 flood and 1650 non-flood points. This division, as indicated in existing literature (Tehrany et al., 2015; Bui et al., 2020; Costache et al., 2020; Wang et al., 2021) is crucial. It allows the training of the models using the training sample, while the validation sample is utilized to confirm the results provided by all models.

4.2.2. Normalization of Data Layers

The initial flood factors possess both data types i.e. categorical and continuous. Hence, it is advisable to utilize numerical values during the training phase of machine learning models. The data layers were normalized between 0.1 and 0.9 in order to be used as input data for the

models. For normalization of the data layers, the following formula has been employed through the raster calculator tool in Arc GIS;

$$y = \frac{(x-\min(d))\times(\max(n)-\min(n))}{\max(d)-\min(d)} + \min(n) \quad (4.1)$$

The parameters in this equation are defined as follows: x represents the current value, y represents the standardized value of x, d shows limit of range value, and n illustrates the standardized limit of range.

4.3. Feature Selection

Various techniques and methodologies have been utilized by previous researchers for feature selection. The flowchart of the proposed methodology is portrayed in Figure 4.1. Firstly, the flood conditioning factors were obtained from the Digital Elevation Model and vector databases and the flood inventory (flood and non-flood points) were divided into training and testing samples. The dataset was tested for econometric problems of multicollinearity, heteroscedasticity, and spatial autocorrelation. The prevalence of these issues in the dataset demonstrated the need for feature selection. Thus, the metaheuristic algorithms were utilized to scrutinize the dataset. These metaheuristics were further tested for performance accuracies by hybridizing each algorithm with SVM and KNN. The highest performance accuracy determined the most appropriate and relevant dataset for hydraulic modeling.

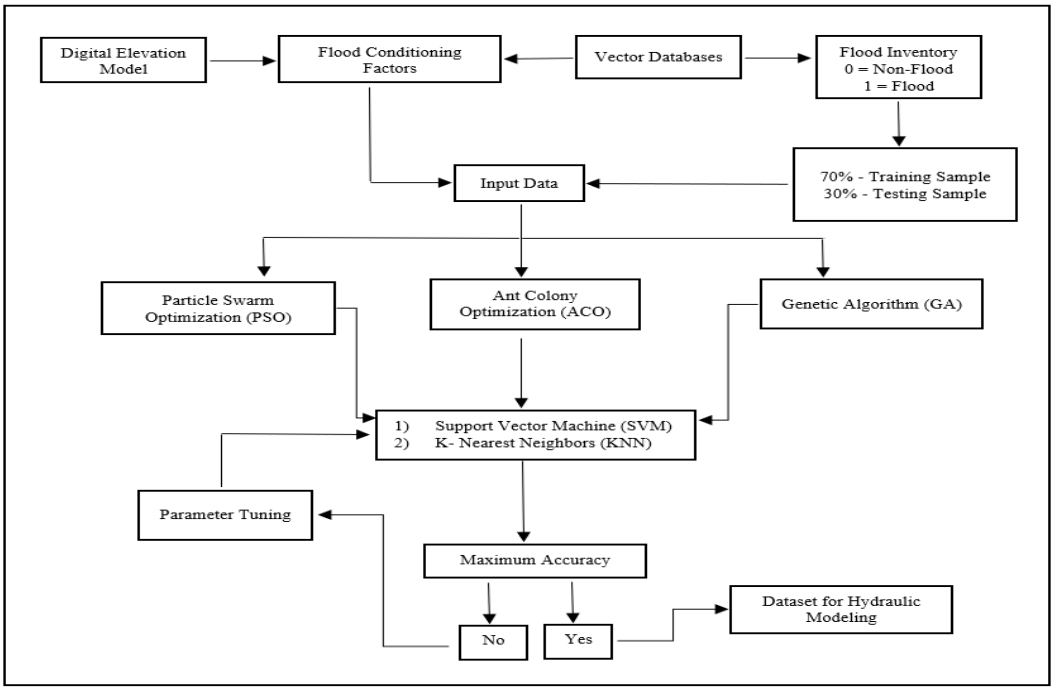


Figure 4. 1: Scheme of Methodological Workflow

4.3.1. Tests for Presence of Multicollinearity, Heteroscedasticity and Autocorrelation

When a study incorporates numerous independent elements, the existence of collinearity might have a substantial impact on the ultimate findings. Hence, it is imperative to determine the multicollinearity among such parameters (Arabameri et al., 2019; Wang et al., 2021). The correlation matrix is utilized in this research to identify independent variables that exhibit multicollinearity. For this work, twenty-one conditioning factors were examined for analysis. Therefore, it is crucial to assess their multi-collinearity. To assess the prevalence of heteroscedasticity and autocorrelation tests, a simple regression analysis was conducted.

4.3.2. Metaheuristics Algorithms and their Parameter Tuning

In the present study, heuristics are employed for feature selection as our dataset had econometric issues of high multicollinearity, heteroscedasticity and spatial autocorrelation which is discussed in detail in the analysis section. Following discussion provides the parameter tunings of metaheuristics.

4.3.2.1. Particle Swarm Optimization (PSO)

The basic update rule for position and speed are depicted in equations (1) and (2), respectively.

$$\varphi_i\,(t+1) = \varphi_i + \vartheta_i(t+1) \tag{4.2}$$

$$\vartheta_i(t+1) = \omega\vartheta_i(t) + c_1 r_1(\rho_i - x_i) + c_2 r_2(g - x_i) \tag{4.3}$$

Here, ω is the inertia weight constant, c_1 is the cognitive constant and c_2 is the social learning constant, likewise, r_1 and r_2 denote random numbers, respectively, ρ_i is the best position of particle *i* and lastly, g is the global best position among all the particles in the swarm.

The simulations of PSO algorithm were performed on R 4.3.0 using the Pso library. The parameter settings are; swarm population sizes (s): 10, 15, 20, 25, 30, 50, 70, 100, 150, and 200, maximum iterations (maxit): 1000, and random seed: 123.

4.3.2.2. Ant Colony Optimization (ACO)

In ACO, a probabilistic transition rule is utilized for selection of the most informative features on the currently selected features. (Basset and L. A. F, 2018). It represents likelihood of an ant to choose to travel from feature i to feature j at the time t:

$$p_{ij}^{k}(t) = \frac{[\tau_{ij}(t)]^{\alpha} \times [\eta_{ij}]^{\beta}}{\Sigma_{l \varepsilon J_i^k} [J_{il}(t)^{\alpha} \times [\eta_{il}]^{\beta}} \quad (4.4)$$

Wherein, *k* shows the number of ants, J_i^k denotes the set of ant *k*'s unvisited features, η_{ij} shows heuristic desire of the chosen feature *j*, when it is at feature *i*, and $\tau ij\ (t)$ illustrates the amount of virtual pheromone on edge1(i, j). Choice of α and β is usually controlled experimentally.

The ACO algorithm simulations were conducted on R 4.3.0 using FsinR and class libraries. The parameter settings are; ant colony population sizes (k): 10, 15, 20, 25, 30, 50, 70, 100, 150, and 200, maximum iterations (maxit): 1000, random seed: 123, α = 0.8, and β=1.

4.3.2.3. Genetic Algorithm (GA)

The simulations were conducted on R 4.3.0 using the GA library. The parameter settings are; population sizes: 10, 15, 20, 25, 30, 50, 70, 100, 150 and 200, maximum iterations (maxit): 1000, random seed: 123, mutation: 0.01, crossover probability: 0.9, selection: gabin_lrSelection and number of bits: 22.

4.3.3. Kernel Functions for SVM

The kernel functions utilized in this research for formulation of hybrid models with SVM are represented as follows:

4.3.3.1. Linear Kernel

$$K(i,j) = x_i . x_j \quad (4.5)$$

4.3.3.2. Radial Basis Function (RBF) Kernel

$$K\ (i,j) = \exp\ (-\gamma\ \left|x_i - \ x_j\right|^2) \quad (4.6)$$

4.3.3.3. Sigmoid Kernel

$$K\ (i,j) = \tanh(\gamma(x_i\ .\ x_j) + r) \tag{4.7}$$

4.3.3.4. Polynomial Kernel

$$K\ (i,j) = (\ \gamma\ (x_i\ .\ x_j) + r\ )^d \tag{4.8}$$

In the above-mentioned equations, γ, d, and r denote the kernel parameters. γ portrays the parameter of RBF and sigmoid kernels whereas d is the functional parameter of the polynomial kernel, while r denotes the residual.

4.3.4. Hybridization of metaheuristics algorithms with SVM and KNN

The flood conditioning factors were incorporated into the spatial database, which was then analyzed using GA-SVM, PSO-SVM, ACO-SVM, GA-KNN, PSO-KNN, and ACO-KNN to assess the most appropriate variables in the causation of flood in the considered study area. Chapter 3 discusses the main sources utilized to construct the spatial layers in this study. Within GIS, each flood conditioning factor was transformed into a grid database, ensuring a uniform spatial resolution of 30x30 meters.

4.4. Decision Trees Modelling

The variables chosen by each metaheuristic algorithm (PSO, ACO, and GA) are subsequently utilized for modelling and contrasted based on performance using various performance indicators. The classification (modelling) process utilizes four base classifiers, namely RF, NBT, LMT, and REPT, together with an ensemble of these classifiers by using a joint bagging-boosting technique. Each decision tree model and its ensemble were evaluated using performance criteria including true positive, true negative, false positive, false negative, sensitivity, specificity, precision, accuracy. Following the training of these models, which involved adjusting their hyperparameters and optimizing them, a performance comparison was conducted by evaluating the differences in sensitivity, specificity, precision, and accuracy. The study was conducted on both the training and validation datasets. The training dataset (70%) was used to assess the model's fitting, while the validation dataset (30%) was used to evaluate the model's generalization capacity. Khwaja et dal. (2020) developed the integarted bagging and boosting technique and applied it for electric load forecasting using the neural networks. The modification of our research is that we have utilized joint bagging boosting technique for

formulating decision trees ensemble models for forecasting flood. Hence, the DGP for forecasting flood using joint bagging boosting is described below.

4.4.1. Training of Bag-boost Hybrid

In the flood forecasting issue, we have used a training data set that includes known previous features X_{tr} and their related flood points Y_{tr}. X_{tr}has a size of $N_F \times N_T$, while Y_{tr} has a size of $N_T \times 1$. The objective is to find a function f(.) that best matches the training data as follows;

$$f: X_{tr} \rightarrow Y_{tr} \tag{4.9}$$

To reduce variance of the projected outcomes, we approximate f(.) by an ensemble of N_{bag} models, denoted as $\mathcal{F}_{bag}$, where the i^{th} model in the ensemble is trained on a bootstrapped data set (X_{tr}^{i}, Y_{tr}^{i}) of the original training data. The bootstrapped dataset size is given by $N_F \times N_{samples}$ for $(X_{tr}^{i}$ and $N_{samples} \times 1$ for Y_{tr}^{i}, where $N_{samples}$ is the number of samples in the set. It can be obtained by randomly sampling with replacing the original data set.

$$\mathcal{F}_{bag} = \bar{f}_1, \bar{f}_2, \dots\dots, \bar{f}_{N_{bag}} \tag{4.10}$$

In the traditional approach, the i^{th} model $\bar{f}_i$ in the ensemble is determined by reducing the error described by the following equation:

$$\min_{\bar{f}_i} \| Y_{tr}^{i} - \bar{f}_i(X_{tr}^{i}) \|_2^2, \quad i = 1, 2, \dots., N_{bag} \tag{4.11}$$

By simultaneously training the models in the ensemble described by equation (4.10) and averaging their output, it is possible to decrease the variance of prediction error. Bagging does not effectively reduce prejudice. In order to address this problem and create individual models inside the ensemble that are sufficiently accurate, we suggest representing each model as an additional ensemble of models. Therefore, equation (4.10) is adjusted as follows:

$$\bar{\mathcal{F}}_{bag} = \bar{\mathcal{F}}_1, \bar{\mathcal{F}}_2, \dots\dots, \bar{\mathcal{F}}_{N_{bag}} \tag{4.12}$$

where the i_{th} model $\bar{\mathcal{F}}_i$ further comprises of an ensemble of models, i.e.,

$$\bar{\mathcal{F}}_i = \tilde{f}_i^1, \tilde{f}_i^2, \dots\dots, \tilde{f}_i^{N_{boost}} \tag{4.13}$$

Boosting is employed to create decision trees (DT) models inside each ensemble model, therefore reducing both bias and variance. Each decision tree (DT) model in the ensemble is trained to reduce the discrepancy between the training data. The following equations represent the training process for each DT model in the proposed ensemble.

$$\tilde{f}_i^1 : X_{tr}^i \rightarrow Y_{tr}^i$$

$$\tilde{f}_i^2 : X_{tr}^i \rightarrow Y_{tr}^i - \alpha \tilde{f}_i^1(X_{tr}^i)$$

$$\vdots$$

$$\tilde{f}_i^{N_{boost}} : X_{tr}^i \rightarrow Y_{tr}^i - \alpha \tilde{f}_i^1(X_{tr}^i) - \alpha \tilde{f}_i^2(X_{tr}^i) \dots\dots - \alpha \tilde{f}_i^{N_{boost}-1}(X_{tr}^i) \quad (4.14)$$

The weight value α, ranging from 0 to 1, guarantees the error gradually diminishes in every model. The concept revolves around the notion of gradually adding more models, as more models are added, the entire model becomes a more robust predictor. The estimation of the boost in the final model $\tilde{f}_i^{N_{boost}}$ is effectively achieved by minimizing the mentioned below error:

$$\min_{\tilde{f}_i^{N_{boost}}} \| Y_{tr}^i - \alpha \tilde{f}_i^1(X_{tr}^i) - \alpha \tilde{f}_i^2(X_{tr}^i) \dots\dots - \alpha \tilde{f}_i^{N_{boost}-1}(X_{tr}^i) - \alpha \tilde{f}_i^{N_{boost}}(X_{tr}^i) \|_2^2 \quad (4.15)$$

4.4.2. Model Performance Evaluation Statistics

The first step in verifying results and evaluating model performance involves calculating statistical metrics such as sensitivity, specificity, accuracy, and precision. These indicators are widely used in previous studies to assess vulnerability to natural hazards (Costache et al., 2020). These metrics reflect the model's effectiveness in classifying flooded and non-flooded pixels. Baratloo et al. (2015) evaluated the model's capability to accurately classify these two pixel types. They discussed that specificity indicates how accurately the model identifies non-flooded grid cells, while sensitivity measures its accuracy in detecting flooded grid cells. Moreover, accuracy represents the proportion of correctly classified grid cells, both flooded and non-flooded. The following equations can be used to calculate these three indicators:

$$Sensitivity = \frac{TP}{TP+FN} \quad (4.16)$$

$$Specificity = \frac{TN}{FP+TN} \quad (4.17)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4.18)$$

$$Precision = \frac{TP}{TP+FP} \quad (4.19)$$

Where: TP (true positive) refers to total grid cells that are correctly categorized, while TN (true negative) represents the total grid cells that are also accurately classified. On the other hand, FP (false positive) indicates the number of grid cells that are wrongly classed as positive, and FN (false negative) represents the grid cells that are wrongly classified as negative.

The second phase of result verification is finalized by constructing the ROC curve, which plots the sensitivity on the Y-axis against the 1-specificity on the X-axis. The success rate of the model's performance is computed by utilizing a training dataset that consists of samples from 70% flooded and non-flooded grid cells. The remaining 30% of the grid cells are utilized to derive a prediction rate that emphasizes the precision of the outcomes. Both scenarios rely on the area under the curve (AUC) as a statistical metric to determine the performance of the models. An AUC number close to 1 indicates a highly efficient model, while an AUC value close to 0 indicates a non-informative model (Shafizadeh-Moghadam et al., 2018). The AUC value can be calculated using the following equation:

$$AUC = \frac{(\Sigma TP+\Sigma TN)}{(P+N)} \quad (4.20)$$

Where; P represents the overall count of flood pixels, while N represents the overall count of non-flood pixels.

4.4.3. Tests for Statistical Significance

For a detailed comparison among classifier ensemble schemes, statistical tests are employed to show that the difference between the classifiers is significant (Garcia et al., 2010). In this regard, non-parametric Friedman test will be utilized. As the null hypothesis of Friedman test is rejected, so the study has conducted a post-hoc test using Nemenyi test to determine that our classifiers are significantly different (J. Demsar, 2006, Parra et al., 2023).

4.4.4. Sensitivity Analysis of the Amount of Flood Data for Generating Training and Testing Dataset

When conducting susceptibility modelling research, it is necessary to divide the landslide data into two separate segments in order to create training and testing datasets. This division is crucial for validating the models. (Chung and Fabbri, 2003). There is a lack of consensus on the optimal amount of geospatial data to use when creating datasets. Pradhan (2013) used a 50/50 split of geospatial data for both training and testing datasets, while Lee and Oh (2012) used a split of 70/30 for training and testing datasets. This study conducted a sensitivity analysis to assess the impact of the quantity of flood data on modelling outcomes. It aimed to determine the optimal ratio for splitting the flood data to generate datasets. Initially, the hydraulic data was divided into several proportions with a 10% interval. Subsequently, the several datasets are created by employing the aforementioned data splitting technique. Subsequently, our bag-boost ensemble model is created and verified using various produced datasets. The AUC values are computed and utilized to validate the efficacy of the bag-boost ensemble model.

4.4.5. Gain Ratio

The Gain Ratio method, a widely used technique for feature selection, was introduced by Quinlan in 1993. The Gain Ratio approach has been employed in this study to assess and choose the appropriate factors that contribute to flood (Pham et al., 2018). In this research, Gain Ratio method has been utilized to analyze the predictive capability of the flood influencing factors by utilizing the stand-alone decision tress models and bag-boost ensemble model. The Gain Ratio of a given training data set, with respect to the class property C, is defined as follows (Quinlan 1993);

$$Entrophy = -\sum_{i=1}^{k} p(c_i)\, log_2(p(c_i)) \tag{4.21}$$

$$Gain\ Ratio\ (f, C) = \frac{Gain(f,C)}{SplitInfo(f,C)} \tag{4.22}$$

Entropy is the level of uncertainty regarding the value of the class attribute C. The probability of C being equal to ci is denoted by $p(c_i)$. Split Info refers to the information that may be gained by dividing the training data f into subsets f_1, f_2, ..., f_m, based on the attribute C. Split Info is computed by a specific formula:

$$Split\ Info(f, C) = -\sum_{j=1}^{m} \frac{|f_j|}{|f|} log_2 \frac{|f_j|}{|f|} \tag{4.23}$$

4.4.6. Spatial Relationship between Flood and Conditioning Factors by Using Frequency Ratio Model

The Frequency Ratio (FR) method is one of the most widely accepted and frequently applied bivariate statistical techniques due to its simplicity and practicality (Ali et al., 2019; Pradhan, 2010). The FR model is based on identifying spatial relationships between dependent and independent variables. Previous studies have shown that FR is a reliable approach for flood susceptibility modelling, effectively used to examine the correlation between selected variables (Rahmati et al., 2016a; Samanta et al., 2018).

In this thesis, the objective is to assess the relationship between flood occurrence points (training datasets) and flood conditioning factors (a set of 14 selected variables). Thus, the flood points are taken as dependent variable, while the conditioning factors are considered independent variables. The FR weights are calculated by determining the ratio between the number of pixels with flood points within each class of a given factor and the total number of pixels in the study area. For variables with continuous or numerical values, the Natural Breaks classification method was applied. The FR coefficients are computed using equation 4.17 as described in the works of Ali et al. (2019), Costache (2020), and Youssef et al. (2015).

$$FR = \frac{\frac{Qpix_{X_i}}{\Sigma pix_{Y_i}}}{\frac{\Sigma Ppix_{X_j}}{\Sigma Tpix_{Y_j}}} \tag{4.24}$$

The Frequency Ratio (FR) of class i and j is calculated as the ratio between the number of pixels containing flood points in class Q ($Qpix_{Xi}$) and the total number of pixels having class Q over the study area ($\sum pix_{Yi}$). Similarly, the total number of pixels containing flood points in class Q ($\sum Ppix_{Xj}$) is divided by the total number of pixels over the study area ($\sum Tpix_{Yj}$).

If the value of FR is larger than 1, it indicates a positive correlation between the class of conditioning variables and training points, suggesting a high vulnerability to floods. However, a value less than 1 suggests a lack of substantial correlation and a reduced vulnerability to floods (Ali et al., 2019; Rahmati et al., 2016a; Samanta et al., 2018).

4.5. The Flood Susceptibility Forecasting

The flood susceptibility maps for 2010, 2022 and 2032 have been created in this research using the bag-boost ensemble model. The 14 variables selected by PSO algorithm are used as

independent variables and flood points. The maps thus produced are categorized on five flood intensities as: (1) extremely high (2) high (3) moderate (4) low (5) extremely low. The susceptibility maps of 2010 and 2022, annual daily data for rainfall and temperature has been utilized with the periods of 1990-2010 and 2000-2022, respectively. The population density rasters are employed for the period of 1990-2022.

For the generation of flood susceptibility map for 2032, flood extent was forecasted based anticipated changes in the variables; rainfall, temperature, population density, and LULC. These variables are forecasted for the year 2032. The simulated annual daily rainfall and temperature forecasts are derived from Representative Concentration Pathways (RCPs) 2.6. The RCP2.6 pathway represents a situation in which there is a determined and rapid reduction in carbon emissions, reflecting significant climate action. Based on this situation, it is probable that the increase in global warming can be restricted to a maximum of 2 °C above the average surface temperature before industrialization. This aligns with the goals set in the Paris Agreement (IPCC 2014).

The population density forecasting is conducted by using Compound Growth Exponential Regression Model (CGERM) simulation and data from 1960 till 2022 has been utilized in this regard (Islam et al., 2023). The predictions are performed by using the following equation;

$$Y_p = [Y_c(1 + B)^n] \quad (4.25)$$

Where, Y_p is the value of the response variable at the projected time, Y_c denotes the actual/ collected value of the response at time "t", "B" is the regression slope of the line or regression coefficients, "n" is the total number of years (projection horizon), i.e. $t_p - t_c$.

The estimation of future LULC exposure is conducted by utilizing the GIS-extension MOLUSCE to simulate the anticipated changes in land use within the research area. The land use changes were determined by using the Sentinel 1 (with 10m resolution) land covers data between 2010 and 2022. The research area utilized two land covers and a Digital Elevation Model (DEM) raster to train an artificial neural network in MOLUSCE for the purpose of identifying potential future land use transitions in 2032. Once the artificial neural network (ANN) had been trained, MOLUSCE was used to model future land use through a cellular automata (CA) simulation for 2032. The MOLUSCE model is executed with both single and

double simulation iterations to produce land use raster dataset for 2032. The land use simulation has been conducted using QGIS 2.18.2 due to the incompatibility of the MOLUSCE GIS-extension with ArcGIS. The projected land use raster dataset has been reclassified and resampled, in Arc GIS for generating flood susceptibility for 2032.

4.6. Correlation between Flood and Land Use Land Cover (LULC)

Two episodes of land use land cover and floods were utilized to assess the correlation between LULC and flood. In this context, the LULC and flood of 2010 and 2022 were employed for the study region.

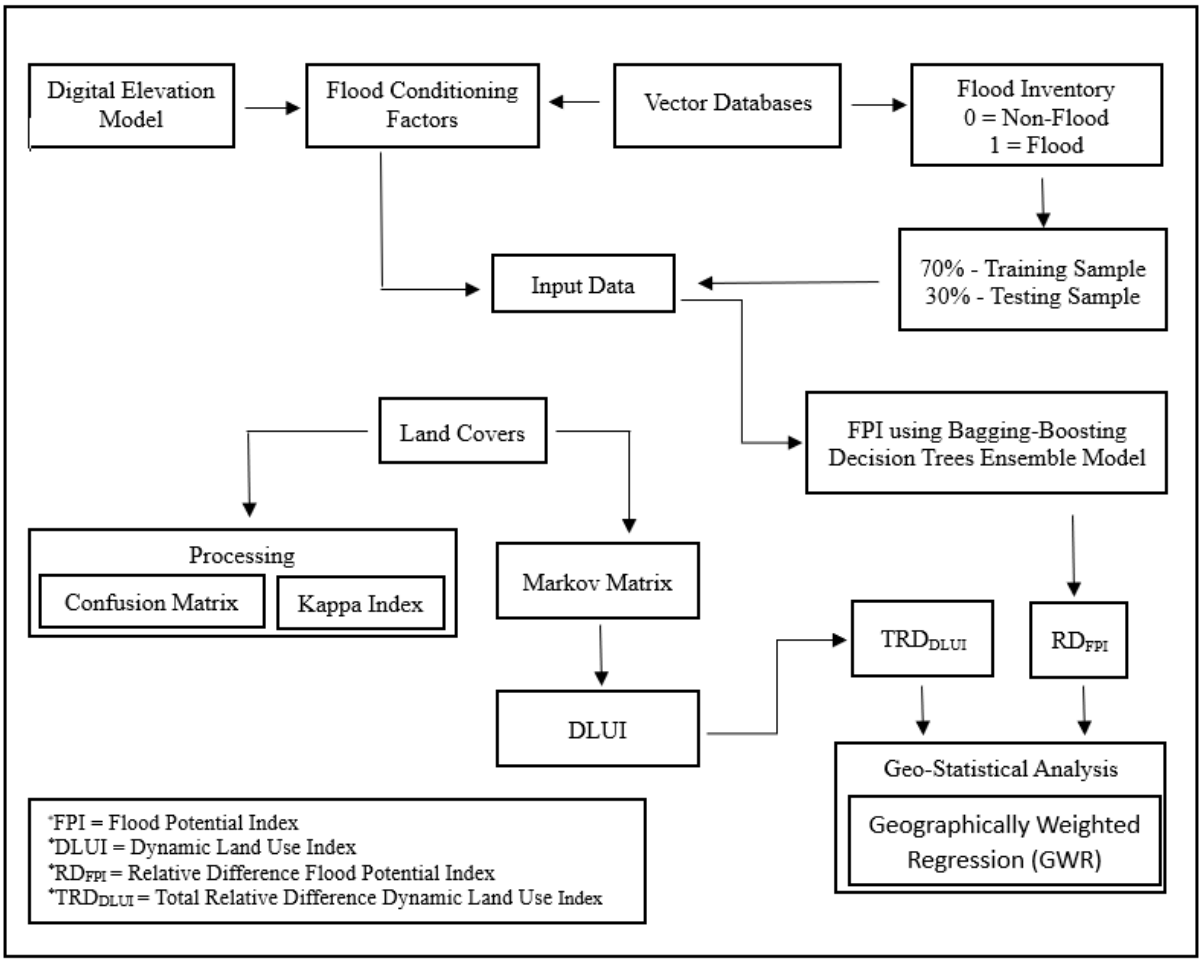


Figure 4. 2: Scheme of Methodological Workflow

4.6.1. Accuracy Assessment of Land Use Land Cover

For the classification data to be useful to detect changes, it is essential to have individual classifications (Owojori et al., 2005). In order to assess the precision of satellite images, the accuracy of the images is evaluated through the utilization of a confusion matrix. Confusion matrices using ground truth testing areas are generated to test the accuracy of classification.

This matrix allows for the calculation of many metrics including overall accuracy, kappa coefficient, producer accuracy, and user accuracy.

$$Overall\ Accuracy = \frac{Total\ Number\ of\ Correctly\ Classified\ Pixels\ (diagonl)}{Total\ Number\ of\ Reference\ Pixels} \times 100 \quad (4.26)$$

$$Producer\ Accuracy = \frac{Number\ of\ Correctly\ Classified\ Pixels\ in\ Each\ Category}{Total\ Number\ if\ Reference\ Pixels\ in\ that\ Category\ (The\ Column\ Total)} \times 100 \quad (4.27)$$

$$User\ Accuracy = \frac{Number\ of\ Correctly\ Classified\ Pixels\ in\ Each\ Category}{Total\ Number\ of\ Classified\ Pixels\ in\ that\ Category\ (The\ Row\ Total} \times 100 \quad (4.28)$$

$$Kappa\ Coefficient = \frac{((TS \times TCS) - \Sigma(Column\ Total \times Row\ Total))}{(TS^2 - \Sigma(Column\ total \times Row\ Total))} \times 100 \quad (4.29)$$

Where; TS implies total sample and TCS is the total corrected sample.

The resultant land use/land covers are used for the calculation of LULC change index (TRD_{SDLUI}) and Flood Potential Index (FPI) for the study time periods (2010-2022).

4.6.2. The Total Relative Difference Dynamic Land Use Indicator

Land-use transitions are employed in the Markov matrix, which are generated through mentioned below steps. Firstly, the raster data files for LULC change are reclassified by utilizing appropriate codes, secondly, the transitions are obtained in ArcGIS software through addition of raster datasets by using the Raster Calculator tool. The changes are then quantized and spatialized through TRD_{DLUI} (Equation (4.30)), which are obtained from an annual ratio of LULC transition, namely the Dynamic Land Use Index. The TRD_{DLUI} index is derived as;

$$TRD_{DLUI} = \frac{\Sigma_{i=1}^{n} |\Delta LU_{i-j}|}{2 \sum LU_i} \times 100 \quad (4.30)$$

Where: TRD_{DLUI} is the TRD obtained from dynamic land use indicator, LU_i shows the area of LULC "i" at the initial date; ΔLU_{i-j} depicts the changed area from land cover "i" to the land cover "j". This indicator calculates the intensity of LULC alterations in all considered spatial locations by showing all changes recorded among various LULCs.

4.6.3. Relative Difference Flood Potential Index

To correlate the variation of flood susceptibilities of the flood years and the LULC transitions for 2010 and 2022, the mean values of flood susceptibility for 2010 and 2022 are

obtained at the grid-cell size of 1 km^2 and flood potential Index (FPI) is thus generated. This procedure is conducted by employing Zonal Statistics tool of Arc GIS software. Afterwards, the FPI values of 2010 and 2022 are integrated into a relative evolution, by using the equation 4.31:

$$RD_{FPI\ (2010-2022)} = \frac{|FPI_2 - FPI_1|}{FPI_1} \times 100 \quad (4.31)$$

Where; RD_{FPI} is the relative difference for the flood potential index, FPI_1 is the flash flood potential index for the starting date (2010), FPI_2 is the flash flood potential index for the end date (2022).

4.6.4. Computation of Geographical Weighted Regression

Due to the heterogeneity of the studied area and spatially varying relationships, the statistical correlation between TRD_{DLUI} and RD_{FPI} indicators is calculated by using Geographical Weighted Regression (GWR). The regression analysis is conducted in Arc GIS by utilizing the two computed indices. This regression is selected instead of Ordinary Least Squares (OLS) due to the spatial heterogeneity. (Fotheringham et al., 2003).

4.7. Quantification of Flood Caused Damages

While developing the methodology, a computational model is created in a GIS context to analyze the expected damage caused by floods. Grid-formatted raster layers are utilized for this objective. A territorial analysis is conducted using ArcGIS software and its Spatial Analyst extension. This involved using several raster datasets to extract information through overlay and combination of themes. The flood susceptibility of 2022 and 2032, computed by using the bag-boost ensemble model has been used in this regard. The database utilized the subsequent information layers:

1) A digital elevation model (DEM) is a representation of the topography of an area using digital data.
2) The flood inundation; determined using hydraulic modeling.
3) The LULC map for various land covers such as built-up (residential, industrial, infrastructural, educational and medical units) and agricultural area.
4) The administrative boundaries map

The shapefile format was transformed into a raster format, with each layer having the same cell size. This was done to ensure a one-to-one correspondence between the cells. This facilitates direct correspondence between cells from different grids and allows for operations to be performed on grid values. In the conversion step from shapefile to another format, a cell dimension of 30 meter was selected, considering the resolution of the data source. This cell size is appropriate for accurately depicting any item or landform with an adequate level of detail.

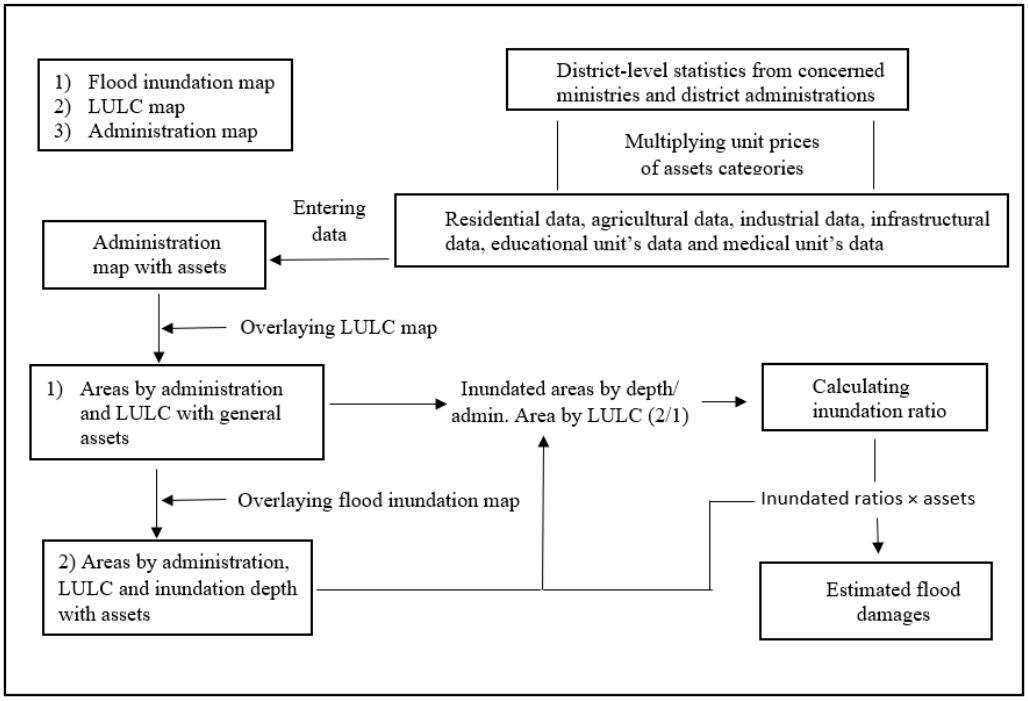


Figure 4. 3: Scheme of Methodological Workflow

For all flooded locations, the ground surface elevation grid (DEM) was subtracted from the computed water surface elevation grid (flood susceptibility), resulting in a grid that displays the floodwater levels above ground surface (m) of the flood depths. Calculation of the predicted economic loss was later carried out by the activities illustrated in Figure 4.3.

The damages thus calculated are for agricultural crop damages, residential buildings, industrial damages, infrastructural damages, educational buildings and medical buildings damages. District-level areas are extracted from land use land cover map of the study area. Agricultural damages are calculated for major kharif crops, cotton, rice and sugarcane owing to the fact that these crops are sown/harvested from June-August, which are the peak flooding months. The cost of damaged crop is calculated based on per unit prices of the crop for normal year. The normal year considered for this research is 2021. The prices of the damaged crops are based on the 2021 prices.

The agricultural damages model utilizes the identical grid as the hybrid bag-boost ensemble model generated flood to calculate the overall agricultural damage, denoted as AD(m), for the three selected months (m). To assess the overall agricultural damages, the agricultural damage per unit area Dua (i, j, m) is initially computed for each grid cell (i, j) during months (m) using Equation 4.32.

$$AD(m) = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} D_{ua}(i,j,m) \,.\, TA(i,j) \tag{4.32}$$

$$D_{ua}(i,j,m) = C_{uw}(i,j) \,.\, Y_{ua}(i,j) \,.\, DC(i,j,m) \tag{4.33}$$

The agricultural damage, denoted as AD (m), is measured in PKR, while the agricultural damage per unit area, denoted as Dua (i, j, m), is clearly expressed in hectors. The parameters n_i and n_j in Equation (4.32) reflect the total number of rows and columns in the grid, respectively. TA (i, j) denotes the total area of the grid cell (i, j) in hectors. In equation (4.33), C_{uw}(i, j) represents the estimated cost per unit weight of the crop in PKR per bale (for cotton) and m. tons (for rice and sugarcane). Y_{ua} (i, j) represents the normal year yield per unit area of the crop in bales and m. tons per hector. DC (i, j, m) represents the dimensionless damage value that corresponds to the flood depth damaged surface for the cell (i, j) and month m, which is generated by the subtracting the flood susceptibility map from DEM.

The research has also conducted forecasting of agricultural damages for the forecasted flood of 2032, generated through our bag-boost model simulations conducted on R Studio software. For the purpose of forecasting, the research has conducted various CGREM model simulations (Islam et al., 2023) to obtain district-level total cultivated area, total normal year production and per unit price of the crop. The model utilized has been depicted in equation 4.25 and the data utilized is for the period of 1990 to 2021.

The flood damaged monetary value of residential buildings, industries and infrastructure has been calculated by utilizing the equations 3.34 (Luino et al., 2006 and Mahmood et al., 2019).

$$E_v = E_{vpe} \,.\, T_e \,.\, DC \tag{3.34}$$

Where, E_v is the economic value of element at risk, E_{vpe} denotes the economic value of the element in PKR, the economic value of residential buildings and infrastructure has been assessed by the construction cost of damaged buildings and infrastructure (Luino et al., 2006)

and for industrial damages, value of fixed assets and GDP at factor prices are utilized to assess the economic value, T_e is the total number of elements in the selected sector and DC is the flood depth damaged surface.

Chapter 5
Feature Selection Using Hybrid Metaheuristic Algorithms and Machine Learning Models

5.1. Introduction

Floods are recognized as highly devastating natural disasters, leading to the loss of millions of lives and causing billions of dollars in economic damages globally. They are caused either due to climatic alterations or are human induced but their destructive power varies geographically due to which spatial analysis of floods is pertinent. Flood occurrences are rarely caused by a single factor; instead, various factors hold varying degrees of importance. While different researchers have emphasized specific factors, some common elements have been highlighted in flood analysis, as pointed out by studies such as Tehrany et al. (2015), Rahmati et al. (2015), Blistanova et al. (2016), and Ali et al. (2019). In hydrology and geospatial analysis, a consensus has not been made on the choice of variables. Thus, feature selection is pertinent to obtain only relevant data subsets that can enhance the performance of classification models.

Feature selection involves choosing relevant features for classification and comprises two steps: (1) exploring subsets of features, and (2) evaluating these subsets to identify the best features. Literature has illustrated three approaches for feature selection which are; filter, embedded, and wrapper methods. Metaheuristic feature selection algorithms stem from the wrapper approach (Agarwal et al., 2022). Various meta-heuristic algorithms are designed and utilized in research to tackle feature selection challenges. These include genetic algorithm (GA) (Zhou and Hua, 2022), simulated annealing (SA), (Araújo et al., 2022), ant colony optimization (ACO) (Hashemi et al., 2022), particle swarm optimization (PSO) (Shanmugam and Preethi, 2019), and more.

Flood modeling and analysis commonly employ hydraulic, statistical, and machine learning models. Hydraulic models, while precise, are influenced by data uncertainties (Rizeei, 2018). Statistical models assume predefined relationships between floods and causal factors, relying on linear interactions in watersheds (Javidan et al., 2020). In contrast, machine learning models primarily consider linear and non-linear relationships and offer advantages like lower computational cost and higher accuracy (Mosavi et al., 2020). Numerous machine learning algorithms have been depicted in previous research for flood modeling, but challenges include overfitting and complex mathematical functions (Wang et al., 2021). Metaheuristic algorithms,

like adaptive neuro-fuzzy inference system (ANFIS) combined with culture (CA), bee (BA), and invasive weed optimization (IWO) algorithms, have been applied for flood susceptibility mapping in different regions (Bui et al., 2019; Termeh et al., 2018; Dodangeh et al., 2020; Wang et al., 2021; Arora et al., 2021; Rahmati et al., 2020; Panahi et al., 2021).

Adhering to Li and Jun (2024), optimization methods can be categorized into traditional and meta-heuristic approaches. Traditional methods like linear programming, non-linear programming, and dynamic programming often converge quickly to local optima but struggle with complex search spaces and large dimensions due to strict constraints, hindering the attainment of global optimal solutions. To address these limitations, intelligent optimization algorithms have gained popularity for their adaptability, ease of use, effectiveness in handling discrete problems, lack of reliance on differentiation, and ability to find global optima. To enhance the performance of machine learning models, numerous meta-heuristic algorithms have been employed to optimize model parameters, yielding more accurate results compared to conventional methods. Meta-heuristic algorithms draw inspiration from natural concepts, and in this study, we focus on metaphor-based metaheuristics, which encompass two paradigms: swarm intelligence-based and evolutionary-based. In hydrological analysis, swarm and evolutionary-based algorithms have been extensively explored in the literature. Thus, Genetic Algorithm (GA) which is an evolutionary-based algorithm has been used. Furthermore, Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) have been employed which belong to the swarm intelligence (SI) category. The choice of PSO and ACO among numerous swarm intelligence-based algorithms has been made due to the reason that they are the most popular and largely used SI algorithms (Abdel-Baseet et al., 2018).

Incorporating machine learning algorithms into flood modeling has transformed the discipline, allowing for the examination of vast datasets and the derivation of valuable insights (Pham et al., 2021). These algorithms are capable of recognizing hidden relationships and patterns in the data that traditional methods cannot identify (Razavi-Termeh et al., 2023). Flood-associated variables involve complex patterns that can be easily recognized by machine and deep-learning algorithms. Moreover, the literature reveals that hybrid machine-learning algorithms have performed better than stand-alone machine-learning models and traditional statistical models. For instance, Wang et al. (2021) forecasted the monthly runoff of the Dahuofang reservoir and compared simple SVM with the hybrid PSO-SVM model and concluded that the hybrid model had better performance capability with more accurate

predictions compared to the simple SVM model. Moreover, Costasche et al. (2020) computed the Flash Flood Potential Index (FFPI) using stand-alone KNN and its hybrid with the AHP model (KNN-AHP). The performance metrics revealed that the hybrid model of KNN produced higher accuracy than the stand-alone KNN model. In this connection, Agarwal et al. (2021), presented a comprehensive literature on feature selection utilizing metaheuristic algorithms for the period of 2009-2019. They investigated that for feature selection using meta-heuristic algorithms, KNN and SVM are the most utilized classifiers and have performed better than other classifiers like Naive Bayesian (NB), Optimum Path Forest (OPF), Random Forest (RF), ID3, Artificial Neural Network (ANN), C4.5, Kernel Extreme Learning Machine (KLM) and Fuzzy rule-based (FR). Referring to the applicability and performance accuracy of hybrids of SVM and KNN with meta-heuristics in hydraulic analysis we have employed these models in our study.

Hence, in the present chapter, metaheuristic algorithms are used to attain the optimal data subset that can be used for flood susceptibility and hazard mapping. To the best of our knowledge, metaheuristic algorithms have not been used for feature selection in hydrology. For this purpose, the research has utilized three algorithms, GA, PSO, and ACO. To evaluate their performance, they are hybridized with two machine learning models, SVM and KNN, for feature selection among the selected dataset. The goal of formulating hybrid models GA-SVM, PSO-SVM, ACO-SVM, GA-KNN, PSO-SVM, and ACO-KNN is to enhance performance accuracy and reduce the number of features. Thus, the research aims to propose a methodology through which the most relevant flood-associated variables can be obtained for the study area. The technical steps for the attainment of this objective are (1) selection of features by using different population sizes of meta-heuristic algorithms (2) assessment of performance accuracies by hybridizing each meta-heuristic with SVM using the kernel trick (3) evaluation of performance accuracies of meta-heuristics by hybridizing them with KNN with altering k-values. To the best of our knowledge, metaheuristic algorithms have been used in hydrology using geo-environmental, human-induced, and topographic variables for mere classification and forecasting purposes. The novelty of the research is that these algorithms are being used for feature selection for geo-environmental, human-induced, and topographic datasets by hybridizing these algorithms (GA, PSO, and ACO) with machine learning models (SVM and KNN).

The rest of the chapter is arranged as section 2 comprises tests and analysis of tests for multicollinearity, heteroscedasticity and autocorrelation, section 3 feature selection models, section 4 illustrates the model performances and section 5 carries a conclusion.

5.2. Tests for Multicollinearity, Heteroscedasticity, and Spatial Autocorrelation

In Figure 5.1, the presence of blue and red colors indicates positive and negative correlations between the variables. The darkness of color and size of the circles reflect a high pairwise correlation. On the right side of the correlogram, the legend color portrays the pairwise correlation. Numerous dark-colored circles in blue and red are evident, indicating strong pairwise correlations. Figure 3 illustrates the existence of high multicollinearity among the covariates.

To assess the prevalence of heteroscedasticity and autocorrelation tests, a simple regression analysis was conducted. The results of the Breusch-Pagan test revealed the existence of heteroscedasticity with a p-value equivalent to 0.098. Moreover, Moran's I test was also conducted to assess of autocorrelation problem in the dataset and it revealed the presence of autocorrelation with a p-value equal to 0.298.

5.2.1. Analysis of Results

The correlation matrix, Breusch-Pagan test result, and Moran's I test result show the presence of multicollinearity, heteroscedasticity, and spatial autocorrelation problems in the dataset. Therefore, to improve the model performance and efficiency, we have utilized heuristic methods; to quickly find optimal or near-optimal solutions. Meta-heuristic algorithms provide a general framework to strategically design heuristics to achieve improved solutions and computational efficacy (Lin and Gen, 2009). For this reason, meta-heuristics are used, in this study, to find the optimal solution or the optimal dataset that is free from these econometric issues. They are further hybridized with machine learning models to assess their performance accuracies.

5.3. Feature Selection

This section introduces a hybrid approach to feature selection. Three feature selection techniques – PSO, ACO, and GA – are employed to extract a subset of highly representative features. This subset can enhance the classification performance in the subsequent modeling

stage. Parameters of all feature selection techniques are fine-tuned, and the selected feature subset is then applied to SVM and KNN classifications, separately. To assess the classification accuracies, a hold-out evaluation method is adopted. The dataset is divided into two parts: 70% is utilized for training, and the remaining 30% is used for testing. Along with the best-selected features, the output includes the most suitable feature selection technique.

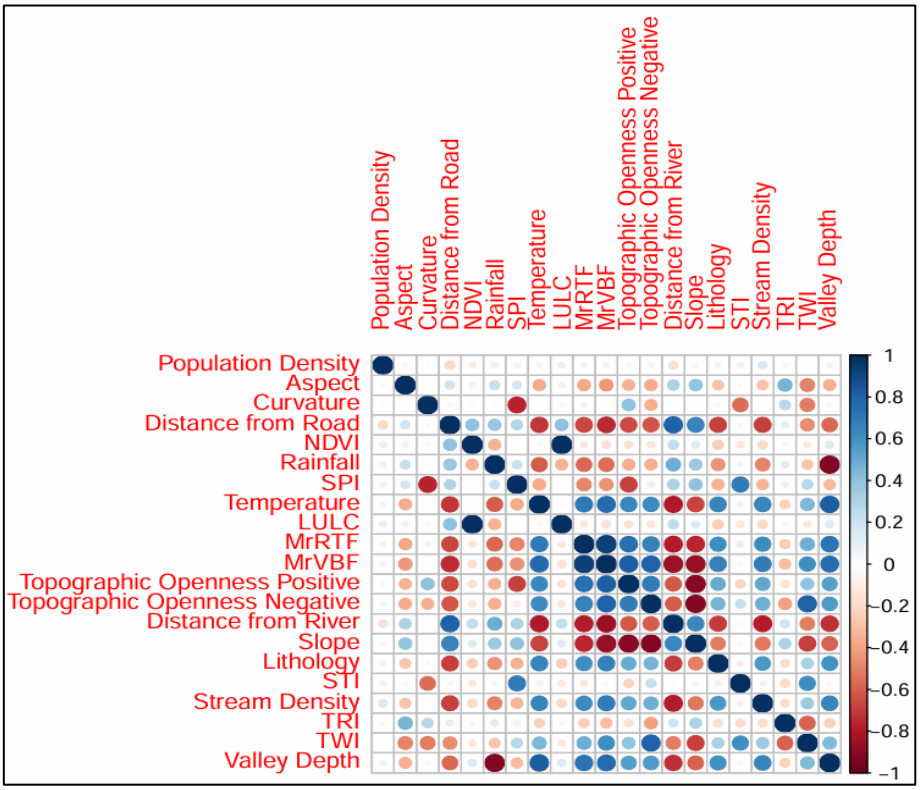


Figure 5. 1: Correlation Matrix

Adhering to Ullah et al. (2017), multicollinearity affects only standard errors and not the prediction capacity of a model. As, the sole purpose of the thesis is prediction rather than estimation; thus, the issue of multicollinearity cannot distort the model. Therefore, multicollinearity is not a significant concern. In their paper, Ullah et al. (2017) discuss various methods that can be used to address the problem of multicollinearity which include; using Principle Component Analysis (PCA), Partial Least Square (PLS) and ridge regression, increasing the sample size, dropping the variables, regularization penalties and decision trees. In the machine learning decision trees, the tree splitting resolves the issue of multicollinearity as these models are robust to such issues (Breiman, 2017). Secondly, feature selection retains the most potential variables and drop the rest depending upon the level of multicollinearity (Dodangeh et al., 2020). The cross-validation through parameter tuning has been conducted in the thesis to optimize the predictive accuracy of the model. Hence, as the thesis is mainly concerned with the flood prediction analysis, so the factors; slope, topographic openness

positive and topographic openness negative have been retained despite the existence of multicollinearity in these factors.

5.3.1. Feature Selection by Varying Population Sizes

For feature selection, GA, PSO, and ACO were fine-tuned, and 1000 iterations were performed for each model to obtain a representative data subset. Table 5.1 depicts ten models constructed with varying population sizes for each metaheuristic algorithm and the number of features obtained are recorded. In the table, it can be depicted that ACO has selected a maximum number of attributes, dropped only rainfall, and selected all the best features at a population size of 10, 25, 70, and 150. Secondly, PSO selected 14 features at a population size of 20. It selected population density, rainfall, SPI, temperature, LULC, topographic openness positive, topographic openness negative, distance from the river, slope, lithology, stream density, STI, TWI, and TRI. Thirdly, GA selected 11 features at most with a population size of 30. It selected population density, curvature, distance from the road, NDVI, rainfall, temperature, topographic openness positive, topographic openness negative, lithology, STI, and TRI.

Table 5. 1: Feature Selection by Varying Population Sizes

Model	**Population Size**	**Features selected**		
		GA	**PSO**	**ACO**
1	10	8	7	**20**
2	15	7	9	19
3	20	7	**14**	19
4	25	6	12	20
5	30	**11**	13	19
6	50	10	10	20
7	70	7	9	20
8	100	9	11	19
9	150	7	8	20
10	200	9	12	19

Figure 5.2 illustrates the performance of the three algorithms through the fitness function. The figure portrays that PSO performed the best compared to ACO and GA. PSO obtained the best solution at about 400 iterations, GA after about 700 iterations and ACO at 600 iterations. It shows that PSO performed better among the selected metaheuristic algorithms. The fitness function metric reveals the best performance of the PSO algorithm as it has obtained optimal solution at the lowest number of iterations but to verify the performance accuracies of these algorithms they are further hybridized with machine learning algorithms;

SVM and KNN. The SVM and KNN models' performances are discussed in the following section.

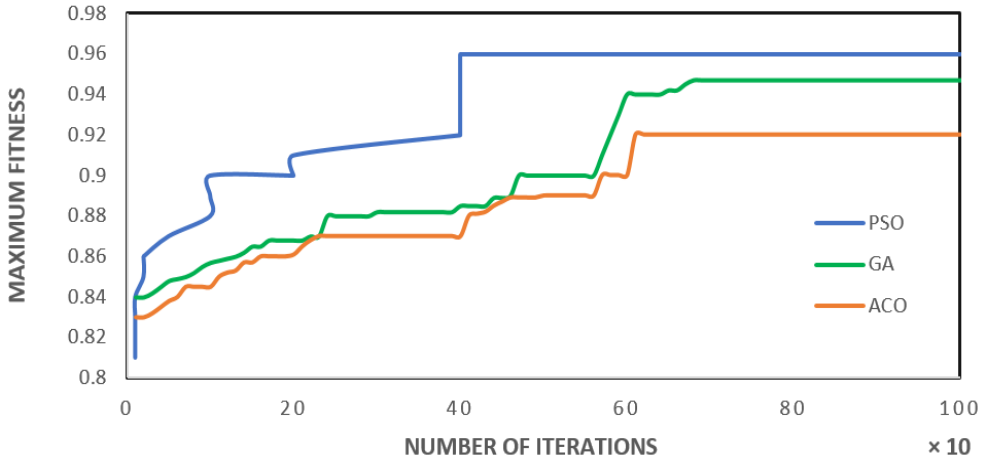


Figure 5. 2: Fitness Function of Metaheuristic Algorithms

5.4. Model Performance

To further evaluate the performance of these algorithms they are hybridized with two machine learning models i.e. SVM and KNN and simulations have been performed to obtain maximum accuracy.

5.4.1. Support Vector Machine

In Tables 5.2, 5.3, and 5.4, SVM accuracies of GA, PSO, and ACO models, respectively, have been assessed using four kernel distributions namely, linear, RBF, sigmoid, and polynomial. For both RBF and sigmoid kernels, σ and d are taken from 0.1 – 0.9 and for polynomial d = 1, 3, 5, 7, 9, and 11 are utilized for assessment of SVM accuracies.

5.4.1.1. Performance of Different SVM Kernels with GA Algorithm

As far as the GA-SVM model is concerned, the parameters σ = 0.4, d = 0.1, and d = 3 are utilized as they revealed the highest accuracies among the rest parameters. Among them, comparatively, RBF has revealed the highest accuracy in model 5 i.e. 96.7% for the training dataset and 99.8% for the validation dataset and sigmoid with the lowest accuracies. The performance output is depicted in Table 5.2.

5.4.1.2. Performance of Different SVM Kernels with PSO Algorithm

Likewise, for PSO, the parameters σ = 0.4, d = 0.7, and d = 3 are employed for RBF, sigmoid, and polynomial distributions. The results of SVM-PSO model are illustrated in Table 5.3. For PSO as well, RBF has shown the highest accuracy compared to linear, sigmoid, and polynomial distributions at model 3 with swarm size of 20 i.e. 99.3 % and 98.7% for training and validation datasets, respectively.

Table 5. 2: Performance of Different SVM Kernels with GA Algorithm (In %)

Model	**Population Size**	**SVM Accuracy**							
		Linear		**RBF Sig=0.4**		**Sigmoid d=0.1**		**Polynomial d=3**	
		T	**V**	**T**	**V**	**T**	**V**	**T**	**V**
1	10	0.902	0.847	0.924	0.871	0.795	0.758	0.867	0.764
2	15	0.935	0.997	0.946	0.997	0.837	0.817	0.937	0.991
3	20	0.938	0.997	0.948	0.998	0.856	0.819	0.923	0.995
4	25	0.947	0.932	0.935	0.947	0.817	0.857	0.941	0.947
5	30	**0.958**	**0.948**	**0.967**	**0.998**	0.847	0.894	**0.949**	**0.948**
6	50	0.937	0.932	0.942	0.957	0.813	0.813	0.937	0.927
7	70	0.927	0.957	0.904	0.986	0.884	0.867	0.923	0.924
8	100	0.903	0.963	0.957	0.927	0.881	0.865	0.927	0.917
9	150	0.914	0.947	0.926	0.914	0.817	0.875	0.947	0.943
10	200	0.916	0.941	0.957	0.943	**0.896**	**0.827**	0.936	0.948

T= Training dataset, V= Validation dataset

5.4.1.3. Performance of Different SVM Kernels with ACO Algorithm

Interestingly, ACO revealed the same accuracies for ant sizes equal to 10, 25, 50, 70, and 150 for all the distributions with σ = 0.4, d= 0.8, and d = 1, showing the highest accuracies for RBF, sigmoid, and polynomial distributions, respectively. Comparatively, RBF showed highest accuracy of 97.9 % and 97.8% for training and validation datasets, respectively. The results for ACO-SVM are depicted in Table 5.4.

Table 5. 3: Performance of Different SVM Kernels with PSO Algorithm (In %)

Model	**Population Size**	**SVM Accuracy**							
		Linear		**RBF Sig=0.4**		**Sigmoid d=0.7**		**Polynomial d=3**	
		T	**V**	**T**	**V**	**T**	**V**	**T**	**V**
1	10	0.912	0.901	0.929	0.975	0.817	0.896	0.947	0.945
2	15	0.925	0.978	0.967	0.957	0.883	0.853	0.928	0.978
3	20	0.958	0.962	**0.993**	**0.987**	0.847	0.874	0.939	0.966
4	25	0.942	0.937	0.991	0.964	0.836	0.869	0.937	0.936
5	30	0.968	0.956	0.992	0.963	0.812	0.817	**0.967**	**0.952**
6	50	0.947	0.947	0.913	0.957	**0.883**	**0.875**	0.917	0.965
7	70	**0.983**	**0.974**	0.923	0.936	0.847	0.842	0.903	0.967
8	100	0.942	0.953	0.956	0.958	0.829	0.825	0.942	0.946

9	150	0.946	0.947	0.927	0.927	0.863	0.873	0.921	0.914
10	200	0.954	0.938	0.993	0.968	0.867	0.937	0.943	0.942

T= Training dataset, V= Validation dataset

5.4.1.4. Performance Analysis of SVM with Metaheuristic Algorithms

It can be assessed that for the selected metaheuristic algorithms, GA, PSO, and ACO, RBF shows the highest accuracy when hybrid models are formulated with SVM. This conclusion is in line with the results provided by Afifi et al. (2013) and Kancherla et al. (2019). Therefore, to evaluate SVM accuracy, RBF distribution can be utilized for the selected algorithms. While comparing RBF, for all three algorithms, it can be seen that PSO has illustrated the highest accuracy. Previous research by Dibike et al. (2001) in rainfall-runoff modeling found that the RBF kernel outperformed other kernels. Similarly, in reservoir monthly inflow forecasts, Yang et al. (2017) demonstrated that SVR with the RBF kernel provided more accurate predictions compared to other kernels. Numerous prior studies in hydrological forecasting, such as those conducted by Yu X. et al. (2004) and Kancherla et al. (2019), have consistently shown that the RBF kernel ensures satisfactory and robust performance.

Table 5. 4: Performance of Different SVM Kernels with ACO Algorithm (In %)

Model	**Population Size**	**SVM Accuracy**							
		Linear		**RBF Sig=0.4**		**Sigmoid d=0.8**		**Polynomial d=1**	
		T	**V**	**T**	**V**	**T**	**V**	**T**	**V**
1	10	**0.927**	**0.915**	**0.979**	**0.978**	0.844	0.896	**0.967**	**0.971**
2	15	0.923	0.917	0.971	0.975	**0.847**	**0.892**	0.962	0.975
3	20	0.923	0.917	0.971	0.975	0.847	0.892	0.962	0.975
4	25	0.927	0.915	0.979	0.978	0.844	0.896	0.962	0.971
5	30	0.923	0.917	0.971	0.975	0.847	0.892	0.962	0.975
6	50	0.927	0.915	0.979	0.978	0.844	0.896	0.962	0.971
7	70	0.927	0.915	0.979	0.978	0.844	0.896	0.962	0.971
8	100	0.923	0.917	0.971	0.975	0.847	0.892	0.962	0.975
9	150	0.927	0.915	0.979	0.978	0.844	0.896	0.962	0.971
10	200	0.923	0.917	0.971	0.975	0.847	0.892	0.962	0.975

T= Training dataset, V= Validation dataset

5.4.2. K-Nearest Neighbor

In the present study, KNN has also been used for the assessment of accuracies of the metaheuristic algorithms. For this purpose, k = 10, 50, 100, 150, 200, 250, 300, 350, and 400 have been used for all algorithms in each model. The k-values are chosen between 10 and 400, due to the reason that k-values less than 10 had very insignificant output results, with very little

variation in the performance metrics, and the *k*-value of 400 acquired the maximum ties. The maximum ties arise due to the binary nature of the dependent variable (flood in our case), when there are too many neighbors equidistant to the target point, such that the algorithm cannot choose only *k* of them. Results revealed that *k* = 50 and 100 obtained the highest accuracies. At *k* = 10, the accuracy was low, it increased to its highest at *k* = 50 and 100 and reduced afterward.

5.4.2.1. Performance Analysis of KNN with Metaheuristic Algorithms

Like in the case of SVM, PSO-KNN has revealed the highest accuracy of 99.8%, compared to the GA-KNN and ACO-KNN. In comparison to SVM, KNN has performed better. This portrays that, for feature selection in the geospatial dataset, PSO-KNN performs better and can be used to attain the most representative set of features. Ay et al. (2023) proposed several hybrid metaheuristic algorithms with SVM and KNN for feature selection and concluded that KNN performs better with metaheuristic algorithms in comparison to SVM. Gauhar et al. (2021) utilized KNN for flood prediction and evaluated performance by varying *k*-values. The results presented by them also depict that accuracy of KNN increase from low to higher value, attain a maximum value, and then start declining. The reason for this has been provided by Varda and Subrahmanyam (2021) that in the KNN model when *k* is small, the algorithm becomes sensitive to outliers. Conversely, if *k* is large, the neighborhood might encompass an excessive number of points from different classes. In our research, *k* larger than 400 encountered this problem of maximum ties but this *k*-value maximum ties issue may vary for other datasets or different study areas. Therefore, selecting an appropriate *k* value is vital and must be chosen vigilantly. The performance of KNN is depicted in table 5.5.

Table 5. 5: Performance of KNN with Different K-Values (In %)

	K Values											
	GA				PSO				ACO			
Models	K=50		K=100		K=50		K=100		K=50		K=100	
	T	V	T	V	T	V	T	V	T	V	T	V
1	0.962	0.969	0.912	0.911	0.967	0.947	0.950	0.957	**0.967**	**0.941**	0.958	0.988
2	0.923	0.912	0.923	0.910	0.981	0.972	0.981	0.983	0.961	0.945	0.947	0.938
3	0.917	0.914	0.978	0.937	**0.998**	**0.990**	0.983	0.925	0.961	0.945	0.947	0.938
4	0.975	0.932	0.956	0.982	0.982	0.952	0.971	0.943	0.967	0.941	0.958	0.988
5	**0.987**	**0.963**	0.978	0.910	0.970	0.953	0.973	0.958	0.961	0.945	0.947	0.938
6	0.927	0.919	0.973	0.971	0.926	0.978	0.986	0.978	0.967	0.941	0.958	0.988
7	0.973	0.978	0.962	0.947	0.924	0.926	0.928	0.917	0.967	0.941	0.958	0.988
8	0.963	0.945	0.951	0.958	0.937	0.917	0.939	0.935	0.961	0.945	0.947	0.938
9	0.958	0.946	0.949	0.946	0.958	0.948	0.957	0.957	0.967	0.941	0.958	0.988
10	0.945	0.921	0.967	0.963	0.926	0.917	0.913	0.901	0.961	0.945	0.947	0.938

T= Training dataset, V= Validation dataset

5.5. The Dataset and the Research Framework

The results of the training and validation dataset for SVM and KNN with the best performance outcomes have shown that the PSO algorithm has highlighted the best and the most pertinent dataset that can be utilized for prediction of floods in the lower Indus basin. Thus, the 14 factors, population density, rainfall, SPI, temperature, LULC, topographic openness positive, topographic openness negative, distance from the river, slope, lithology, stream density, STI, TWI, and TRI, are utilized for the prediction and analysis of floods in the thesis in the next chapters. The major purpose of the thesis is to perform flood prediction analysis, so, these variables are comprised of a combination of environmental, topographic and human-induced factors that are important for flood prediction analysis as highlighted by Tehrany et al. (2015), Bui et al. (2018), Chem et al. (2016), Nanditha et al. (2023) Mahmood et al. (2020) Mukharjee and Singh (2020) and Rahman et al (2021) with disagreements on the choice of variables.

5.6. Conclusion

It is an arduous task to select the best representative features for spatial analysis of hydraulic research in the presence of multicollinearity, heteroscedasticity, and autocorrelation. The application of such models is common in time series data while rare studies have utilized these techniques for spatial analysis using hydraulic datasets. Following are the key findings of the chapter; 1) PSO performed the best and selected the most relevant variables with the least number of iterations, compared to the rest models. 2) Among the hybrid models of SVM, the results revealed that for spatial datasets, with autocorrelation, heteroscedasticity, and autocorrelation problems, PSO-SVM with RBF kernel performed the best with the highest accuracy of 99.3%. Thus, it is suggested that while performing feature selection for spatial datasets, PSO-SVM with RBF kernel can be used. 3) Similarly, PSO-KNN revealed the highest accuracy of 99.8% with a *k*-value of 50. It is further assessed that for KNN, the lower *k*-values showed lower accuracies, then attained a maximum value and finally started declining. Therefore, it is recommended that for PSO-KNN, the k-value should not be the lowest or the highest, rather some middle value be considered. 4) While comparing the performance of PSO-SVM and PSO-KNN, the PSO-KNN performed better. This asserts that PSO-KNN can be employed for feature selection from topographic, geo-environmental, and human-induced datasets possessing high multicollinearity, heteroscedasticity, and autocorrelation problems. In this way, a relevant subset of variables can be selected for accurate flood susceptibility

mapping, hazard and risk assessments, and economic damages calculation that can provide more accurate and precise predictions for the lower Indus basin. As this chapter of the thesis discusses about the representable variables for flood modelling and provides a dataset that can be utilized for flood modelling, hence, the next chapter is mainly concerned about the best model identification, among the decision trees models, that can be utilized for flood susceptibility modelling and flood hazard assessments.

Chapter 6
Comparative Assessment of Decision Trees Models and Ensemble Model

6.1. Introduction

Ensemble machine learning algorithms have gained popularity in machine learning due to their strong performance and ability to handle noisy data. These algorithms enhance the precision of predictions by amalgamating the results of multiple feeble decision models (Zou, 2021; Sagi and Rokach, 2018). This strategy mitigates the likelihood of overfitting the data when using only one algorithm. Aggregating the outcomes of different algorithms also reduces the unpredictability in prediction results caused by variations in data. This results in improved ability to generalize and adapt to new data (Dong et al., 2020; Dietterich, 2000). In addition, ensemble algorithms generally comprise numerous decision trees, with each tree selecting features at split nodes to minimize impurity or maximize information gain. This repeated process of feature selection, which is optimized for performance, allows the algorithms to accurately measure the impact of each feature on the model's performance (Guan et al., 2014). Feature importance offers a clear comprehension of the model's decision-making process, improving the model's interpretability and clarifying how it predicts outcomes using various features. When negative samples are randomly and equally selected, the feature importance accurately represents the distribution patterns of positive samples. This enhances the comprehension of flood vulnerability and offers valuable insights for optimizing the training dataset (Bolón-Canedo and Alonso-Betanzos, 2019). Prior research has conducted a comparative analysis of ensemble learning and conventional machine learning algorithms in the context of flood vulnerability assessment, focusing on accuracy as the primary criterion. Nevertheless, it is necessary to provide additional clarification on the benefits of ensemble algorithms in terms of their interpretability and resilience.

There is a wide range of statistical and machine learning techniques that can be used for flood susceptibility modelling. Statistical models used for flood prediction encompass the frequency ratio method (Lee et al., 2012; Youssef et al., 2016) and the weights-of-evidence approach (Tehrany et al., 2014; Youssef et al., 2015b). Several techniques for making decisions based on various factors have been proposed by Papaioannou et al. (2015), Stefanidis and Stathis (2013), and Youssef et al. (2011b). Machine learning techniques, including artificial neural networks (Radmehr and Araghinejad, 2014), logistic regression (Youssef et al., 2015a), support vector machines (Tehrany et al., 2014), and decision trees (Tehrany et al., 2013), have

been studied in recent years for flood modelling and have shown promising outcomes. Decision Trees (DT) is a highly effective tool for flood susceptibility mapping and has demonstrated strong predictive capabilities (Tehrany et al., 2013). Nevertheless, the application of DT models for flash flood evaluation remains restricted. The decision tree (DT) offers a clear and hierarchical structure with rules that are straightforward to understand (Tien Bui et al., 2016a). The DT method has several advantages. Firstly, it is a statistical analysis that does not make any assumptions about statistical distribution. Secondly, it is capable of handling data from different scales. Thirdly, it allows for the identification of homogeneous groups with different susceptibility levels. Lastly, it aids in the creation of rules for predicting complex relationships (Tehrany et al., 2013). The Decision Tree (DT) can also be utilized for real-time flood forecasting in relation to the increase in water level and water flow (Han et al., 2002). Logistic Model Trees (LMT), Reduced Error Pruning Trees (REPT), Naïve Bayes Trees (NBT), and Random Forest (RF) are sophisticated decision tree approaches. Hence, the primary aim of this chapter is to utilize the LMT, REPT, NBT, RF models and their ensemble to evaluate the outcomes to determine the most effective model for flash flood vulnerability evaluation.

Therefore, in this chapter, features selected by each metaheuristic algorithm has been further used for modelling and compared with each other on the bases of performance by utilizing various performance metrics. For this purpose, 14 features obtained from PSO, 11 variables from GA and 20 factors from ACO models were fed, separately, in each machine learning model (RF, LMT, REPT and NBT) and their ensemble (RF-LMT-REPT-NBT). After training of these models based on their hyper parameter tuning and best optimization, output performance comparison was performed by considering the differences in models' sensitivity, specificity, precision and accuracy. This analysis was performed on both training and validation datasets as the training dataset depicts the model's fitting and validation dataset represents model's generalization ability. The purpose behind performing this comparative modeling is to chalk-out the best model that can be employed for flood susceptibility and hazard mapping and flood risk analysis.

The chapter is organized in following sections; section 1 describes the models' performance using the selected features obtained by each hybrid metaheuristic algorithm and the next section illustrates the impacts of altering the training and validation data sample on the model performance.

6.2. Model Performance Using the Selected Features Obtained by Each Hybrid Metaheuristic Algorithm

For the comparison of decision trees models and their ensemble, each model has been assessed through performance metrics such as true positive, true negative, false positive, false negative, sensitivity, specificity, precision and accuracy. Nguyen et al. (2021) stated that the three-way data split is a robust approach for model development, but it is not always mandatory. If the model demonstrates stable performance, then simplifying the data splitting strategy can be both efficient and effective. As the results are almost stable across different performance metrics and data splits so the data is not divided into three splits i.e. training, validation and testing. Hence, the research considers only two data splits. As the results are almost stable so there is no need of under/over fitting checks in the analysis. The following discussion provides detailed analysis in this context.

6.2.1. Model Performance for Ensemble and Benchmark Models using PSO Algorithm Selected Variables

Table 6.1 depicts that for the PSO selected variables, the ensemble performed the best with 0.995 accuracy and 0.992 precision for training dataset and 0.981 and 0.989 respectively for validation dataset. Followed by REPT model which illustrated 0.979 and 0.973 accuracies and 0.991 and 0.985 precision rates for training and validation datasets, respectively. Likewise, NBT, LMT and RF models showed 0.977, 0.958 and 0.968 accuracies, respectively, for training datasets and 0.967, 0.942 and 0.956 accuracy levels for validation data. Furthermore, they showed 0.989, 0.985 and 0.984, and 0.988, 0.959 and 0.978 precision rates for training and validation datasets, respectively. The performance of these models portrays that ensemble performed the best with highest accuracy and precision values, followed by NBT, REPT, RF and LMT. The LMT showed the least accuracy level. This result is analogous to the study conducted by Khosravi et al. (2018), in which they considered four decision trees models (REPT, NBT, LMT and Altering Decision Trees (ADT)) to assess flood susceptibility and concluded that NBT performed better than LMT and REPT.

Table 6. 1: Model Performance for Ensemble and Benchmark Models using PSO Algorithm Selected Variables

	RF		**LMT**		**NBT**		**REPT**		**Ensemble**	
	T	**V**	**T**	**V**	**T**	**V**	**T**	**V**	**T**	**V**
True Positive	6197	2167	6006	2153	6405	2198	6358	2197	6403	2207
True Negative	1254	989	1374	957	1133	1012	1164	993	1221	1029
False Positive	35	49	93	93	59	34	69	27	35	25
False Negative	214	95	227	97	103	56	109	83	41	39

Sensitivity	0.967	0.958	0.964	0.957	0.984	0.975	0.983	0.964	0.994	0.983
Specificity	0.972	0.953	0.937	0.911	0.951	0.967	0.944	0.974	0.973	0.976
Precision	0.984	0.978	0.985	0.959	0.991	0.985	0.989	0.988	0.992	0.989
Accuracy	0.968	0.956	0.958	0.942	0.979	0.973	0.977	0.967	0.995	0.981

T= Training Dataset, V= Validation Dataset

6.2.1.1. Receiver Operating Characteristic Curve Using PSO Selected Variables

The receiver operating characteristics curve (ROC) has been obtained for both training and validation datasets. In figure 6.1, it can be observed in both training and validation datasets that the ensemble model achieved the highest performance, which can be analyzed by its highest area under the curve (AUC). The outputs of sensitivity and specificity of ensemble model depicts that the number of correctly classified flood pixels is 99.4% and 97.3% for training datasets, respectively, and 98.3% and 97.6% for validation datasets, respectively. The figure also illustrates the highest performance of ensemble model as it shows highest AUC i.e., 99.5% and 96.5% for training and validation datasets, respectively. Followed by NBT with 95.8% AUC for training and 96.2% for validation data, REPT with 93.7% and 95.9% AUC for train and test datasets, RF with 92.9% and 95.1% AUC for training and testing data and lastly, LMT model showed least performance in terms of sensitivity, specificity and AUC showing 89.3% for training and 86.2% for validation datasets.

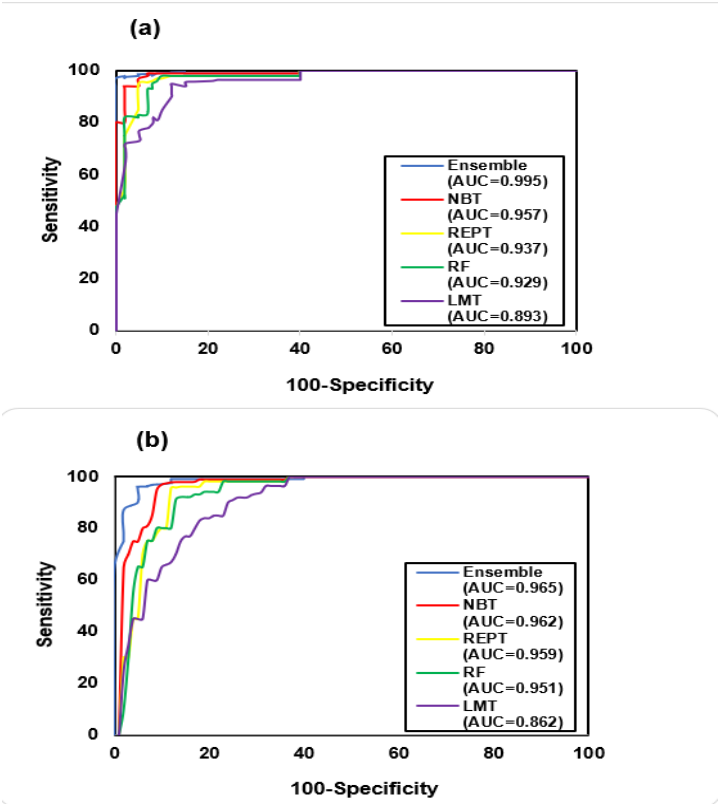


Figure 6. 1: ROC Curves Using PSO Algorithm Selected Variables (a. Training Dataset, b. Validation Dataset)

Table 6.1 and the graphical analysis 6.1 illustrate that with 70/30 data splits for training and testing, the ensemble model has performed the best in terms of sensitivity, specificity, precision, and accuracy. Among the selected individual models, NBT performed better, followed by REPT, RF and the least LMT. This depicts that for flood risk, hazard or susceptibility mapping, an ensemble of REPT, NBT, RF and LMT may be used for higher and better performance instead of individual machine learning models.

6.2.2. Model Performance for Ensemble and Benchmark Models using GA Algorithm Selected variables

Table 6.2 portrays the performance of models with the GA selected variables. With these variables, the ensemble, again, performed the best with 94.5% accuracy and 96.5% precision for training dataset and 96.8% and 97.5% respectively for validation dataset. The second-best results are produced by NBT model which showed 91.1% and 89.8% accuracies and 94.2% and 92.4% precision rates for training and validation datasets, respectively.

Table 6. 2: Model Performance for Ensemble and Benchmark Models using GA Algorithm Selected Variables

	Random Forest (RF)		**LMT**		**NBT**		**REPT**		**Ensemble**	
	T	**V**	**T**	**V**	**T**	**V**	**T**	**V**	**T**	**V**
True Positive	5676	1978	4928	1974	6347	2143	5820	1995	6374	2167
True Negative	1079	987	1473	904	1136	974	1378	1110	1123	1025
False Positive	527	162	647	143	98	87	228	101	98	74
False Negative	418	173	652	279	119	96	274	94	105	34
Sensitivity	0.831	0.819	0.893	0.876	0.947	0.919	0.936	0.930	0.969	0.977
Specificity	0.671	0.809	0.694	0.863	0.759	0.859	0.801	0.857	0.8332	0.949
Precision	0.815	0.904	0.883	0.932	0.942	0.924	0.942	0.925	0.965	0.975
Accuracy	0.837	0.858	0.871	0.872	0.911	0.898	0.906	0.905	0.945	0.968

T= Training Dataset, V= Validation Dataset

Similarly, REPT, LMT and RF models showed 90.6%, 87.1% and 83.7% accuracies, respectively, for training datasets and 90.5%, 87.2% and 85.8% accuracy levels for validation data. Furthermore, REPT, LMT and RF showed 94.2% and 92.5%, 88.3% and 93.2%, and 81.5%, 90.4% precision rates for training and validation datasets, respectively. The performance results of these models illustrate that ensemble model performed the best with the best accuracy and precision rates, followed by NBT, REPT, LMT and RF. The RF showed the least accuracy level. Similarly, Sajithra and Ramyachitra (2021) performed a comparative analysis of tree classifiers for disease datasets and concluded that LMT performed better than RF. Moreover, Khosravi et al. (2019) performed a flood susceptibility analysis in China by

employing various data mining models and revealed that LMT performed better as compared to RF.

6.2.2.1. Receiver Operating Characteristic Curve Using GA Selected Variables

The receiver operating characteristics curve (ROC) of the models with GA variables shows that in both training and validation datasets the ensemble model has the highest performance, which is depicted by its highest area under the curve (AUC). The sensitivity and specificity of ensemble model portrays that the number of correctly classified flood pixels is 96.9% and 83.3% for training datasets, respectively, and 97.7% and 94.9% for validation datasets, respectively. Moreover, the figure illustrates that ensemble has achieved the highest performance in terms of AUC i.e., 93% and 85.9% for training and validation datasets, respectively. Followed by NBT with 91.3% AUC for training and 81.9% for validation data, REPT with 88.3% and 79.0% AUC for train and test datasets, LMT with 82.9% and 76.8% AUC for training and testing data and lastly, RF model obtained least performance in terms of sensitivity, specificity and AUC showing 89.3% for training and 86.2% for validation datasets.

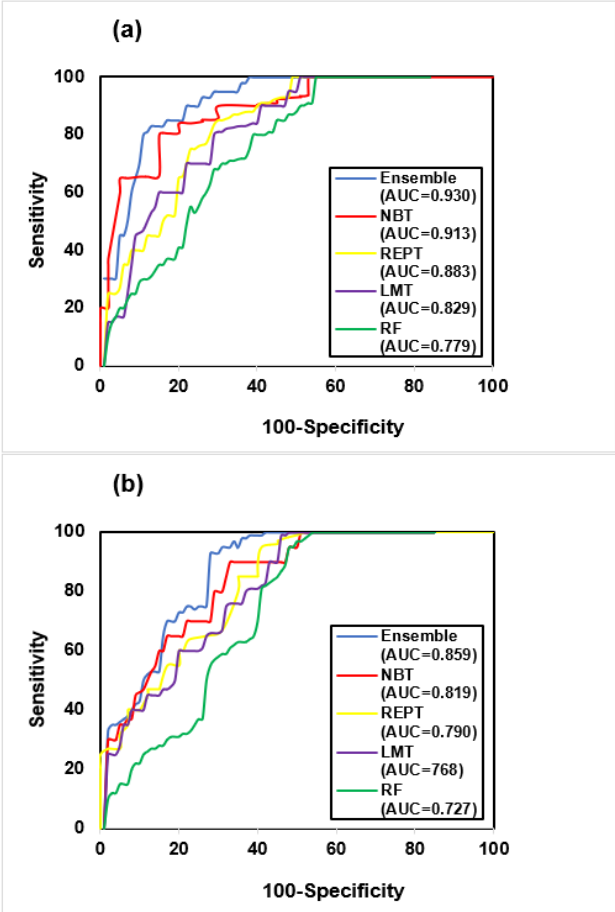


Figure 6. 2: ROC Curves Using GA Algorithm Selected Variables (a. Training Dataset, b. Dataset)

6.2.3. Model Performance for Ensemble and Benchmark Models using ACO Algorithm Selected variables

Table 6.3 illustrates the performance of models with the ACO selected variables. With these variables, the ensemble model, again, showed the best performance with 97.3% accuracy and 98.4% precision rate for training dataset and 96.8% and 96.7% respectively for validation dataset. Furthermore, the NBT model showed 97.1% and 94.4% accuracies and 98.4% and 96% precision rates for training and validation datasets, respectively. Likewise, REPT, RF and LMT models showed 93.4%, 91.6% and 84.8% accuracies, respectively, for training datasets and 94%, 89.1% and 88.9% accuracy levels for validation data. Furthermore, REPT, LMT and RF showed 96.2% and 95.1%, 94.2% and 95.1%, and 93%, 90.1% precision rates for training and validation datasets, respectively. The performance analysis of these models demonstrate that ensemble model has performed best with highest accuracy level and precision rates, followed by NBT, REPT, RF and LMT. LMT model showed the least performance levels. The same result is demonstrated by the study presented by Rai and Sharma (2021) which undertakes a classification of Av labeling technique by utilizing various decision trees models. They concluded that RF model performs better than LMT model.

Table 6. 3: Model Performance for Ensemble and Benchmark Models using ACO Algorithm Selected Variables

	Random Forest (RF)		**LMT**		**REPT**		**NBT**		**Ensemble**	
	T	**V**	**T**	**V**	**T**	**V**	**T**	**V**	**T**	**V**
True Positive	5781	1874	5173	1932	6347	2143	5820	1995	6374	2167
True Negative	1278	1067	1357	1004	1136	974	1378	1110	1123	1025
False Positive	352	140	551	210	98	87	228	101	98	74
False Negative	289	219	619	154	119	96	274	94	105	34
Sensitivity	0.952	0.895	0.893	0.926	0.981	0.957	0.955	0.955	0.983	0.984
Specificity	0.784	0.884	0.711	0.827	0.920	0.918	0.858	0.916	0.919	0.932
Precision	0.942	0.930	0.903	0.901	0.984	0.960	0.962	0.951	0.984	0.967
Accuracy	0.916	0.891	0.848	0.889	0.971	0.944	0.934	0.940	0.973	0.968

T= Training Dataset, V= Validation Dataset

6.2.3.1. Receiver Operating Characteristic Curve Using ACO Selected Variables

The receiver operating characteristics curve (ROC) of the machine models with the features obtained by ACO algorithm shows that in both training and validation datasets the ensemble model has achieved the best performance, as it possesses highest area under the curve (AUC). The sensitivity and specificity of the ensemble model shows that the number of correctly classified flood pixels is 98.3% and 91.9% for training datasets, respectively, and, 98.4% and 93.2% for validation datasets, respectively.

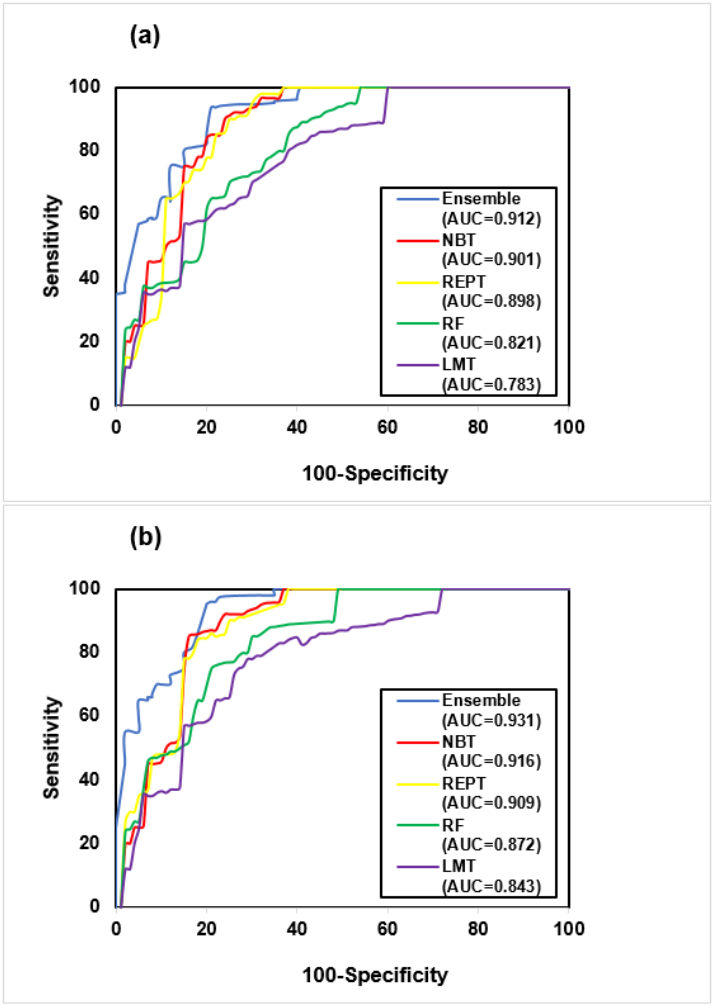


Figure 6. 3: ROC Curves Using ACO Algorithm Selected Variables (a. Training Dataset, b. Validation Dataset)

Similarly, the figure depicts that ensemble has achieved the highest performance in terms of AUC as well i.e., 91.2% and 93.2% for training and validation datasets, respectively. Followed by NBT with 90.1% AUC for training and 91.6% for validation data, REPT with 89.8% and 90.9% AUC for training and testing datasets, RF with 82.1% and 87.2% AUC for training and testing data and lastly, LMT model obtained least performance in terms of sensitivity, specificity and AUC showing 78.3% for training and 84.3% for validation datasets.

6.2.4. Tests of Statistical Significance

The research has utilized those variables which are selected by employing each hybrid metaheuristic algorithm, to assess the performance of the ensemble model and the standalone benchmark models. In this regard, the tables 6.1, 6.2 and 6.3 depict the results obtained from all classification models. In order to ensure that the validation test is not influenced by random chance, the research assessed the significance of these results using the Friedman test. The

study has solely evaluated the significant differences among all classifiers resulting from the PSO feature selection, as this outcome is the most favorable. The null hypothesis states that there are no significant variations in accuracy among the classifiers, while the alternative hypothesis suggests that there are substantial differences in accuracy among the classifiers. At significance level of 0.95, the output of classifier significance using Friedman rank sum test revealed χ_F^2 value equal to 5 and p-value equal to 0.0253, thus, the null hypothesis is rejected and there exist no significant variations in accuracy among the classifiers. As this test only presents statistical distinctions for all the models, it does not offer a comparison between each pair of these models (Beasley and Zumbo, 2003). Consequently, the Neymenyi test is also employed in this investigation.

The outcome of Friedman Rank sum test suggests that there are notable differences among classifiers. Nevertheless, due to the conservative nature of this result, we employ a more robust post hoc test known as the Nemenyi test to compare all classifiers against one another. At the confidence level of 0.05, the Neymenyi test has p-value of 0.8765. This depicts that there exist no significant variations in accuracy among the classifiers.

6.3. The Sensitivity Analysis of Data with Various Data Splits

While conducting flood hazard modelling research, it is difficult to validate the models without dividing the data on flood conditioning factors into separate training and testing datasets (Chung and Fabbri 2003). However, there is no agreement on the optimal amount of dataset to use for creating training and testing datasets. For example, Pradhan (2013) used an equal split of 50/50 for training and validation datasets, while Saro Lee and Oh (2012) used a split of 70/30 for training and testing datasets.

In this research, feature selection was performed to identify the most suitable dataset that can contribute to flood causation in the Lower Indus basin and potentially impact the modelling outcomes. However, it is also crucial to determine the optimal ratio for data splits in order to generate datasets that yield the most accurate flood susceptibility or flood hazard map. Firstly, flood elements were divided into different proportions with a 10% difference between each. Furthermore, these datasets were acquired through the utilization of these specific data divisions. Subsequently, the ensemble model was utilized to perform ROC analysis on both the training and testing datasets. This analysis aimed to assess the data sensitivity in each data split and ultimately create the final flood susceptibility map. The AUC values were produced and

used to validate the performance of the ensemble model. The sensitivity analysis results demonstrate that the area under the curve (AUC) values of the ensemble model vary dramatically for each dataset that was generated.

Table 6.4: The Performance of Ensemble Model with various Training and Validation Data Splits

S. No.	Datasets	Dataset Split	Total Data in Each Splits	AUC
1	Training	10%	9900	0.992
	Validation	90%	1100	0.767
2	Training	20%	2200	0.991
	Validation	80%	8800	0.824
3	Training	30%	3300	0.983
	Validation	70%	7700	0.839
4	Training	40%	4400	0.978
	Validation	60%	6600	0.841
5	Training	50%	5500	0.972
	Validation	50%	5500	0.921
6	Training	60%	6600	0.964
	Validation	40%	4400	0.927
7	Training	70%	7700	0.995
	Validation	30%	3300	0.965
8	Training	80%	8800	0.984
	Validation	20%	2200	0.962
9	Training	90%	9900	0.957
	Validation	10%	1100	0.910

This illustrates that the choice of a suitable data division for the training and testing datasets has a substantial impact on the modelling outcomes. The results demonstrate that when the flood factors' data is split with a ratio of 70/30, the ensemble model achieves the highest AUC values for both the training and validation datasets. In the present research, with 70/30 data splits, the ensemble model obtained 99.5% and 96.5% of AUC for training and testing datasets, respectively. Rest all the data splits obtained lower AUC either for training data sample or validation data sample. For instance, 10/90 data split for training and validation samples, respectively, obtained a higher AUC of 99.2% for training sample but for validation sample it obtained the least AUC level of 76.7%. Thus, the data samples of 70/30, for training and validation samples, respectively, provides the best and accurate results. This result is in line with the study presented by Pham et al. (2018). Table 6.4 and figure 6.4 illustrate the impact of various data splits on ensemble model performance. Therefore, it can be deduced that the splitting ratio of 70/30 yields the optimal performance ratio for creating training and testing datasets for flood susceptibility mapping in this study.

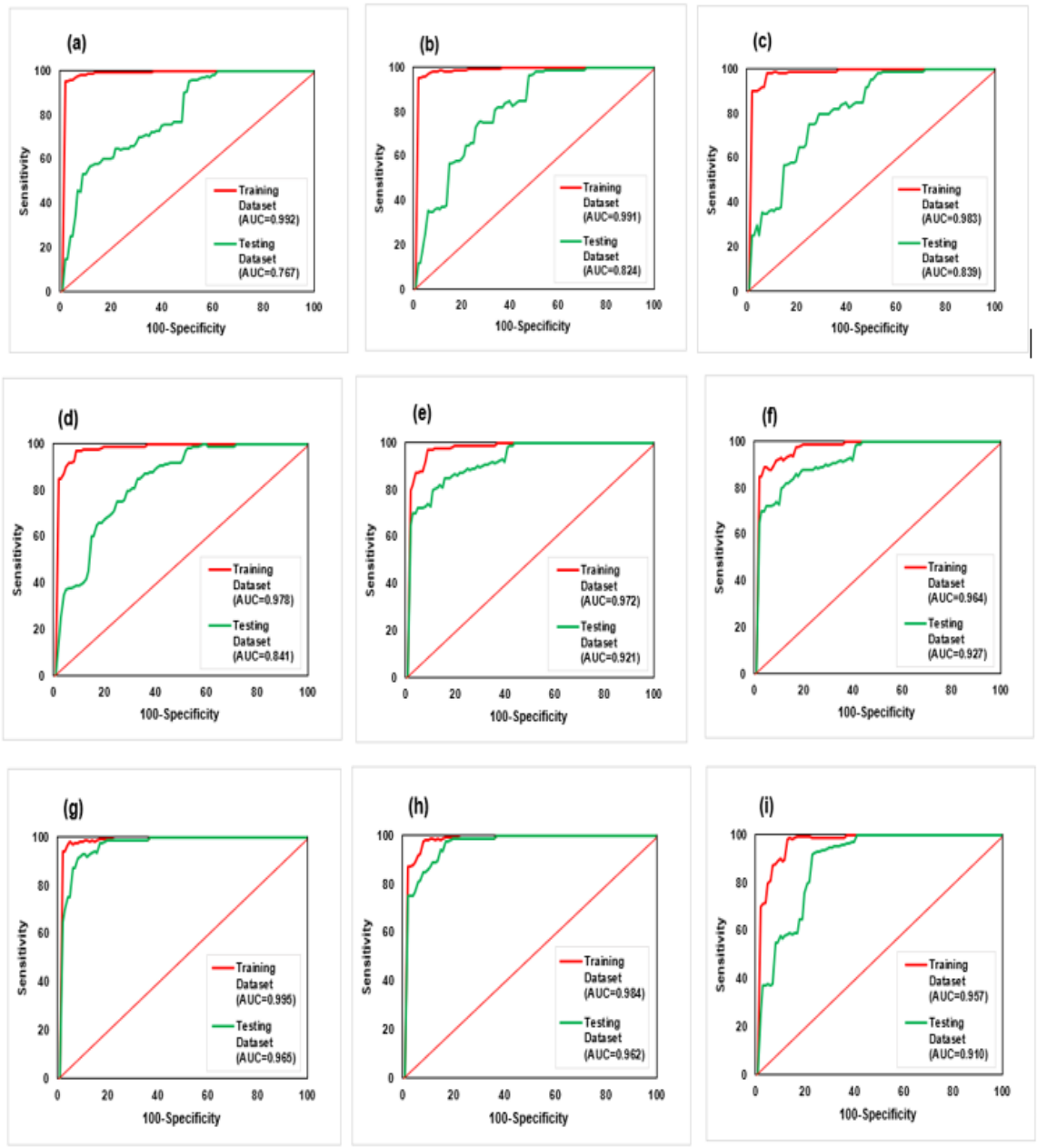


Figure 6. 4: ROC Curves with Training and Validation Data Splits (a. 10%/90%, b. 20%/80%, c. 30%/70%, d. 40%/60%, e. 50%/50%, f. 60%/40%, g. 70%/30%, h. 80%/20%, i. 90%/10%)

6.4. Conclusion

In this chapter, the model performance has been compared and analyzed for the ensemble model and the independent models by utilizing various statistical metrics. The five classification models have been compared among themselves by utilizing three different datasets i.e. obtained by simulation results of feature selection. The key results of the models reveal that 1) the ensemble model has performed the best with all the datasets and LMT showed the least accuracies and areas under the curve for PSO and ACO variables and RF had the least accuracy and area under the curve for GA variables. 2) When the dataset is divided into various

splits, hence; altering the training and validation data range, the data split with 70% training sample and 30% validation sample produces the best results. Thus, it is recommended to utilize 70/30 data split as training and validation sample, respectively.

This chapter has compared decision trees models' performance with their ensemble model and concludes that the ensemble model performs the best among all the selected models. This depicts that this model can be utilized for flood susceptibility mapping. Therefore, the next chapter of the thesis has employed this model for mapping the flood hazard in the lower Indus basin for the year 2022 and 2032.

Chapter 7
The Flood Susceptibility Mapping and Regional Hazard Analysis using Hybrid Bagging Boosting Decision Trees Ensemble Model

7.1. Introduction

Owing to climate change and other environmental conditions, flood disasters have become the most recurrent natural occurrence. The susceptibility of most countries to flood dangers poses a substantial threat to human life globally, resulting in various forms of devastation, including physical, social, and economic consequences (Hussain et al., 2011; Quan et al, 2021; Wang et al., 2021). While floods can cause damage in various locations, the agricultural sector and infrastructure located near rivers are particularly susceptible due to the widespread impact of floods on agricultural land. This could be attributed to inadequate cartography or preventative efforts, or the inclusion of various factors such as drainage density, slope, and so on (Huang et al., 2019; Wahla et al., 2021; Chen et al., 2021). Based on the historical data over the past three decades, Pakistan has annual occurrences of flooding Tariq et al., 2019; Waqas et al., 2021). The country's susceptibility to hazards can be attributed to several factors, including its large latitudinal breadth, geographical location, presence of three mountain systems, diverse temperature fluctuations, and regional earth morphology (Tariq et al., 2022; Zou et al., 2022). In addition to physiographic factors, there are several other factors that contribute to susceptibility. These include the rapid population growth, a large population living below the poverty line, inadequate disaster management measures at the local and national level (including the absence of accurate flood mapping), a low economic growth rate, and a lack of education on pre- and post-disaster management (Yin et al., 2022; Zhan et al., 2022). Disaster impacts exhibit non-uniform distribution and possess diverse characteristics Avand et al., 2021). There is a direct correlation between susceptibility and disasters, meaning that the more vulnerable a population is, the greater the impact of disasters will be. The most vulnerable individuals, including women, young people, and the disabled, are disproportionately impacted by these circumstances due to their limited resources and capabilities (Tehrany et al., 2014; Thomas et al., 2017].

Flood modelling and analysis commonly employ hydraulic, statistical, and machine learning models. Hydraulic models, while precise, are influenced by data uncertainties (Rizeei, 2018). Statistical models assume predefined relationships between floods and causal factors, relying on linear interactions in watersheds (Javidan et al., 2020). In contrast, machine learning

models primarily consider linear and non-linear relationships and offer advantages like lower computational cost and higher accuracy (Mosavi et al., 2020). Numerous machine learning algorithms have been depicted in previous research for flood modeling, but challenges include overfitting and complex mathematical functions (Wang et al., 2021). Metaheuristic algorithms, like adaptive neuro-fuzzy inference system (ANFIS) combined with culture (CA), bee (BA), and invasive weed optimization (IWO) algorithms, have been applied for flood susceptibility mapping in different regions (Bui et al., 2019; Termeh et al., 2018; Dodangeh et al., 2020; Wang et al., 2021; Arora et al., 2021; Rahmati et al., 2020; Panahi et al., 2021)

Pakistan has had several episodes of strong monsoonal rainfall since mid-June 2022, primarily due to a powerful low-pressure system (Mallapaty, 2022). Moreover, the occurrence of heavy precipitation in Pakistan is linked to the presence of La Nina, as stated by Adnan et al. (2021) and Ali et al. (2020). In 2022, the eastern Pacific experienced a cool sea surface temperature. The presence of La Nina worsened the precipitation event, along with the strengthening of low-pressure systems (L Otto et al., 2022). Seven glacial lake outbursts caused by summer heatwaves resulted in an elevated flow rate in the higher tributaries of the Indus river (Jones, 2022; UNDP, 2022). According to Nanditha et al. (2023), the flood in 2022 exceeded the highest rate of water flow during the destructive floods in Pakistan in 2010. In addition, the 2022 event shares similarities with the 2010 event in terms of the presence of La-Nina and Rossby formations in the high-altitude jet streams (Aziz, 2022; Di Capua et al., 2021; Hong et al., 2011). In 2010, a trough formed in the upper troposphere over the Khyber Pakhtunkhwa and northwest Baluchistan due to the presence of the mid-latitudinal jet stream. In August 2022, a comparable system was created in northwest Pakistan (L Otto et al., 2022). An analysis has been conducted on the contribution of human-induced warming to the significant floods that occurred in 2010 and 2022. Hirabayashi et al. (2021) found that the 2010 flood event was made worse by human activities, but Christidis et al. (2013) did not provide any conclusive evidence linking the precipitation event to climate change. It is essential to comprehend the main cause of floods in order to accurately assess the impact of human-induced global warming on the recurrence of catastrophic disasters. This knowledge is necessary for developing timely adaptation strategies for the future.

Therefore, in this chapter, the ensemble model (RF-LMT-NBT-REPT) that has been selected in the previous chapter, due to its best performance, is utilized to conduct the susceptibility mapping of the lower Indus basin for 2022 flood. The mapping is based on the

14 selected flood contributing factors which are utilized to calculate the flood extent through the ensemble model.

The rest of the chapter is organized as section 2 illustrates the relative importance of the flood contributing factors, discusses the spatial relationship between flood factors and flood by utilizing frequency ratio model and provides the flood susceptibility map of the lower Indus basin and demonstrates the impact of flood intensity and extent at district level for 2022 and predicted 2032 flood.

7.2. Flood Susceptibility and Analysis

The study area was hit by a destructive flood in 2022. The following sections provide detailed discussion on 2022 flood. Moreover, ten-years ahead prediction has also been provided in the next section. In this regard, section 7.2.1 provides details of 2022 flood and 7.2.2 discusses the predicted flood of 2032.

7.2.1 Analysis of 2022 Flood

The 2022 flood in lower Indus basin has been analyzed and discussed in detail in the following sections.

7.2.1.1. Relative Importance of Conditioning Factors in Flood Causation

The Gain Ratio approach is employed in this study to assess and choose the appropriate flood influencing elements for flood analysis in the study area. This method improves the predictive capability of models by eliminating factors that have little predictive value. Table 7.1 depicts the relative importance of the flood conditioning factors (for 2022 flood in the study area) among the 14 selected variables by employing all five models. It is evident from the table that rainfall has the highest impact on flood for all models and population density has the least contribution in flood causation for all models except for NBT. The model NBT showed lowest gain ratio for topographic openness negative i.e. 0.862 instead of population density as shown by the rest models.

In a nutshell, the gain ratio output shows that all the selected variables have varying degrees of contribution in flood as none of the GR value is zero. Moreover, the general illustration is that the variables like rainfall, LULC, temperature, slope and stream power index are pertinent contributors with gain ratios higher than 90% and openness negative and

population density are least contributors with gain ratios lower than 90% for all the models. Figure 7.1 shows the bar charts for the predictive ability of the conditioning factors using all five models using the gain ratio method. An analogous variable sensitivity analysis has been presented by Yaseen et al. (2022) for flood susceptibility mapping of Karachi. They concluded that in Karachi, the highest flood contributing factors in 2022 flood were land use land cover, rainfall and elevation.

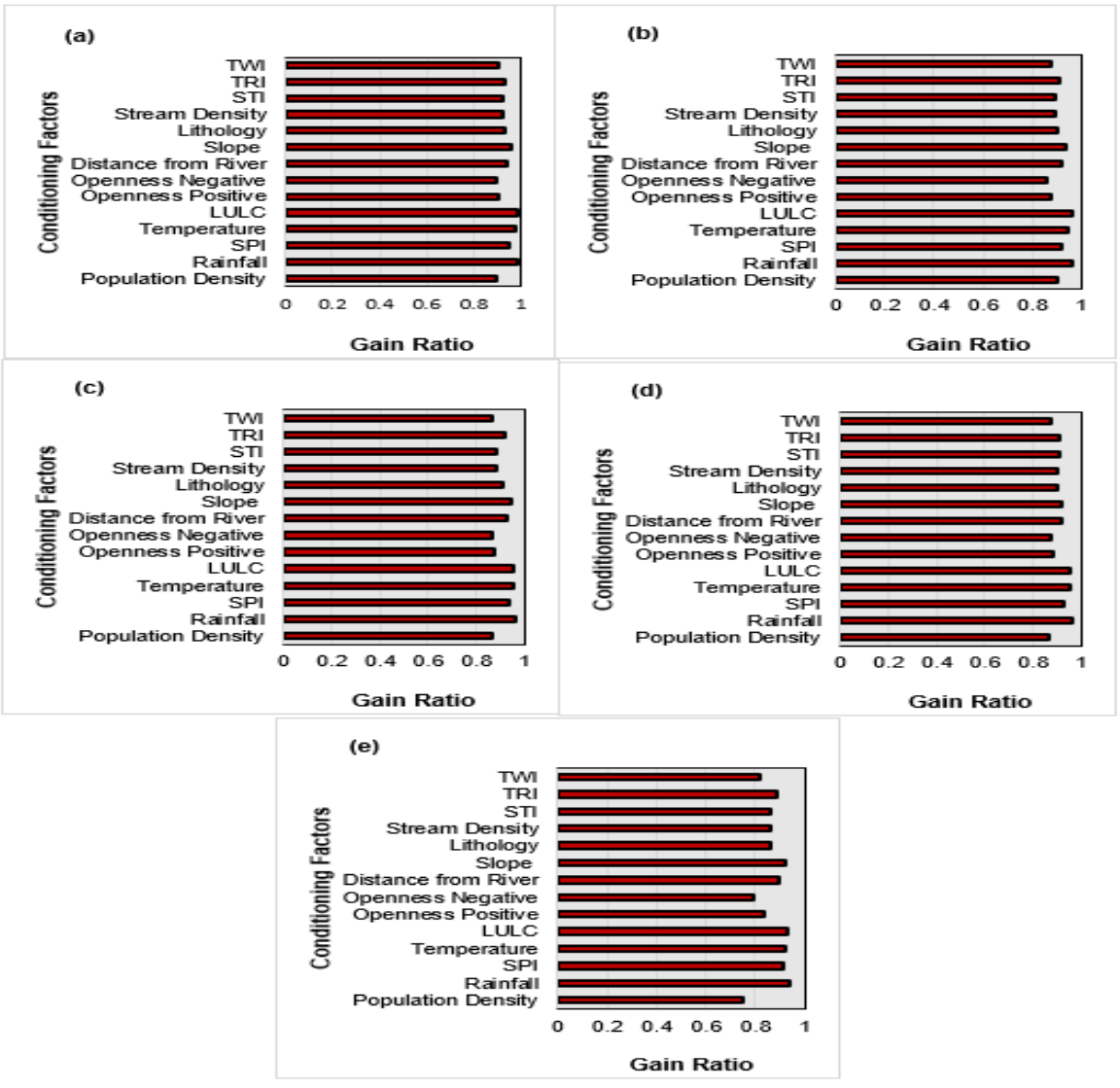
Table 7. 1: Relative Importance of Conditioning Factors in 2022 Flood Causation

S. No.	Flood Conditioning Factors	Gain Ratio (GR)				
		Ensemble	NBT	REPT	RF	LMT
1.	Rainfall	0.985	0.967	0.959	0.967	0.938
2.	Land Use Land Cover	0.983	0.964	0.954	0.958	0.929
3.	Temperature	0.977	0.945	0.951	0.956	0.925
4.	Slope	0.962	0.937	0.947	0.919	0.917
5.	Stream Power Index	0.951	0.921	0.938	0.931	0.912
6.	Distance from River	0.943	0.919	0.932	0.919	0.893
7.	Topographic Ruggedness Index	0.936	0.916	0.918	0.906	0.884
8.	Lithology	0.93	0.903	0.910	0.905	0.862
9.	Stream Density	0.925	0.894	0.887	0.897	0.858
10.	Sediment Transport Index	0.922	0.892	0.885	0.908	0.858
11.	Openness Positive	0.904	0.875	0.872	0.886	0.836
12.	Topographic Wetness Index	0.901	0.875	0.869	0.879	0.817
13.	Openness Negative	0.897	0.862	0.863	0.873	0.795
14.	Population Density	0.894	0.899	0.862	0.870	0.751

7.2.1.2. Spatial Relationship between Conditioning Factors and 2022 Flood

An FR value equal to 0 portrays lack of flood grid cells. In table 7.2 it can be witnessed that none of the FR values are 0, which indicates that the selected variables had significant impact in causation of 2022 flood. In stream power Index, the class boundary between -11.47-10.46 (FR = 1.60) is more susceptible to flood. Likewise, the slope between 0.66 and 7.74 (FR = 1.66) degrees has high flood susceptibility. Moreover, the stream density between 3.38 to 34.43 (FR= 1.02), sediment transport index between 0.1-11.79 (FR=1.16), Topographic Wetness index between the range of 7.67-8.51 (FR=1.35) and topographic ruggedness index of the range 0.67-0.88 (FR=1.42) are more susceptible to flood as they obtained the highest FR values among the class boundaries of the respective variables. For topographic openness

positive and topographic openness negative, the class boundaries 1.55 to1.66 with FR =1.22 and 1.61 to 1.65 with FR=1.28, respectively, showed the highest flood pixels. Further, the areas with population density between 45000 to 60000 is more susceptible to floods as is FR is the highest with a value of 184.56. In land use land cover, the croplands and built up areas (FR=1.84) are highly probable to flood compared to the rest land covers. The areas which are 80000-100000 meters distant from the river are the most probable to floods with FR of 40.53. The districts with rainfall between 3000 to 4000 mm and temperature between 27.7-28.5 have higher flood susceptibility as they show FR= 1.36 and FR=1.55, respectively. Lastly, the lithology of Sindh shows that areas with loamy soil are more flood prone with FR=1.66. This evaluation depicts a comprehensive grasp of the contribution of each class of all the contributing elements to the occurrence of floods in the research area, as shown in table 7.2.



.

Figure 7. 1: The Predictive Capability of Flood Conditioning Factors (a) Ensemble (b) NBT (c) REPT (d) RF (e) LMT

Table 7. 2: Spatial Relationship between Conditioning Factors and 2022 Flood by Frequency Ratio (FR) Model

S. No.	Factor	Class Boundaries	Factor Pixel in Class Domain		Flood Pixel in Class Domain		FR
			No.	%	No.	%	
1	Stream Power Index	(i)-13.81- -11.46	17173616	10.97	8.69	14.81	1.35
		(ii)-11.47-10.46	27360883	17.48	16.38	27.90	1.60
		(iii)-10.47--8.11	22415978	14.32	2.20	3.75	0.26
		(iv)-8.12- -2.59	41312073	26.39	18.89	32.17	1.22
		(v)-2.60- -10.39	48290040	30.85	12.55	21.38	0.69
2	Slope (Degree)	(i) 0- 0.59	44635376	13.37	5.77	9.82	0.73
		(ii) 0.60-0.65	17201182	33.16	4.96	8.44	0.25
		(iii) 0.66 – 1.25	34114962	5.22	5.10	8.68	1.66
		(iv) 1.26-7.74	51442541	33.21	32.45	55.26	1.66
		(v) 7.75-78.14	9158530	15.05	10.46	17.81	1.18
3	Stream Density	(i) 0.001-3.37	21515364	29.12	16.52	28.14	0.97
		(ii) 3.38-6.88	16836947	22.79	13.64	23.23	1.02
		(iii) 6.89-10.66	21248962	28.76	17.02	28.99	1.02
		(iv) 10.67-15.66	12910814	17.47	10.42	17.74	1.02
		(v) 15.67-34.43	1381752	1.87	1.12	1.91	1.02
4	Sediment Transport Index (STI)	(i) 0	108527933	89.28	52.21	88.91	1.00
		(ii) 0.1-11.79	894116	0.74	0.50	0.86	1.16
		(iii) 11.80-23.59	6718929	5.53	3.62	6.16	1.11
		(iv)23.60-47.18	5289096	4.35	2.35	3.99	0.92
		(v)47.19-3007.97	125362	0.10	0.05	0.08	0.81
5	Topographic Wetness Index (TWI)	(i) 1.89-6.82	2813924	1.80	0.35	0.59	0.33
		(ii) 6.83-7.66	57157967	36.51	12.94	22.04	0.60
		(iii) 7.67-8.51	79499943	50.78	40.36	68.74	1.35
		(iv) 8.52-10.28	16865872	10.77	5.02	8.56	0.79
		(v) 10.29-21.53	214885	0.14	0.05	0.08	0.57
6	Topographic Ruggedness Index (TRI)	(i) 0.11-0.33	30194799	20.65	16.30	27.75	1.34
		(ii) 0.34-0.44	35922254	24.56	11.36	19.35	0.79
		(iii) 0.45-0.55	23687923	16.20	5.52	9.40	0.58
		(iv) 0.56-0.66	30310361	20.72	10.62	18.08	0.87
		(v) 0.67-0.88	26137316	17.87	14.92	25.41	1.42
7	Openness positive	(i) 0.71-1.36	714980	0.46	0.09	0.15	0.32
		(ii) 1.37-1.46	2605478	1.66	0.35	0.60	0.36
		(iii) 1.47-1.51	8857763	5.66	0.78	1.34	0.24
		(iv) 1.52-1.54	28525569	18.22	4.55	7.75	0.43
		(v) 1.55-1.66	115848801	74.00	52.95	90.17	1.22
8	Openness negative	(i) 0.69-1.30	335642	0.21	0.04	0.07	0.31
		(ii) 1.31-1.49	6462396	4.13	0.70	1.19	0.29
		(iii) 1.50-1.55	32241284	20.59	5.40	9.20	0.45
		(iv) 1.56-1.60	71037479	45.38	30.31	51.62	1.14
		(v) 1.61-1.65	46475789	29.69	22.27	37.93	1.28
9	Population density	(i) < 15000	117642263	0.76	33.04	56.26	74.32
		(ii) 15000-30000	2732832	0.02	1.82	3.10	176.10
		(iii)30000-450000	23188548	0.15	15.72	26.76	179.34

		(iv) 45000-60000	10785939	0.07	7.52	12.81	184.56
		(v) 60000-80000	1042007	0.01	0.63	1.07	159.37
10	Land Use Land Cover (LULC)	(i) Savannas,	38843684	24.81	6.37	10.85	0.44
		(ii) Grasslands and Permanent Wetlands	22189911	14.18	13.01	22.16	1.56
		(iii) Croplands and Urban buildup lands	43553122	27.82	30.05	51.17	1.84
		(iv) Natural Vegetation and Barren lands	1186205	0.76	0.72	1.22	1.62
		(v) Water bodies and Shrublands	50765371	32.43	8.57	14.59	0.45
11	Distance from River (m)	(i) 0 - 20000	106848894	68.22	4	6.18	0.09
		(ii) > 20000-40000	28023772	17.89	4	6.99	0.39
		(iii)> 40000-60000	14640488	9.35	4	6.54	0.70
		(iv) >60000-80000	4612350	2.94	9	15.66	5.32
		(v)> 80000-100000	2497734	1.59	38	64.63	40.53
12	Rainfall (mm)	(i) 500 -1000	37963661	24.26	18.52	31.53	1.30
		(ii) > 1000 – 2000	20444966	13.06	8.90	15.16	1.06
		(iii) >2000 – 3000	39281895	25.10	13.66	23.25	0.93
		(iv) > 3000 – 4000	37348449	23.86	14.79	25.18	1.36
		(v) > 4000 - 6000	21469847	13.72	2.86	4.87	0.36
13	Temperature (C)	(i) 2.59-26.6	1629994	1.04	0.14	0.24	0.23
		(ii) 26.7-27.0	19968050	12.76	1.55	2.63	0.21
		(iii) 27.1-27.2	17402562	11.12	4.38	7.46	0.67
		(iv) 27.3-27.6	62573185	39.97	20.67	35.20	0.88
		(v) 27.7-28.5	54960498	35.11	31.99	54.47	1.55
14	Lithology	(i) Clay	76676321	47.25	11.45	19.50	0.41
		(ii) Loam	7295255	4.50	4.37	7.44	1.66
		(iii) Silt Loam	72030007	44.39	42.31	72.05	1.62
		(iv) Clay Loam	6264870	3.86	0.60	1.01	0.26

7.2.1.3. Flood Susceptibility Prediction

Figure 7.2 depicts the flood susceptibility map for 2022. In this research, the ensemble model (discussed in previous chapter) has been utilized to construct the 2022 flood susceptibility map by using the 14 selected flood conditioning factors. Moreover, for comparison and validation of the flood extent, a Google Earth Engine (GEE) based flood extent estimation has also been conducted by using flooding satellite imagery of Sentinel 1 for the period of July to August. It is evident from the flood map that northern and western parts of Sindh are affected by flood with extremely high flooding. Table 7.3 shows the district-wise flood extents calculated by using the flood susceptibility map. The table tells that the districts; Dadu, Jamshoro, Kashmore, Khairpur, Larkana, Mitiari, Shaheed Benazirabad, Shikarpur and Sukkur flood inundation of higher than 1000 km^2 area. Shikarpur was the worst affected district with total affected area of 1900 km^2. The total flood affected area in Sindh province was 22100

km^2. This conclusion is analogous with the studies conducted by Sohail and Mohammad (2023) and Qamer et al. (2022). Figure 7.3 depicts the bar chart of flood intensities. It is evident from the bar chart that 57% of the total flood affected area had extremely high flood, 14% of the area suffered through high flood, 12% had moderate inundation, 11% had low and 6% area was extremely low flooded.

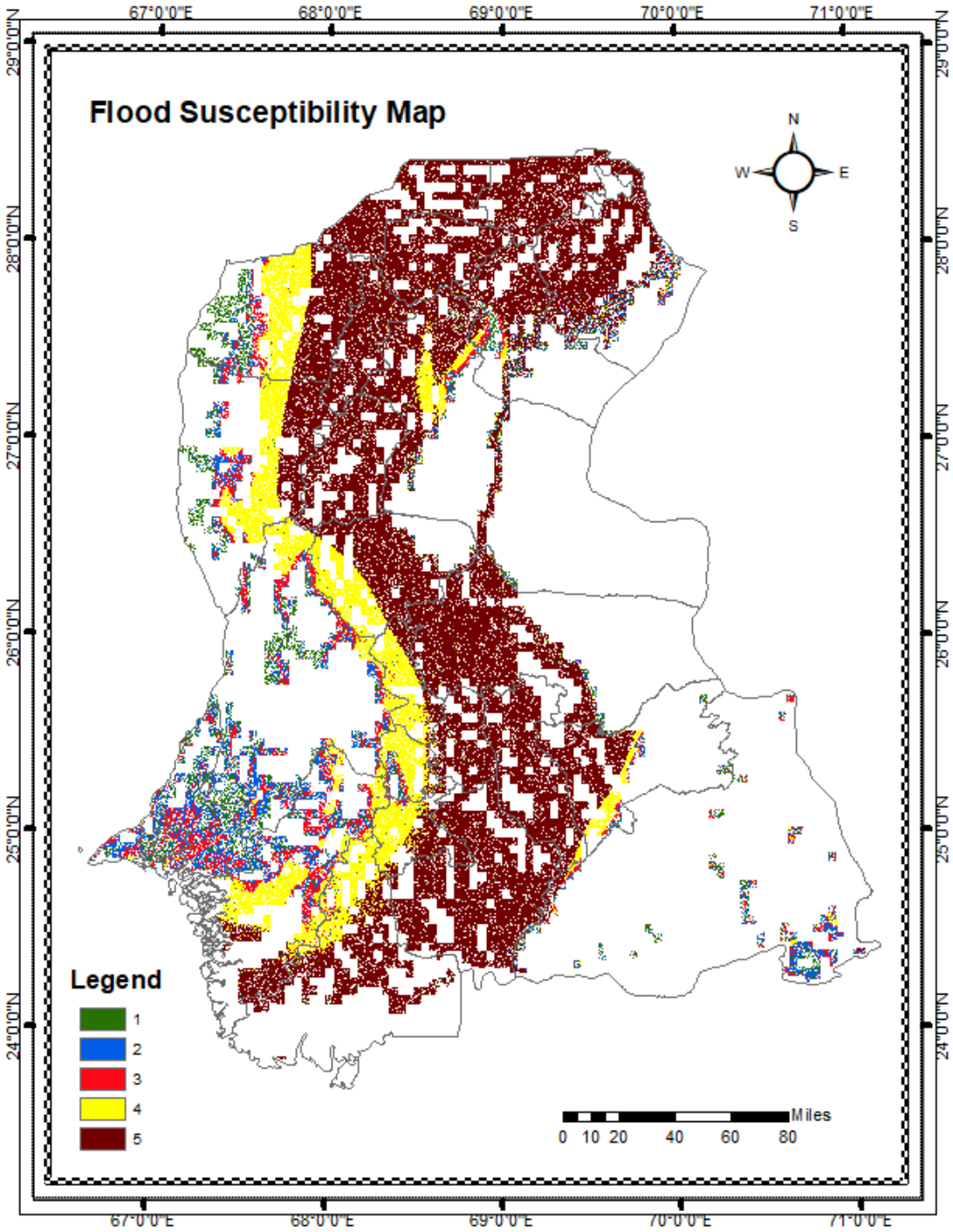


Figure 7. 2: Flood Susceptibility Map of 2022 (1) Extremely Low (2) Low (3) Moderate (4) High (5) Extremely High

Table 7. 3: District-wise Flood Extent (Areas in 100)

Districts	Affected Area of Districts (km^2)					Total Affected Area (km^2)
	Extremely Low	Low	Moderate	High	Extremely High	
Badin	0.0132	1.0059	1.0019	0.0078	2.7009	4.7297
Central Karachi	0.0115	0.005	0.0077	0	0	0.0242
Dadu	2.1858	1.2534	7.1654	2.043	5.1058	17.7534
East Karachi	0.0197	0.0268	0.0438	0.0008	0	0.0911
Ghotki	1.0826	0.1196	1.0408	1.0283	1.6391	4.9104
Hyderabad	0.0263	0.0294	0.0394	0.298	0.034	0.4271
Jacobabad	0.0026	0.0006	0.0004	5.0017	8.0323	13.0376
Jamshoro	1.5003	1.5583	2.4329	1.3945	2.0075	8.8935
Kambar Shahadat Kot	0.2508	1.1033	1.1294	0.754	0.9346	4.1721
Kashmore	0.0074	2.0009	1.0005	5.0006	10.132	18.1414
Khairpur	2.0956	3.0722	6.0683	4.288	2.4411	17.9652
Korangi Karachi	0.0098	0.008	0.0132	0.0065	0	0.0375
Larkana	0.0087	0.0012	1.0007	2.0029	11.9049	14.9184
Malir Karachi	0.2221	0.3798	0.2985	0.002	0	0.9024
Matiari	0.0091	0.0053	1.0163	1.4239	11.1976	13.6522
Mirpur Khas	0.0034	0.0004	0.0027	0.0213	1.3533	1.3811
Naushahro Feroze_	0.0105	0.0012	0.0005	0.0021	1.17	1.1843
Sanghar	0.0512	0.0252	0.0089	0.0142	2.3117	2.4112
Shaheed Benazirabad	0.0287	0.0115	0.0273	0.4631	18.1649	18.6955
Shikarpur	0.0077	0.0019	0.0017	0.009	19.0759	19.0962
South Karachi	0.0289	0.0617	0.0495	0.0063	0	0.1464
Sujawal	1.0219	0.0112	1.0387	2.4698	2.2244	6.766
Sukkur	0.0951	0.0708	0.0545	1.0513	16.9135	18.1852
Tando Allahyar	0.0075	0.0031	0.002	0.0107	0.614	0.6373
Tando Muhammad	0.0075	0.007	0.0206	0.371	0.3507	0.7568
Tharparkar	5.2102	12.1758	3.2035	1.75	0.6707	23.0102
Thatta	0.1413	0.3814	0.4996	1.7564	3.3694	6.1481
Umer Kot	0.026	0.0305	0.036	0.1596	2.689	2.9411
West Karachi	0.0233	0.0327	0.024	0.0015	0	0.0815
Total	14.1087	23.3841	27.2287	31.3383	125.0373	221.0971

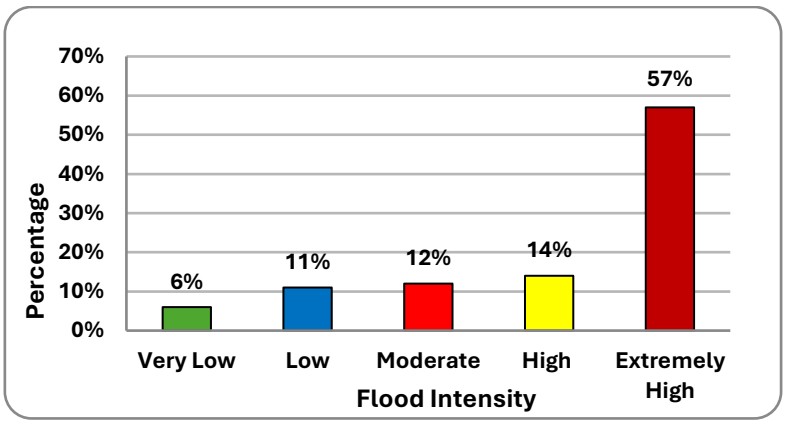


Figure 7. 3: Percentage of 2022 Flood Intensity

7.2.2. Analysis of 2032 Flood

The following sections provide the detailed description and analysis of the predicted flood of 2032.

7.2.2.1. Relative Importance of Conditioning Factors in Flood Causation

Table 7.4 illustrates the relative importance of the 14 flood conditioning factors employed in this study using the ensemble model. The table depicts that, just like the main contributors in 2032 flood, rainfall, LULC and temperature are the major contributors in flood 2032 with GR higher than 96%. Moreover, distance from the river is also a higher contributor with GR 96.7%. Unlike 2022 flood, the lowest contributor in 2032 flood is openness negative rather than population density. Figure 7.4 illustrates the bar chart of 2032 flood contributing factors using ensemble model

Table 7. 4: Relative Importance of Conditioning Factors in 2032 Flood Causation (In%)

S. No.	Flood Conditioning Factors	Gain Ration (GR)
1.	Rainfall	0.989
2.	LULC	0.984
3.	Temperature	0.982
4.	Slope	0.962
5.	SPI	0.946
6.	Distance from River	0.967
7.	TRI	0.927
8.	Lithology	0.938
9.	Stream Density	0.951
10.	STI	0.922
11.	Openness Positive	0.904
12.	TWI	0.901
13.	Openness Negative	0.897
14.	Population Density	0.905

7.2.2.2. Spatial Relationship between Conditioning Factors and 2032 Flood

Table 7.5 shows the flood grid cells of 2032 flood using FR model. The class boundary between -13.81- -11.46 (FR = 2.40) in the stream power index indicates a higher vulnerability to flooding. Similarly, the slope ranging from 0.66 to 7.74 degrees (FR = 1.68) exhibits a significant vulnerability to flooding. Furthermore, the stream density ranging from 3.38 to 6.88 (FR= 1.39), sediment transport index ranging from 0.1 to 11.79 (FR=1.16), Topographic

Wetness index ranging from 7.67 to 8.51 (FR=1.35), and topographic ruggedness index ranging from 0.67 to 0.88 (FR=1.42) exhibit a higher vulnerability to flooding as they have the highest FR values among the class boundaries of their respective variables. The flood pixels in the class boundaries of 1.55 to 1.66 with a topographic openness positive and a FR value of 1.22, and 1.61-1.65 with a topographic openness negative and a FR value 3.01. In addition, areas with a population density ranging from 60000 to 80000 are particularly vulnerable to floods, as indicated by the highest FR value of 128.82.

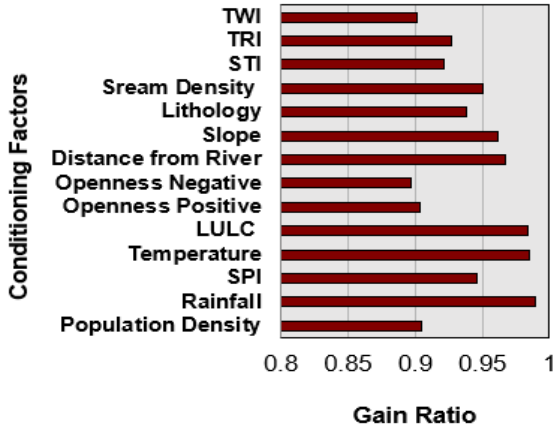


Figure 7. 4: The predictive Capability of 2032 Flood Conditioning Factors Using Ensemble Model

When it comes to land use land cover, the areas used for agriculture and urban areas (FR=1.92) have a significantly higher likelihood of experiencing flooding compared to other types of land cover. The regions located 60000-80000 meters away from the river are the most susceptible to flooding, with a frequency ratio (FR) value of 5.14. The areas that receive rainfall ranging from 1000 to 2000 mm and have temperatures between 27.7 and 28.5 (C^o) exhibit higher sensitivity to flooding, with frequency ratios (FR) of 1.32 and 1.66, respectively. Finally, the geological composition of Sindh indicates that regions with loamy soil have a higher susceptibility to flooding, with a flood risk factor (FR) of 1.66.

This review demonstrates a thorough understanding of the impact of each category of all the contributing factors on the incidence of floods in the research area, as illustrated in table 7.5.

Table 7.5: Spatial Relationship between Conditioning Factors and 2032 Flood by Frequency Ratio (FR) Model

S. No.	Factor	Class Boundaries	Factor Pixel in Class Domain		Flood Pixel in Class Domain		FR
			No.	%	No.	%	
1	Stream Power Index	(i)-13.81- -11.46	17173616	10.97	8.69	14.82	2.40
		(ii)-11.47-10.46	27360883	17.48	16.37	27.91	1.63
		(iii)-10.47--8.11	22415978	14.32	2.19	3.74	0.20
		(iv)-8.12- -2.59	41312073	26.39	18.87	32.18	1.23
		(v)-2.60- -10.39	48290040	30.85	12.52	21.35	0.19
2	Slope (Degree)	(i) 0- 0.59	44635376	13.37	21.92	37.38	1.31
		(ii) 0.60-0.65	17201182	33.16	10.81	18.43	1.68
		(iii) 0.66 – 1.25	34114962	5.22	19.46	33.18	1.52
		(iv) 1.26-7.74	51442541	33.21	5.58	9.51	0.29
		(v) 7.75-78.14	9158530	15.05	0.88	1.50	0.26
3	Stream Density	(i) 0.001-3.37	21515364	29.12	16.50	28.13	0.97
		(ii) 3.38-6.88	16836947	22.79	13.62	23.22	1.39
		(iii) 6.89-10.66	21248962	28.76	17.01	29.00	1.04
		(iv) 10.67-15.66	12910814	17.47	10.40	17.74	1.53
		(v) 15.67-34.43	1381752	1.87	1.12	1.91	1.02
4	Sediment Transport Index (STI)	(i) 0	108527933	89.28	52.14	88.91	1.00
		(ii) 0.1-11.79	894116	0.74	0.50	0.86	1.16
		(iii) 11.80-23.59	6718929	5.53	3.61	6.16	1.11
		(iv)23.60-47.18	5289096	4.35	2.34	3.99	0.92
		(v)47.19-3007.97	125362	0.10	0.05	0.08	0.81
5	Topographic Wetness Index (TWI)	(i) 1.89-6.82	2813924	1.80	0.34	0.59	0.33
		(ii) 6.83-7.66	57157967	36.51	12.91	22.01	0.60
		(iii) 7.67-8.51	79499943	50.78	40.33	68.76	1.35
		(iv) 8.52-10.28	16865872	10.77	5.02	8.56	0.79
		(v) 10.29-21.53	214885	0.14	0.05	0.08	0.57
6	Topographic Ruggedness Index (TRI)	(i) 0.11-0.33	30194799	20.65	16.28	27.77	1.34
		(ii) 0.34-0.44	35922254	24.56	11.35	19.35	0.79
		(iii) 0.45-0.55	23687923	16.20	5.51	9.39	0.58
		(iv) 0.56-0.66	30310361	20.72	10.60	18.08	0.87
		(v) 0.67-0.88	26137316	17.87	14.91	25.42	1.42
7	Openness positive	(i) 0.71-1.36	714980	0.46	0.08	0.14	0.31
		(ii) 1.37-1.46	2605478	1.66	0.35	0.59	0.36
		(iii) 1.47-1.51	8857763	5.66	0.77	1.32	0.23
		(iv) 1.52-1.54	28525569	18.22	4.53	7.73	0.42
		(v) 1.55-1.66	115848801	74.00	52.90	90.21	1.22

8	Openness negative	(i) 0.69-1.30 (ii) 1.31-1.49 (iii) 1.50-1.55 (iv) 1.56-1.60 (v) 1.61-1.65	335642 6462396 32241284 71037479 46475789	0.21 4.13 20.59 45.38 29.69	0.08 0.32 0.64 5.11 52.48	0.14 0.55 1.09 8.72 89.50	0.67 0.13 0.05 0.19 3.01
9	Population density	(i) < 15000 (ii) 15000-30000 (iii)30000-450000 (iv) 45000-60000 (v) 60000-80000	117642263 2732832 23188548 10785939 1042007	0.76 0.02 0.15 0.07 0.01	0.04 0.13 0.19 1.09 57.19	0.07 0.22 0.32 1.86 97.53	2.16 3.20 10.87 105.59 128.82
10	Land Use Land Cover (LULC)	(i) Savannas, (ii) Grasslands and Permanent Wetlands (iii) Croplands and Urban buildup lands (iv) Natural Vegetation and Barren lands (v) Water bodies and Shrublands	38843684 22189911 43553122 1186205 50765371	24.81 14.18 27.82 0.76 32.43	6.34 13.01 30.04 0.72 8.54	10.81 22.18 51.22 1.23 14.57	0.39 1.56 1.92 1.37 0.57
11	Distance from River (m)	(i) 0 - 20000 (ii) > 20000-40000 (iii)> 40000-60000 (iv) >60000-80000 (v)> 80000-100000	106848894 28023772 14640488 4612350 2497734	68.22 17.89 9.35 2.94 1.59	2.97 42.49 2.07 8.87 2.20	5.07 72.50 3.54 15.14 3.76	0.07 4.05 0.38 5.14 2.36
12	Rainfall (mm)	(i) 500 -1000 (ii) > 1000 – 2000 (iii) >2000 – 3000 (iv) > 3000 – 4000 (v) > 4000 - 6000	37963661 20444966 39281895 37348449 21469847	24.26 13.06 25.10 23.86 13.72	18.30 10.14 15.50 13.86 0.85	31.21 17.28 26.44 23.63 1.44	1.29 1.32 1.05 0.99 0.11
13	Temperature (C)	(i) 2.59-26.6 (ii) 26.7-27.0 (iii) 27.1-27.2 (iv) 27.3-27.6 (v) 27.7-28.5	1629994 19968050 17402562 62573185 54960498	1.04 12.76 11.12 39.97 35.11	1.02 12.27 10.43 12.15 22.78	1.73 20.93 17.79 20.72 38.84	1.11 1.64 1.60 0.52 1.66
14	Lithology	(i) Clay (ii) Loam (iii) Silt Loam (iv) Clay Loam	76676321 7295255 72030007 6264870	47.25 4.50 44.39 3.86	11.39 4.37 42.30 0.60	19.41 7.45 72.12 1.02	0.41 1.66 1.62 0.26

7.2.2.3. Flood Susceptibility Prediction

The flood susceptibility map for 2032 is illustrated in Figure 7.5. The present study employs the ensemble model, previously explained in the preceding chapter, to create the flood susceptibility map for the year 2032. This is achieved by utilizing 14 carefully chosen flood conditioning elements. The flood map clearly indicates that the northern and western regions of Sindh have been severely impacted by floodwaters, experiencing both extremely high and high levels of flooding. The flood extents for each district, as determined by the flood susceptibility map, are presented in Table 7.6. According to the table, the districts of Dadu, Ghotki, Jamshoro, Kashmore, Khairpur, Larkana, Mitiari, Shaheed Benazirabad, Shikarpur, Tharparkar and Sukkur will experience flood inundation areas of greater than 1000 km2. Kashmore may experience the most severe impact, with a total damaged area of 2300 km2. The extent of the flood-affected territory in the Sindh province is anticipated to be 22500 square kilometers.

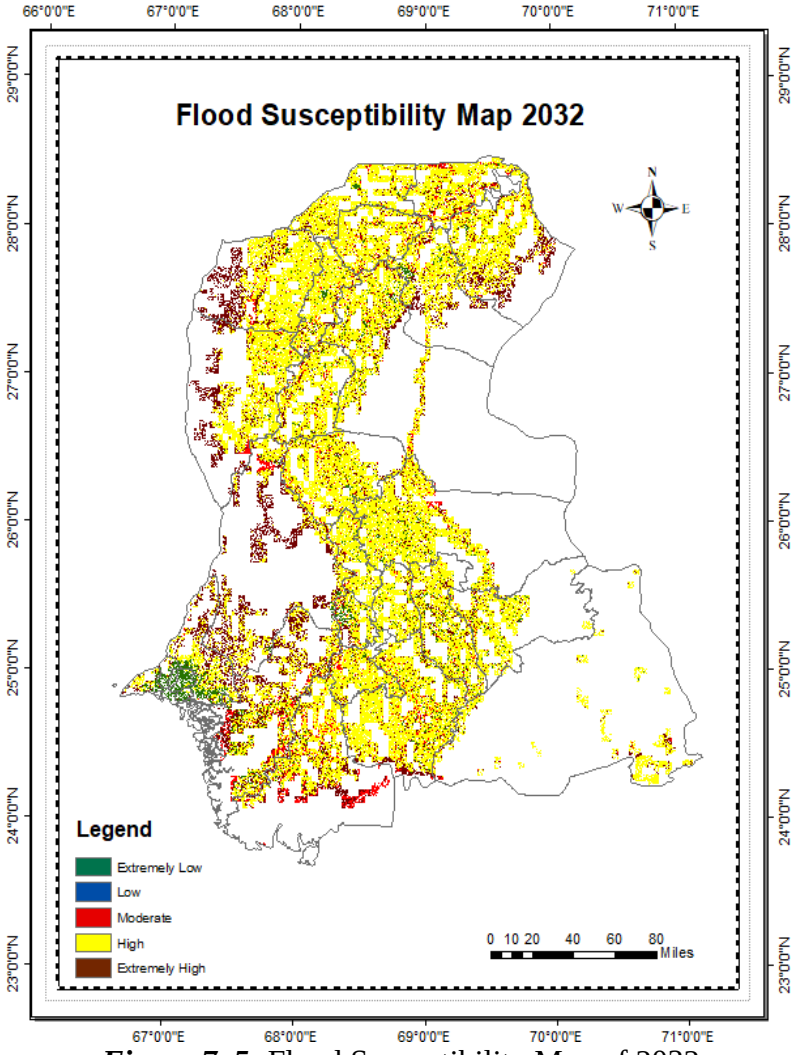


Figure 7. 5: Flood Susceptibility Map of 2032

Figure 7.6 illustrates the bar chart representing flood intensity. The bar chart clearly demonstrates that 33% of the overall flood affected region will experience an extremely high level of flooding, while 44% of the territory may endure a high level of flooding. Additionally, 5% of the areas can have a moderate level of inundation, 10% with a low level of flooding, and 7% of the areas will have minimally inundated.

Table 7. 6: District-wise Flood Extent (Areas in 100)

Districts	Affected Area of Districts (km^2)					Total Affected Area (km^2)
	Extremely Low	Low	Moderate	High	Extremely High	
Badin	0.12	2.15	0.12	0.21	0.12	2.73
Central Karachi	0.02	0.00	0.00	0.00	0.00	0.02
Dadu	0.04	1.41	0.02	0.48	9.05	11.00
East Karachi	0.08	0.01	0.00	0.00	0.00	0.09
Ghotki	0.04	1.37	0.09	8.38	13.04	22.91
Hyderabad	0.07	0.29	0.02	0.04	0.01	0.43
Jacobabad	0.03	0.90	0.02	12.07	0.02	13.03
Jamshoro	0.04	0.78	0.02	1.00	0.05	1.89
Kambar Shahadat Kot	0.04	1.54	0.02	0.52	8.05	10.17
Kashmore	0.03	0.90	3.03	14.10	5.08	23.14
Khairpur	0.10	1.54	5.11	10.17	1.05	17.96
Korangi Karachi	0.03	0.01	0.00	0.00	0.00	0.04
Larkana	0.04	0.76	0.02	8.06	5.03	13.92
Malir Karachi	0.18	0.40	0.03	0.27	0.01	0.88
Matiari	0.03	0.55	0.02	6.03	10.01	16.65
Mirpur Khas	0.10	1.08	0.07	0.11	0.02	1.38
Naushahro Feroze_	0.04	0.98	0.06	0.06	0.03	1.18
Sanghar	0.11	2.04	0.09	0.11	0.06	2.41
Shaheed Benazirabad	0.06	1.41	0.07	3.11	11.05	15.70
Shikarpur	0.03	0.91	1.04	14.07	5.04	21.10
South Karachi	0.06	0.06	0.01	0.02	0.00	0.15
Sujawal	0.05	1.07	0.08	0.30	0.27	1.77
Sukkur	0.05	0.85	0.04	12.21	9.03	22.19
Tando Allahyar	0.04	0.55	0.03	0.02	0.00	0.64
Tando Muhammad	0.03	0.63	0.03	0.03	0.03	0.76
Tharparkar	0.01	0.65	2.01	9.07	3.01	14.75
Thatta	0.12	1.09	0.12	0.63	0.19	2.15
Umer Kot	0.05	0.77	0.03	5.08	0.01	5.94
West Karachi	0.06	0.01	0.00	0.01	0.00	0.08
Total	1.73	24.68	12.21	106.16	80.28	225.05

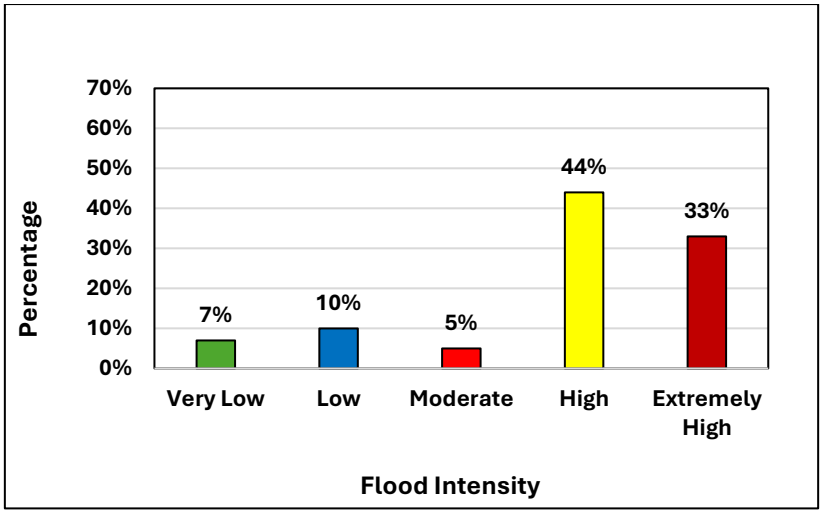


Figure 7. 6: The Percentage of 2032 Flood Intensity

7.3. Conclusion

Owing to the topographic, environmental and anthropogenic conditions of the lower Indus basin, it is prone to massive floods which is evident by the floods of 2010 and 2022 in this area. This chapter has discussed in detail (1) the contribution of flood conditioning factors, and their relative importance, (2) relation of flood factors with flood using FR model, (3) the flood susceptibility and extent at district level for 2022 flood and predicted 2032 flood. The key findings of the chapter are (1) Rainfall, LULC, temperature, and slope were the major contributors of 2022 flood where as for flood 2032 the expected major contributors are rainfall, LULC, temperature and distance from river. (2) In flood 2022, the highest flood hit districts were Dadu, Jamshoro, Kashmore, Khairpur, Larkana, Mitiari, Shaheed Benazirabad, Shikarpur and Sukkur with flood extent greater than 1000 km^2 area. Shikarpur experienced the most severe impact, with a total damaged area of 1900 km^2. The extent of the flood-affected territory in the Sindh province was 22,100 square kilometers, whereas, it is predicted that in 2032 flood, the districts of Dadu, Ghotki, Jamshoro, Kashmore, Khairpur, Larkana, Mitiari, Shaheed Benazirabad, Shikarpur, Tharparkar, and Sukkur would encounter flood inundation areas exceeding 1000 km^2. Kashmore is projected to endure the most substantial consequences, with a damaged region spanning 2300 square kilometers. The flood-affected area in the Sindh province is expected to cover a total of 22,500 square kilometers. In this connection, the proceeding chapter discusses the relation between LULC transitions and floods and compares and contrasts the flood of 2010 and 2022.

This chapter discussed the flood hazards at district level through the simulated flood model. The extent and depth of flood are different, as shown by the results, for each district

owing to the differences in the LULC. Thus, it is pertinent to measure the correlation between the LULC and causation of flood hazards. The next chapter discusses the correlation between flood and LULC transition.

Chapter 8
Spatial Measurement of Correlation between Flood and Land Use Land Cover Change (2010-2022)

8.1. Introduction

Rise in occurrence and intensity of hydrological threats, for instance sudden floods, can be attributed to the recent years of global climate alteration (Arnell and Gosling, 2016), as well as alterations in LULC and land management practices. Riverine floods, adhering to Hossain et al. (2013), result from the accumulation of water in rivers above their capacity due to intense precipitation and have detrimental effects on its social and economic standing due to floods. Therefore, it is crucial to examine and supervise the regions that have a significant likelihood of water accumulation. Various authors have addressed the subject of water accumulation and riverine floods in their scientific research (Hasan et al., 2023). Simultaneously, several studies have examined the relationship between changes in land use and various aspects of hydrological hazards. These include the increase in flood occurrences resulting from the change of natural vegetation to farmland, the rise in flood volume due to urban expansion, the intensification of runoff caused by deforestation, and the reduction in peak discharge frequency through afforestation (Rogger et al, 2017; Dang and Kumar, 2017). The diverse range of studies directions are linked to multiple causes that contribute to changes in land use, including urban expansion, agricultural intensification, deforestation, afforestation, and land abandonment, among others (Munteanu et al., 2017; Gan et al., 2018; Lieskovský et al., 2018). The significant impact of these changes in LULC on the intensification of hydrological dangers lies in resulting modification of the quantitative associations between elements of the water cycle, like interception, infiltration, or evaporation (Chen et al., 2009; Ali et al., 2011). The presence of vegetation plays a significant role in this context by exerting a strong influence on the process of evapotranspiration (Mao and Cherkauer, 2009; Chen et al., 2011) which in turn affects the balance of soil water.

Indus is a prominent river in Pakistan. The lower Indus plain has an area of 14.09 million hectares, which is occupied by the Sindh province. The province mostly uses its territory for agriculture, forestry, and pasture. The Indus River flows across Sindh, spanning a distance of 865 km from Guddu Barrage to the Arabian Sea. It serves as the sole water source for agriculture, forestry, and human consumption. The riverine tract and delta created by the River Indus hold great importance in the economy and environment of the Sindh province. The

country has had substantial economic growth over time as a result of agricultural expansion and later industrial development. The lower Indus basin is prone to riverine floods.

Geoinformation technologies, such as GIS and remote sensing, are valuable instruments for evaluating LULC changes and the potential for flash floods (Khosravi et al., 2016). Progress in remote sensing has facilitated the acquisition of satellite images with frequent revisits and varying spatial resolutions, depending on the sensor, at any location on the globe. Landsat sensors have a spatial resolution that is adequate to accurately describe the processes that impact land-use. GIS plays a crucial role in processing remotely sensed data to generate raster and vector data files, such as flood inventory and its predictors, for specific modelling methodologies.

Hence, this chapter intends to assess and evaluate the land use land cover (LULC) changes in the lower Indus basin over the period of 2010 to 2022 and the impact of this LULC transition in causation of riverine flood of 2022. The dataset used for the prediction of 2010 and 2022 floods in this objective is composed of 14 variables that are selected after feature selection. For this purpose, advanced machine learning models have been used which reduce the issue of heteroscedasticity to much extent. So, the redundant variables were removed in the first step. The chosen variables are relevant to the geography, anthropology and environment of the Sindh province. Secondly, the data was normalized. Normalization of the data mitigates the issue of heteroscedasticity. Thirdly, for modelling of the dataset, the ensemble model (RF-LMT-NBT-REPT) is used and as discussed by Breiman (2017) decision trees models are robust to the issues of heteroscedasticity and heterogeneity of data. Moreover, machine learning models are the modified version of the traditional models so they can tackle such issues automatically rather than traditional models. Thus, the models utilized in this objective addresses these properties.

The rest of the chapter has been divided into mentioned below sections; section 1 assesses the image accuracy of LULC, section 2 evaluates the LULC transition from 2010 to 2022, section 3 provides flood risk assessment of the lower Indus basin, and section 4 illustrates the correlation between LULC changes and causation of riverine floods in the study area.

8.2. Accuracy Assessment of LULC Imagery

In order for the classification data to be useful for change detection, it is imperative to have individual classifications (Owojori and Xie, 2005). In this research, the classification

accuracy is evaluated by constructing a confusion matrix. Therefore, using the data of confusion matrix depicted in table 8.1, the kappa index and classification accuracy are computed. The evaluation was conducted utilizing a random error matrix, and the stratified random approach was utilized to display the various land use and land cover (LULC) classes in the study area. A total of 500 sites were randomly selected on the classified map and compared with the ground truth data.

Table 8. 1: Confusion Matrix for LULC Classification Accuracy Assessment

Year	Overall Acc. (%)	Kappa Index (%)	Class	Ground Truth Samples (Pixels)								T.C. Pixel	User Acc. (%)
				S.H.	S.V.	G.L.	P.W.	C.L.	B.A.	B.	W.B.		
2010	92.47	91.28	S.H.	195	2	1	0	17	0	0	0	215	90.70
			S.V.	19	237	0	0	0	0	9	0	265	89.43
			G.L.	22	1	319	0	4	0	0	0	346	92.20
			P.W.	0	0	0	358	0	0	0	31	389	92.03
			C.L.	12	2	0	0	427	0	0	0	441	96.83
			B.A.	0	0	0	0	3	197	0	0	200	98.50
			B.	3	7	11	0	0	0	215	0	236	91.10
			W.B.	0	0	0	27	0	0	0	153	180	85.00
			T.C. Pixel	251	249	331	385	451	197	224	184	2272	
			Prod. Acc. (%)	77.69	95.18	96.37	92.99	94.68	100.00	95.98	83.15		

Year	Overall Acc. (%)	Kappa Index	Class	Ground Truth Samples (Pixels)								T.C. Pixel	User Acc. (%)
				S.H.	S.V.	G.L.	P.W.	C.L.	B.A.	B.	W.B.		
2022	92.99	91.83	S.H.	145	1	3	0	9	0	0	0	158	91.77
			S.V.	13	178	4	0	0	0	1	0	196	90.82
			G.L.	19	0	295	0	2	0	0	0	316	93.35
			P.W.	0	0	0	421	0	0	0	26	447	94.18
			C.L.	12	8	4	0	367	0	0	0	391	93.86
			B.A.	0	0	0	0	1	248	0	0	249	99.60
			B.	1	23	0	0	0	0	129	0	153	84.31
			W.B.	0	0	0	34	0	0	0	352	386	91.19
			T.C. Pixel	190	210	306	455	379	248	130	378	2296	
			Prod. Acc. (%)	76.32	84.76	96.41	92.53	96.83	100.00	99.23	93.12		

S.H.- Shrublands, S.V.- Savannas, G.L.- Grass Lands, P.W.- Permanent wetlands, C.L. Crop Lands, B.A.- Built-up areas, B.- Barren, W.B.- Waterbodies, T.C. Pixel- Total Class Pixel, User Acc.- User Accuracy, Prod. Acc.- Producer Accuracy

In the year 2010, the overall classification accuracy is 92.47%, and the Kappa index is 91.28%. In 2022, the overall classification accuracy is 92.99%, and the Kappa index is 91.83%. It is important to note that the values of user and producer accuracies are obtained for both analyzed years. Therefore, in 2010, the accuracy of the user varied from 85% for water bodies to 98.50% for built-up regions, whereas, the producer accuracy varied form 77.69% for shrublands to 100% for built-up areas. When considering the year 2022, we can see that the user accuracy varied from 84.31% for barren land to 99.60% for built-up lands. Similarly, the accuracy of the producer ranged from 76.32% for shrublands to 100% for built-up areas.

8.3. Land Use Land Cover Change Analysis

The LULC changes from 2010 to 2022 in the lower Indus basin by utilizing satellite imagery classification is discussed below.

8.3.1. Transition Detection

The Markov matrix (Table 8.2) has been utilized to determine the direction and extent of each LULC conversion. The matrix indicates that the most changed classes are shrublands, grass lands, crop lands, and barren areas. Shrublands and grassland areas experienced a general expansion, primarily through the conversion of barren, built-up, croplands and savannas. However, some losses also occurred due to conversion of shrublands to grasslands and barren lands and grasslands to shrublands and croplands. The overall equilibrium indicated a net growth of 2530 sq.km in shrublands and a decline of 5160 sq.km of grass lands during the course of the analyzed period. The transitional croplands underwent a gain of 4053 sq.km and an abandonment of 3361 sq.km of an area due to conversion of croplands to shrublands and barren areas. The overall transition of crop lands is a gain of 692 sq. km mainly due to conversion of grasslands to croplands. The barren lands depicted a total gain of 1904 sq.km mainly due to conversion of shrublands to barren lands. The built-up area showed an opposite trend compared to the regions shrublands, croplands and barren areas, with a decline of approximately 244 sq.km, mainly due to transition of built-up lands to shrublands, barren and croplands. The largest expansion of urban built-up areas are revealed in Karachi and Hyderabad (shown with pink color).

Based on the Markov matrix, the total area of developed land fell by around 3600 sq.km. The primary transformation that occurred is the conversion of land into shrublands, croplands and barren lands, including an area of total 5126 sq.km, primarily as a result of deforestation which is evident from huge loss area of grasslands and permanent vegetation areas (5160 sq.km). These transitions emphasize significant and intricate changes in land use and land cover that happened in the Sindh province from 2010 to 2022, including all land use and land cover categories. This circumstance supports our prediction that there were substantial alterations in runoff and an increased possibility for floods. The detailed transition map has been shown in figure 8.1. In the transition map, the yellow, red, purple and pink colors show the transitions towards shrublands, barren lands, croplands, and urban built-up areas, respectively.

Table 8. 2: Markov Matrix for the Period 2010-2022

2010/2022	S.H.	S.V.	G.L.	P.W.	C.L.	B.A.	B.	W.B.	Area Loss (sq.km)
S.H.	---	1.88	1755.84	0.45	218.78	22.32	3425.24	4.96	**5429.46**
S.V.	1.01	---	111.68	0	120.20	0.76	2.35	0.05	**236.05**
G.L.	6204.06	69.13	---	17.01	3633.35	30.41	918.73	57.14	**10929.82**
P.W.	0	0	0	---	0	0.18	23.56	29.33	**53.07**
C.L.	90.70	37.32	3068.81	1.99	---	53.06	100.96	8.64	**3361.49**
B.A.	24.57	1.03	24.23	0.48	32.45	---	5.50	272.19	**360.44**
B.	1635.95	0.72	767.07	54.30	41.30	5.78	---	272.19	**2777.32**
W.B.	3.19	0.37	41.21	38.39	7.82	3.15	205.38	---	**299.52**
Area Gain (sq.km.)	**7959.47**	**110.46**	**5768.84**	**112.63**	**4053.90**	**115.66**	**4681.72**	**644.50**	

8.3.2. Spatial Measurement of Magnitude of LULC Transformation

To accurately measure the changes in land use and land cover, the TR-DLUI has been calculated at a grid-cell level of 1 m^2 (Figure 8.2). This index intends to spatialize and quantize the LULC changes from 2010 to 2022. The spatial distribution of this index emphasizes the variations in the magnitude of changes in LULC in research area. An overall trend indicates a gradual rise in the intensity of LULC change from north to south. This change coincided with the transition from grasslands and permanent vegetation areas to the shrublands, crop-lands and barren area. The deforestation contributed to this transition. Adhering to Tagar and Shah (2015) the construction of dams in upper areas of basin, population explosion, mismanagement of wetlands and unmanaged industrial wastes led to this transformation which in-turn poses a great environmental degradation and adverse climatic conditions. Furthermore, Abbasi et al. (2011) asserted that human activities, such as livestock farming, excessive grazing, heavy loads, and unauthorized tree cutting, have played a significant role in causing deforestation issue. The riverine forests are undergoing desertification, with a significant portion of the area being converted for agricultural use.

Null values of TR-DULI indicator, shows no land transition in LULC, accounted for only 0-5.9% of the total area. These null values were unevenly distributed and consisted of multiple dispersed locations within the lower Indus basin. The grid cells showing area of 6-15.9% are mainly comprised of transformed lands in Karachi (Central, West, East, Korangi and Malir) and Mirpurkhas. These values are evenly distributed across the entire area and depicts transformation towards urban lands, shrublands, grasslands and croplands took place. Approximately 9% of the region saw changes in the range of 16% to 25.9%. These changes

resulted in Badin, Umerkot, Shaheed Benazirabad, Kambar Shahdat Kot and Sukkur districts, and the conversion of lands to croplands, shrublands and barren lands took place. About 9% conversions took place in districts of Sujawal, Thatta, Sanghar, Ghotki and Dadu and the conversions are barren lands, crop lands and permanent wetlands (only in Sujawal). The largest transformation of 20% took place in districts of South Karachi, Tharparker, Tando Muhammad Khan, Hyderabad, Jamshoro, Mitiari, Tando Allahyar, Khairpur, Naushero Feroze, Larkana, Jacobabad, Kashmore and Shikarpur. This transformation led to expansion in croplands and barren lands.

These transformations tell the sensitivity of each district to flood. Thus, it can be assessed that deforestation and conversion of permanent vegetation areas to croplands, shrublands and barren areas make the lower Indus basin prone to floods.

8.4. Flood Risk Assessment

The following discussion provides the details of 2010 and 2022 floods in the lower Indus basin. A comparison has also been provided in this regard between the two flood episodes.

8.4.1. Flood Conditioning Factors and their Relative Importance

Table 8.3 illustrates a comparison between flood contributing factors for 2010 and 2022 flood. It is evident from the table that major contributors of flood in 2010 were slope, distance from river, LULC, stream power index and rainfall, whereas, for 2022 flood, the major contributors were rainfall, LULC, temperature, slope and stream power index. Adhering to Hashmi et al. (2012), the major cause of 2010 flood in the lower Indus basin was rainfall, deforestation and lack of dams and barrages or the water storage capacity. Moreover, the official 2010 flood report published by the Ministry of Water and Power states that the major cause of 2010 flood was heavy and massive precipitation and the lack of storage capacity in the upper and lower Indus basins. Further, Yaseen et al. (2022) have conducted a variable sensitivity study to assess flood susceptibility mapping in Karachi. The researchers determined that the primary elements leading to the 2022 flood in Karachi were land use land cover, rainfall, and elevation. Figure 8.3 depicts the bar charts for the statistics provided in table 8.3.

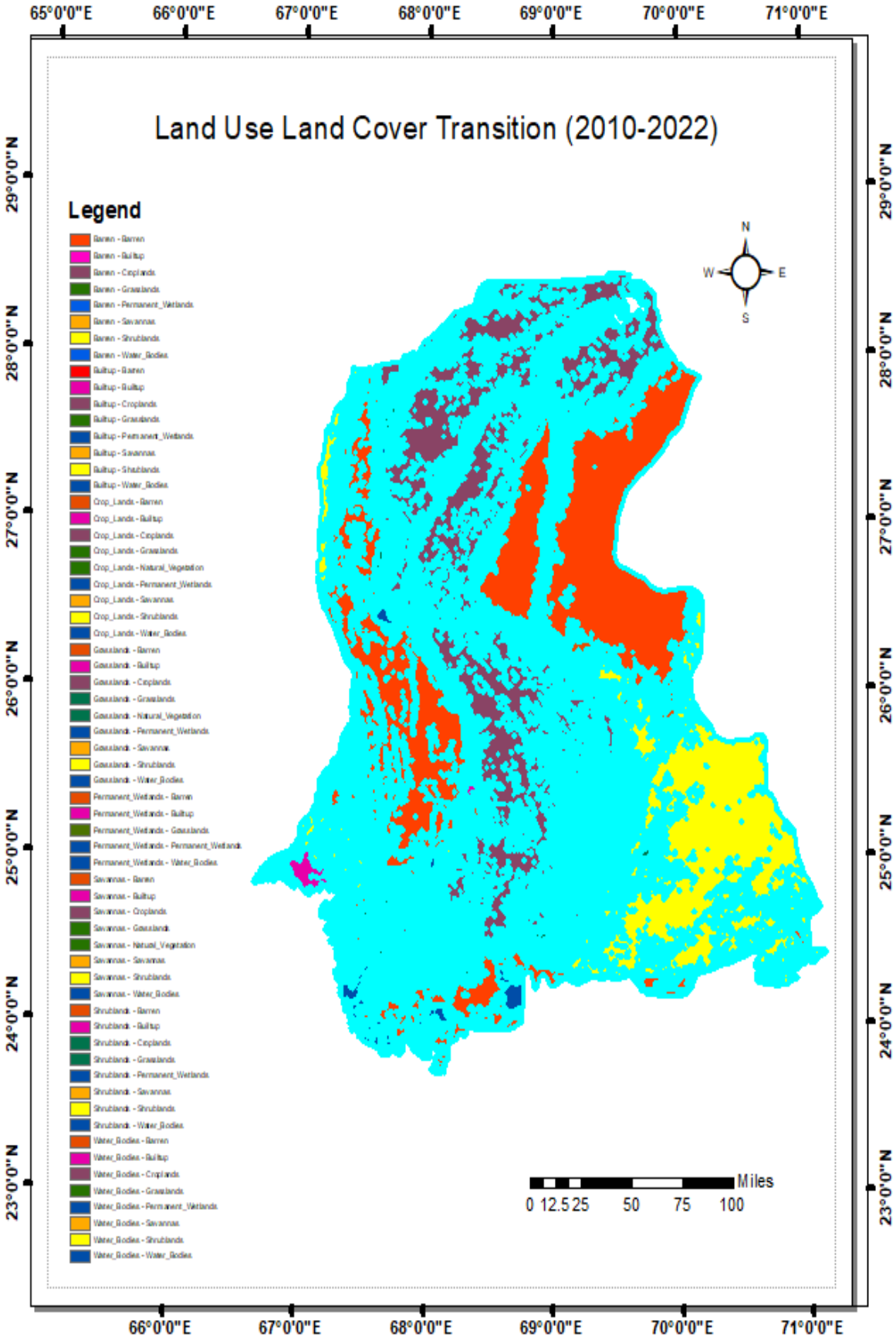


Figure 8. 1: Land Use Land Cover Transition

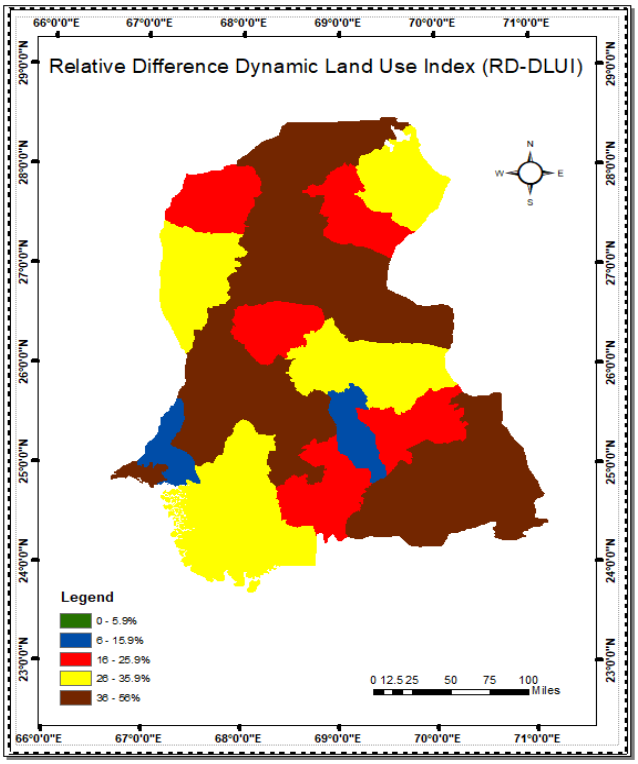


Figure 8. 2: The Total Relative Difference Dynamic Land Use Index (Values/km2)

Table 8. 3: Relative Importance of Conditioning Factors

S.No.	Flood Conditioning Factors	Gain Ratio (GR)	
		2022	2010
1.	Rainfall	0.985	0.925
2.	Land Use Land Cover	0.983	0.951
3.	Temperature	0.977	0.916
4.	Slope	0.962	0.977
5.	Stream Power Index	0.951	0.947
6.	Distance from River	0.943	0.958
7.	Topographic Ruggedness Index	0.936	0.925
8.	Lithology	0.93	0.895
9.	Stream Density	0.925	0.961
10.	Sediment Transport Index	0.922	0.903
11.	Openness Positive	0.904	0.876
12.	Topographic Wetness Index	0.901	0.894
13.	Openness Negative	0.897	0.849
14.	Population Density	0.894	0.856

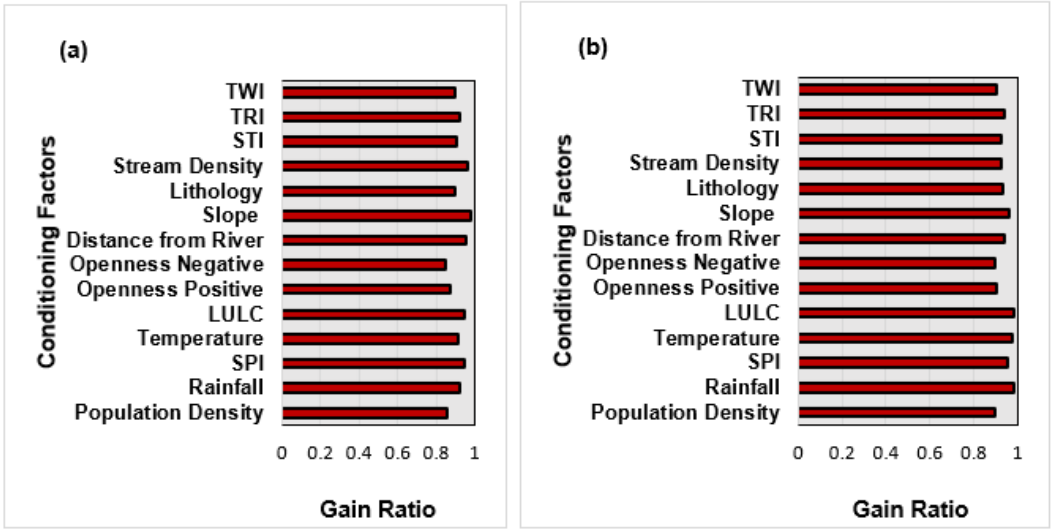


Figure 8.3: Relative Importance of Conditioning Factors in Causing Flood (a) Flood 2010 (b) Flood 2022

8.4.2. A Comparison of Spatial Relationship between Conditioning Factors and Flood Episodes of 2010 and 2022

A FR value of 0 signifies the lack of flood grid cells. The data in Table 8.4 shows that all of the FR values are non-zero, indicating that the chosen variables had a substantial influence on the occurrence of the 2010 and 2022 flood in the study area. The class bboundary ranging from -13.81- -11.46 (FR = 1.74) shows highest 2010 flood grids while -11.47-10.46 has the highest 2022 flood grids (FR = 1.60) in the stream power index. Similarly, the slope between 0.66 and 7.74 (FR = 1.66) degrees demonstrates a notable susceptibility to 2022 flooding whereas, that of between 7.75-78.14 shows the highest susceptibility for 2010 flood. Additionally, the stream density, sediment transport index, Topographic Wetness index, and topographic ruggedness index all show a higher susceptibility to flooding due to their high FR values compared to the class boundaries of their respective variables. The stream density ranges from 3.38 to 34.43 (FR= 1.02) for 2022 and between 15.67-34.43 (FR =9.02) for 2010 flood, the sediment transport index ranges from 0.1 to 11.79 (FR=1.16) for 2022 and between 11.80-23.59 (FR = 161.32) for 2010 flood, the Topographic Wetness index ranges from 7.67 to 8.51 (FR=1.35) for 2022 and between 10.29-21.53 (FR = 13) for 2010 flooding, and the topographic ruggedness index ranges for 2022 and 2010 flood from 0.67 to 0.88 (FR=1.42) and 0.11-0.33 (FR = 1.47), respectively. The most extensive flood pixels for openness positive are seen inside the class limits ranging from 0.71-1.36 with FR value equal to 41.68 for 2010 and 1.55 to 1.66 with FR value of 1.22 for 2022. Additionally, significant flood pixels were also found within the class boundaries of 1.61 to 1.65 for negative topographic openness with FR value of 1.28

for 2022 and within the range of 1.31-1.49 (FR =12.73) for flooding of 2010. Furthermore, for 2022 flood, the areas with a population density between 45000 and 60000 are very susceptible to flooding, as evidenced by the highest FR rating of 184.56 and for 2010 flood, the population dense areas within 15000-30000 density were more susceptible (FR = 2.369). When considering land use and land cover, croplands and built up regions, FR=1.84 for 2022 and FR= 7.87 for 2010 are significantly more prone to floods than other forms of land covers. The areas situated within a range of 80000-100000 meters (2022) and 20000-40000 (2010) from the river are highly prone to flooding, with a FR values of 40.53 and 7.68, respectively. Districts that have rainfall levels ranging from 3000 to 4000 mm (2022) and 4000 – 6000 mm (2010) with temperatures between 27.7 and 28.5 degrees Celsius (2022) and 26.7-27.0 degrees Celsius (2010) are more prone to flooding. These districts have FR of 1.36 and 1.55 for 2022 flood and 5.90 and 2.94 for 2010 flood, respectively. Ultimately, the geological makeup of Sindh suggests that areas with loamy soil are more prone to floods, with FR of 1.66 and 17.59 for 2022 and 2010 floods, respectively. This assessment exhibits a comprehensive comprehension of the influence of each category of all the contributing elements on the occurrence of floods in the research region, as depicted in table 8.4.

Table 8.4: Spatial Relationship between Conditioning Factors and 2022 Flood by Frequency Ratio (FR) Model

S. No.	**Factor**	**Class Boundaries**	**Factor Pixel in Class Domain**	**Flood Pixel in Class Domain**	**FR**	**Flood Pixel in Class Domain**	**FR**
				2010		**2022**	
1	Stream Power Index (SPI)	(i)-13.81- -11.46	17173616	10.56	1.74	8.69	1.35
		(ii)-11.47-10.46	27360883	16.67	1.72	16.38	1.60
		(iii)-10.47--8.11	22415978	10.52	1.32	2.20	0.26
		(iv)-8.12- -2.59	41312073	3.33	0.23	18.89	1.22
		(v)-2.60- -10.39	48290040	14.37	0.84	12.55	0.69
2	Slope (Degree)	(i) 0- 0.59	44635376	29.15	1.8	5.77	0.73
		(ii) 0.60-0.65	17201182	14.32	2.4	4.96	0.25
		(iii) 0.66 – 1.25	34114962	1.95	0.2	5.10	1.66
		(iv) 1.26-7.74	51442541	1.11	0.1	32.45	1.66
		(v) 7.75-78.14	9158530	8.90	2.7	10.46	1.18
3	Stream Density	(i) 0.001-3.37	21515364	3.57	0.22	16.52	0.97
		(ii) 3.38-6.88	16836947	31.99	2.53	13.64	1.02
		(iii) 6.89-10.66	21248962	8.81	0.55	17.02	1.02
		(iv) 10.67-15.66	12910814	1.72	0.18	10.42	1.02
		(v) 15.67-34.43	1381752	9.35	9.02	1.12	1.02
4	Sediment Transport Index (STI)	(i) 0	108527933	5.32	0.11	52.21	1.00
		(ii) 0.1-11.79	894116	11.44	28.07	0.50	1.16
		(iii) 11.80-23.59	6718929	19.89	6.49	3.62	1.11

		(iv)23.60-47.18	5289096	9.56	3.96	2.35	0.92
		(v)47.19-3007.97	125362	9.22	161.32	0.05	0.81
5	Topographic	(i) 1.89-6.82	2813924	11.25	11.29	0.35	0.33
	Wetness Index	(ii) 6.83-7.66	57157967	4.59	0.23	12.94	0.60
	(TWI)	(iii) 7.67-8.51	79499943	21.72	0.77	40.36	1.35
		(iv) 8.52-10.28	16865872	16.88	2.83	5.02	0.79
		(v) 10.29-21.53	214885	0.99	13.00	0.05	0.57
6	Topographic	(i) 0.11-0.33	30194799	16.81	1.47	16.30	1.34
	Ruggedness	(ii) 0.34-0.44	35922254	10.03	0.74	11.36	0.79
	Index (TRI)	(iii) 0.45-0.55	23687923	2.82	0.31	5.52	0.58
		(iv) 0.56-0.66	30310361	14.97	1.30	10.62	0.87
		(v) 0.67-0.88	26137316	10.81	1.09	14.92	1.42
7	Openness positive	(i) 0.71-1.36	714980	10.55	41.68	0.09	0.32
		(ii) 1.37-1.46	2605478	6.61	7.17	0.35	0.36
		(iii) 1.47-1.51	8857763	15.82	5.05	0.78	0.24
		(iv) 1.52-1.54	28525569	11.33	1.12	4.55	0.43
		(v) 1.55-1.66	115848801	11.11	0.27	52.95	1.22
8	Openness	(i) 0.69-1.30	335642	0.29	2.45	0.04	0.31
	negative	(ii) 1.31-1.49	6462396	29.13	12.73	0.70	0.29
		(iii) 1.50-1.55	32241284	1.12	0.10	5.40	0.45
		(iv) 1.56-1.60	71037479	16.67	0.66	30.31	1.14
		(v) 1.61-1.65	46475789	8.22	0.50	22.27	1.28
9	Population	(i) < 15000	117642263	10.61	0.84	33.04	74.32
	density	(ii) 15000-30000	2732832	10.14	2.39	1.82	176.10
		(iii)30000-450000	23188548	10.74	0.68	15.72	179.34
		(iv) 45000-60000	10785939	12.83	0.86	7.52	184.56
		(v) 60000-80000	1042007	11.10	1.44	0.63	159.37
10	Land Use Land	i) Savannas	38843684	5.44	3.07	6.37	0.44
	Cover (LULC)	(ii) Grasslands and Permanent Wetlands	22189911	9.96	0.30	13.01	1.56
		(iii) Croplands and Urban buildup lands	43553122	12.67		30.05	1.84
		(iv) Natural Vegetation and Barren lands	1186205	21.11	1.34	0.72	1.62
		(v) Water bodies and Shrublands	50765371	6.26	2.06	8.57	0.45
11	Distance from	(i) 0 - 20000	106848894	10.66	0.28	4	0.09
	River (m)	(ii) > 20000-40000	28023772	16.67	7.68	4	0.39
		(iii)> 40000-60000	14640488	10.98	6.12	4	0.70
		(iv) >60000-80000	4612350	10.90	2.68	9	5.32
		(v)> 80000-100000	2497734	6.22	1.04	38	40.53
12		(i)500 -1000	37963661	11.23	1.41	18.52	1.30
	Rainfall (mm)	(ii) > 1000 – 2000	20444966	11.09	0.61	8.90	1.16
		(iii) >2000 – 3000	39281895	11.11	0.62	13.66	0.93
		(iv) > 3000 – 4000	37348449	11.04	1.16	14.79	1.36
		(v) > 4000 - 6000	21469847	10.96	5.90	2.86	0.36
13	Temperature	(i) 2.59-26.6	1629994	8.47	0.90	0.14	0.23
		(ii) 26.7-27.0	19968050	20.23	2.42	1.55	0.21
		(iii) 27.1-27.2	17402562	10.18	0.97	4.38	0.67
		(iv) 27.3-27.6	62573185	6.33	0.44	20.67	0.88

		v) 27.7-28.5	54960498	10.22	0.79	31.99	1.55
14	Lithology	(i) Clay	76676321	10.15	0.39	11.45	0.41
		(ii) Loam	7295255	17.59	7.06	4.37	1.66
		(iii) Silt Loam	72030007	16.67	0.68	42.31	1.62
		(iv) Clay Loam	6264870	11.02	5.15	0.60	0.26

8.4.3. Flood Susceptibility Prediction and Analysis

The ensemble model (NBT-REPT-RF-LMT) has been employed for flood susceptibility assessment for the floods of 2010 and 2022. The dataset for flood points was equally divided into two samples each consisting 5500 flood points and 5500 non-flood points for both the floods separately. The formulated maps in figure 8.4, through the stated model, have been used for evaluation of the flood intensity in the lower Indus basin at district level for the two major riverine floods. Moreover, to assess the model's prediction capability, AUC analysis has been conducted in figure 8.5. The flood of 2010 shows 97.8% and 95.1% of AUC for training and testing datasets, respectively, while, 2022 shows 99.5% and 96.5% of AUC for training and validation datasets.

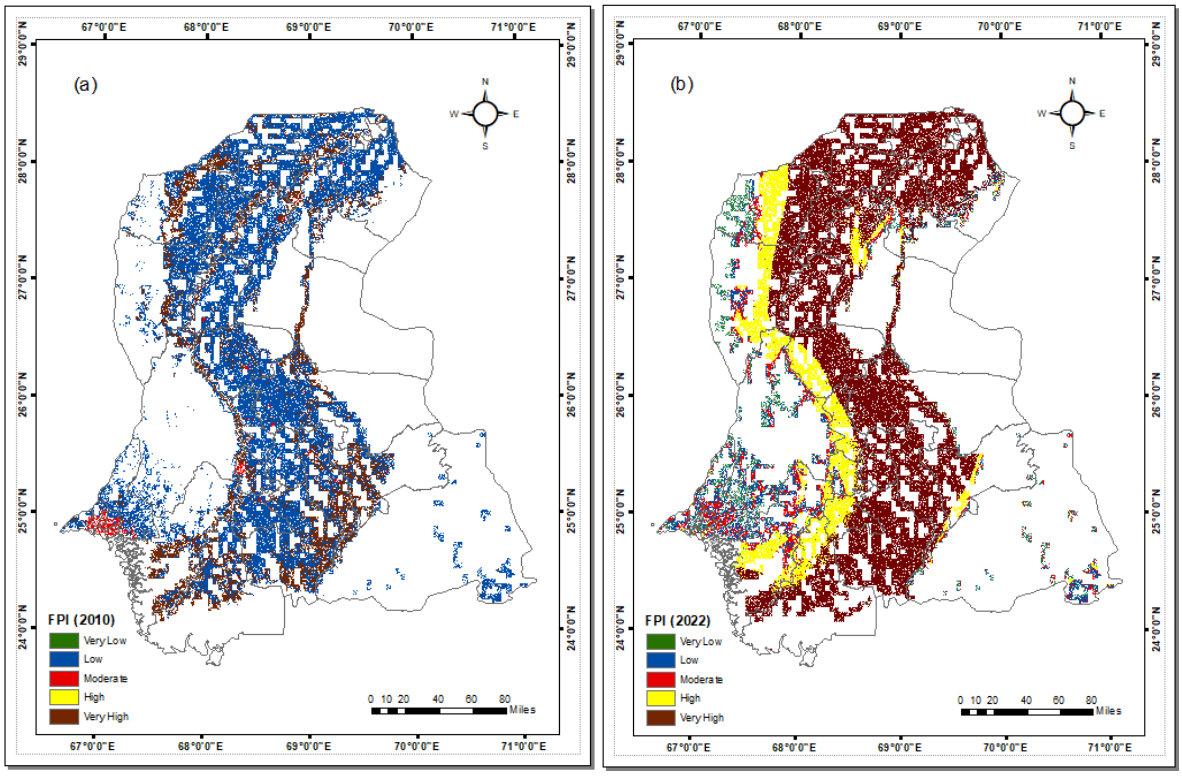


Figure 8. 4: Flood Susceptibility (a) Map for 2010 and (b) Map for 2022

8.4.4. Spatial Analysis of Magnitude of Flood Potential

The RD-FPI has been computed at a grid-cell resolution of 1 m^2 (Figure 8.6) in order to precisely assess the variations in the two flood episodes. The purpose of this index is to assess

and calculate the spatial and quantitative alterations in floods of 2010 and 2022. The spatial calculations of this index show the variations in the extent of changes in flood intensity in the lower Indus basin. Generally, the pattern depicts a progressive rise in the intensity of flood (low to extremely high) from the northern regions to the southern districts. This alteration occurred simultaneously with the shift from grasslands and areas of permanent vegetation to shrublands, cultivated lands, and barren places. This transformation was facilitated by the deforestation. As a result, there is significant environmental degradation and adverse climatic conditions that led to amplify the flood intensity. Sujawal, Dadu, Thatta, Jamshoro, and Sanghar had 21-27% of altered flood intensity. The flood inundations in the figure 8.4 depicts that these districts have experienced high and very high flood in 2022 which was low in 2010 except for the areas which are nearby Indus river (these areas had very high flood). Tharparkar and Khaipur had 28-34% variation in flood. The flood inundation of Tharparkar reveals that it had only low flood in 2010 whereas in 2022, it faced flood inundations ranging from very low to high flood. Khairpur had low and very high flood in 2010 but in 2022 it was hit by very high flood. Jacobabad, Kashmore, Larkana, Shikarpur, Noushero Feroz, Mitiari, Tando Allahyar, Tando Muhammad Khan, Mirpurkhas and Karachi (except South Karachi) had 7-13% variations. Ghotki, Kambar Shahdat Kot, Sukkur, Shaheed Benazirabad, Badin and Umer Kot had 14-20% flood variation between the two years. Lastly, 0-6% change can be seen in South Karachi and Hyderabad.

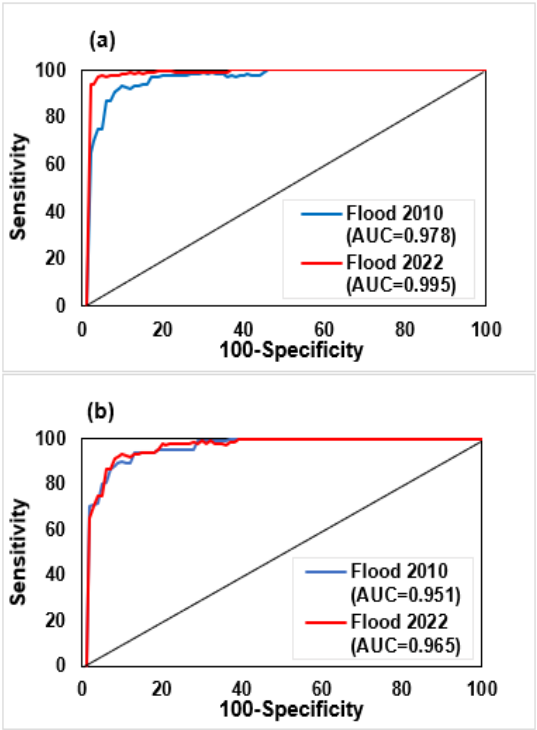


Figure 8. 5: ROC Curve (a) Training dataset (b) Testing dataset

8.5. Statistical Analysis

Figure 8.7 illustrates the association between alterations in the Land Use and Land Cover (LULC) and the likelihood for riverine flooding. Upon examination of Figure 8.7 it is evident that the areas in close proximity to the Indus River had significantly low correlation coefficient values (ranging from 0 to 0.250). This suggests that there is minimal link between land use and land cover (LULC) changes and the occurrence of floods in these regions. Due to the proximity of the lands adjacent to the river, the change in land use and land cover (LULC) has minimal impact on the occurrence of floods along the river's banks. Areas that are not in close proximity to Indus, showed correlation coefficients between 0.251 and 0.500, indicating weak linkage, and between 0.501 and 0.750, indicating moderate association. The districts that are located far away from the Indus demonstrated a significant correlation coefficient (0.751-1.00), indicating a high level of association between land use and land cover changes and the risk for riverine flooding.

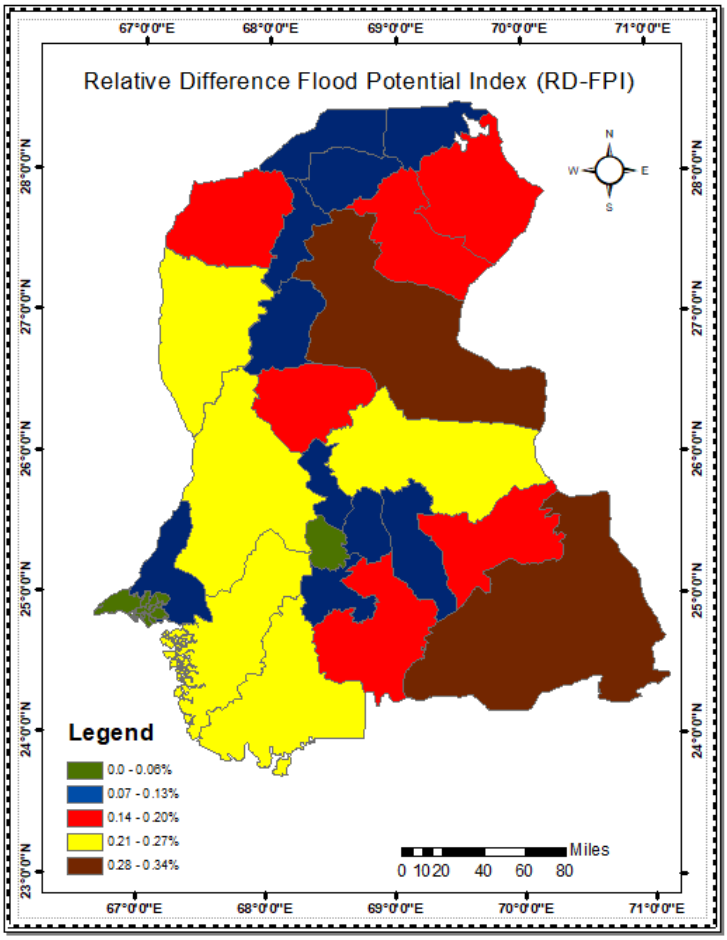


Figure 8. 6: Relative Difference Flood Potential Index

The GWR values were further confirmed using Moran's I index. The Moran's I statistic is used to quantify the spatial autocorrelation by evaluating the degree of similarity between a spatial entity and its neighboring objects. The positive value of Moran's index (0.0125) indicates that the coefficient values are grouped. The z-score corresponds to the calculated standard deviation of 2.691. The p-value represents the probability that random processes are the cause of the observed spatial pattern. The p-value is 0.0398, indicating that we cannot reject the null hypothesis.

The finding indicates that the correlation pattern was concentrated, and there were numerous places within the research area that exhibited similar correlations. The geographic pattern found in the study area is typically not attributable to random chance.

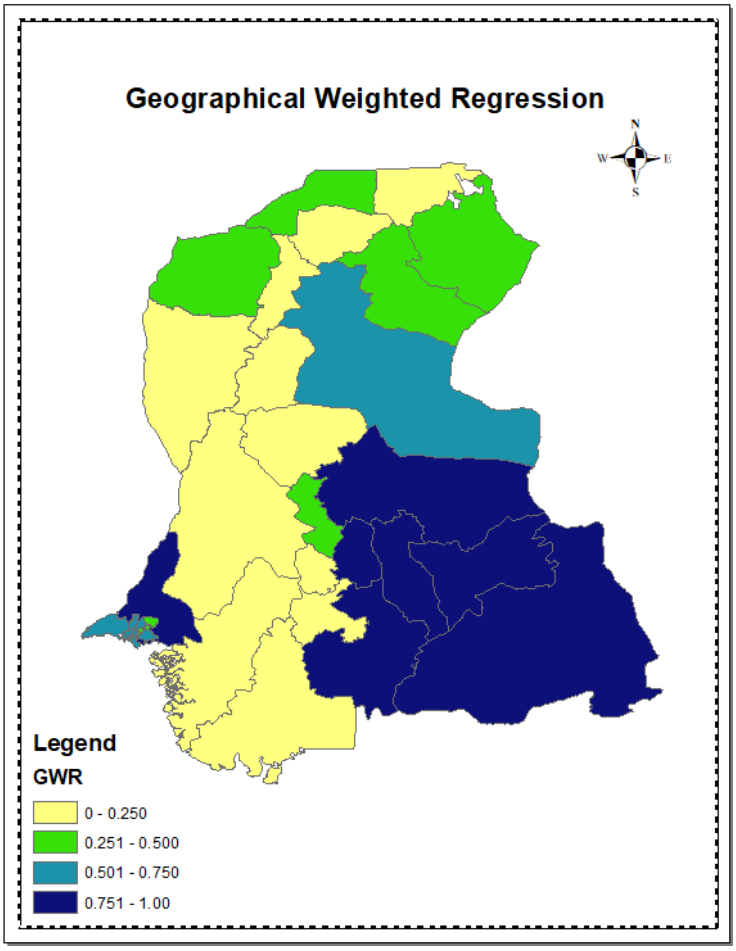


Figure 8. 7: Geographical Weighted Regression

8.6. Conclusion

Pakistan lacks adequate data for this research endeavor. Floods are an inevitable natural occurrence. Nevertheless, by acquiring a comprehensive understanding of the geographical attributes of floods, it becomes feasible to mitigate the

detrimental effects caused by floods. This chapter investigated the relationship between changes in land use and land cover (LULC) and changes in the risk for riverine flooding along the lower Indus basin from 2010 to 2022. By analyzing the changes in land use and land cover (LULC), it is possible to predict flood-prone locations. This knowledge can be used to effectively reduce the danger of flooding.

The findings of this chapter indicate that changes in land use and land cover (LULC), especially in areas located at a considerable distance from the river, significantly influence floods in rivers. Hence, it may be concluded that changes in land use and land cover (LULC) have a substantial impact on the increase in flooding in rivers. Policymakers are expected to utilize the framework to effectively mitigate flood risk through appropriate planning. The chapter advocates the integration of machine learning ensemble model (NBT-REPT-RF-LMT) with GWR (Geographically Weighted Regression) to perform relational statistical analysis, specifically addressing spatial nonlinearity. This methodology may assist other researchers to ascertain the correlation between intricate and nonlinear spatial variables, including not just land use and land cover changes but also flood potential.

This chapter has provided the justification for LULC changes over the years and the causation of flood in 2022. The next chapter investigates the quantitative assets damages caused due to flood. For this purpose, the research utilizes only physical or tangible assets for the monetary assessment of flood damages.

Chapter 9
Quantification of Flood Associated Economic Damages

9.1. Introduction

The true advantages of doing economic research on flood damage mitigation initiatives may be challenging to ascertain, both prior to and during their initiation, as these benefits may not be readily apparent or tangible. Estimating the probable actual losses caused by floods is a crucial and delicate aspect of reducing flood damage. These forecasts are directly beneficial to the affected parties and communities. An extensive analysis is necessary for flood damage assessment, as it encompasses both engineering and economic factors. Engineering studies typically involve conducting hydraulic and hydrologic analyses to anticipate floodplain inundation in protected lowlands. These analyses are essential for flood mitigation planning and assessing flood damage. Economic studies measure the extent of flood damage by analyzing different statistical data and utilizing indices that act as substitutes for economic value in areas impacted by floods (Yi et al., 2010).

Adhering to Ministry of Planning Development and Special Initiatives report on 2022 floods[11], Pakistan's economy is anticipated to be significantly adversely affected by the floods. The total damages amount to PRs3.2 trillion, which is equivalent to 4.8% of the gross domestic product (GDP) for fiscal year (FY) 2022, ending in June 2022[12]. Within the various economic sectors, agriculture and industry have experience almost 25% of the damages each, while services are expected to face over 50% of the damages. The whole loss is approximated at PRs3.3 trillion. The predicted recovery and reconstruction needs are expected to be substantial, amounting to 1.6 times the budgeted national development spending for FY2022[13]. According to Qamer et al., The agricultural region encompasses around 4.9 million hectares, with food crops occupying 60% and cash crops occupying 25% of the total cultivated land. The primary agricultural products cultivated in the Sindh region during the summer (kharif) season include rice, sugarcane, cotton, beans, various vegetables like as tomato, chilli, onion, okra, and cucurbitaceous vegetables, as well as fruits like mango, banana, and papaya. The entire cultivated area is primarily dominated by three cereal and industrial crops: rice, cotton, and sugarcane. These crops make up the majority of the cultivated land. In contrast, horticulture

[11] Pakistan Floods 2022, Post Disaster Needs Assessment
[12] GDP at market prices at current prices for FY2022 (PRs66.9 trillion).
[13] According to the FY2022 Federal Budget, the total national Public Sector Development Programme allocation is PRs2,158 billion

crops only account for a little portion, approximately 7%. At the national level, the province contributes 16% to wheat output, 42% to rice production, 23% to cotton production, and 31% to sugarcane production.

Thus, there is a need to assess the economic damages caused due to floods by using advanced hydraulic modelling. This chapter aims to estimate and evaluate the flood associated agricultural, industrial, residential, infrastructural, medical units and educational institutions damages by employing the inundation depths obtained from our simulated ensemble model at district-level.

The chapter is divided into following sections; first section discusses the economic damages, of each category, caused due to 2022 flood and the next section provides damages assessment for projected 2032 flood.

9.2. Quantification of Flood Damages for 2022 Flood

In this section, the flood extents and depths, obtained by simulating the ensemble model (NBT-REPT-RF-LMT) have been utilized for quantification of the flood various economic flood damages. For this purpose, the simulated floods and the land use dataset were overlaid in GIS to determine the assets affected by flooding and their related depths of inundation. The following sections provide damages assessment of 2022 flood.

9.2.1. District-level Agricultural Losses Estimation

In this research, three major Kharif crops have been considered for the estimation of agricultural losses owing to the fact that these crops are sown/harvested in June-August, which were the peak 2022 flooding months (Qamer et al., 2022). In this context, cotton, sugarcane and rice are considered for damages quantification and evaluation.

9.2.1.1. Cotton Crop Damages

Table 9.1 depicts the district-level cotton crop damages. In the table total cotton cultivated area is illustrated at district level. The total cotton cultivated area in Sindh is 47,4818 hectors. The figures show that Ghotki has the highest cotton cultivated area, followed by Khairpur and Sanghar with 94,173, 82,890, 78,405 hectors area, respectively. The table provides total normal year production of cotton. 2021 is the normal year considered for this research. The figures show that Sanghar had the highest cotton production (9,97,679 bales),

Ghotki had 6,04,891 bales and Khairpur produced 4,17,744 bales of cotton in the 2021. The normal year's cotton production is overlaid on the model estimated inundation depths to calculate the total flood damaged crop. The total estimated flood affected cotton crop was 34,42,718 bales and Sanghar had the largest crop damage of 9,75,358 bales. The total estimated economic value (PKR) of the cotton crop is 108325139805 with a GDP loss of 1949852516. The total damaged cotton crop due to flood was 88.24% of the total normal year's production which depicts a huge loss to the economy of Pakistan in terms of cotton crop damages. An analogous cotton crop damages have been presented by Qamer et al., 2022. This research provides an additional district-level analysis of total estimated flood damaged crop and GDP losses (in PKR). Figure 9.1 depicts the cotton crop flood damage estimates. The figure confirms the data illustrated in the table 9.1.

Table 9. 1: Cotton Crop Damages

Districts	Total Crop Cultivated Area (Hectors)	Total Normal Year Production (Bales)	Total Estimated Flood Damaged Crop (Bales)	Total Estimated Flood Damaged Crop (PKR)	Total Estimated GDP Loss (PKR)
Badin	6216	87290	85287	2683555455	48303998
Central Karachi	0	0	0	0	0
Dadu	11205	57462	27581	867836165	15621051
East Karachi	0	0	0	0	0
Ghotki	94173	604891	518314	16308750010	293557500
Hyderabad	4542	71116	63348	1993244820	35878407
Jacobabad		0	0	0	0
Jamshoro	12201	54633	46438	1461171670	26301090
Kambar Shahdad Kot	0	0	0	0	0
Kashmore	0	0	0	0	0
Khairpur	82890	417744	370164	11647210260	209649785
Korangi Karachi	0	0	0	0	0
Larkana	136	744	59.52	1872797	33710
Malir Karachi	374	1104	264.96	8336966	150065
Matiari	35792	221262	206301	6491260965	116842697
Mirpur Khas	19722	190046	189069	5949056085	107083010
Naushahro Feroze	23624	132776	132776	4177796840	75200343
Sanghar	78405	997679	975358	30689639470	552413510
Shaheed Benazir Abad	63585	329900	293869	9246588085	166438586
Shikarpur	82	175	18	550638	9911
South Karachi	0	0	0	0	0
Sujawal	0	0	0	0	0
Sukkur	26015	197237	172960	5442186400	97959355
Tando Allahyar	1003	216743	137156	4315613540	77681044
Tando Muhammad Khan	5564	15790	12632	397465880	7154386
Tharparkar	269	637	44	1403024	25254
Thatta	2499	68950	35858	1128271970	20308895
Umer Kot	6521	235220	175221	5513328765	99239918
West Karachi	0	0	0	0	0
Total	**474818**	**3901399**	**3442718**	**108325139805**	**1949852516**

Normal Year = 2021, Price of 1 bale of cotton = 31,465 (price data obtained from cotton policy analysis of 2022-2023, Agriculture Policy Institute, Ministry of National Food Security and Research)

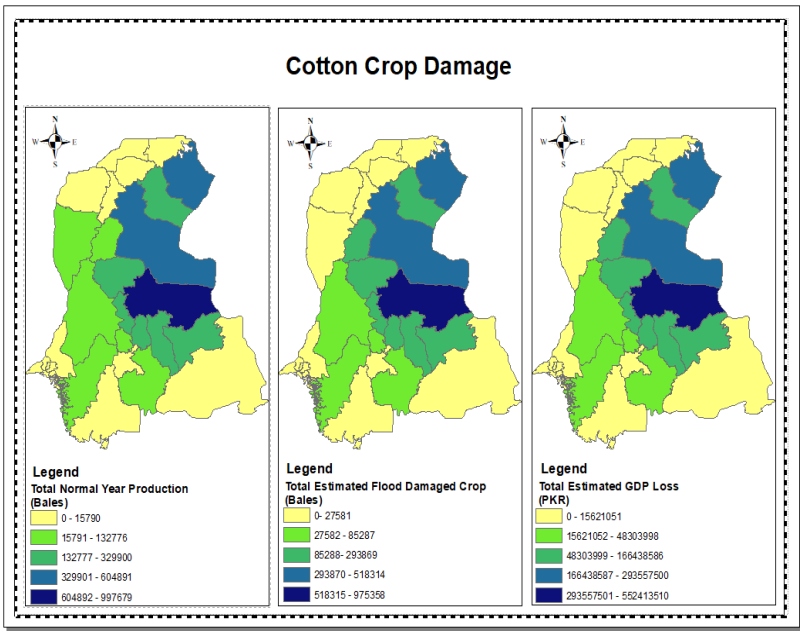


Figure 9. 1: Cotton Crop Damages

9.2.1.2. Rice Crop Damages

Table 9.2 illustrates the extent of the rice crop damages at district level. The table displays the total area of rice cultivation at the district level. The total area of rice cultivation in Sindh is 7,08,983 hectares. The data indicates that Shikarpur has the most cultivated area for rice, with Larkana and Jacobabad following closely behind with areas of 1,11,811, 1,06,000, and 1,03,721 hectares, respectively. The table presents the aggregate annual production of cotton in 2021 year. The year 2021 is the standard reference for this research. The data indicates that Badin had the largest rice output with a total of 3,87,067 m. tons. Larkana followed with 3,58,252 m. tons, while Jacobabad produced 3,54,570 m. tons of rice in 2021. The annual rice yield is overlaid on the model's projected inundation depths to determine the overall extent of crop damage caused by flooding. The overall projected damage to the rice crop caused by the flood was 18,604,12 m. tons, with Larkana experiencing the highest crop loss of 3,26,601 m. tons. The damaged rice crop had an estimated economic value of PKR 66,92,83,21,700, resulting in a GDP loss of PKR 80,31,39,860. The flood caused a significant loss to Pakistan's economy in terms of rice crop damages, with 77% of the standard year's production being affected. Qamer et al., 2022 have revealed similar impacts to the rice crop. This study offers an additional analysis at the district level of the total estimated losses in agricultural damage and GDP resulting from floods, measured in PKR. Figure 9.2 illustrates the estimated extent of flood damage to the rice crop. The graphic representation of the data is presented in table 9.2.

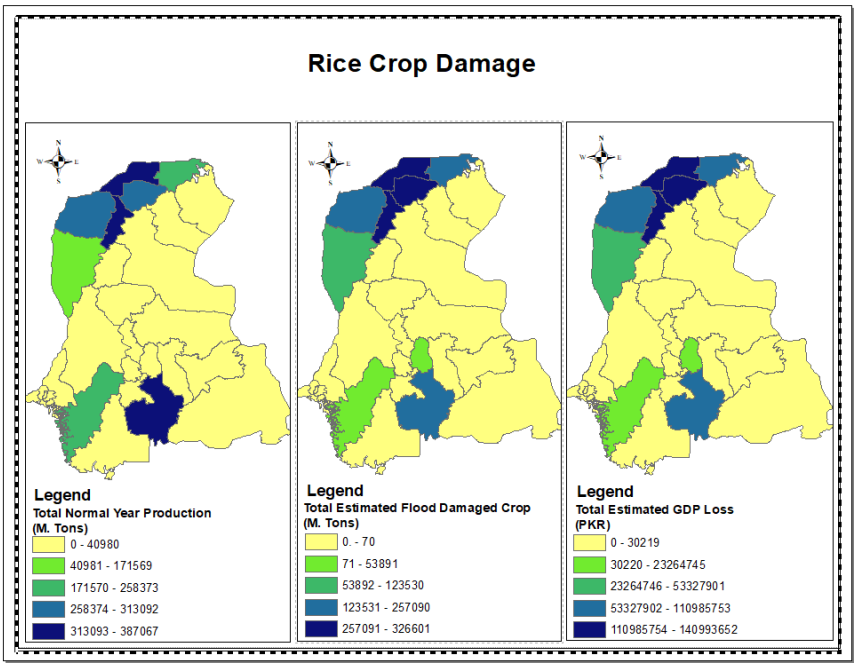


Figure 9. 2: Rice Crop Damages

Table 9. 2: Rice Crop Damages

Districts	Total Crop Cultivated Area (Hectors)	Total Normal Year Production (M. Tons)	Total Estimated Flood Damaged Crop (M. Tons)	Total Estimated Flood Damaged Crop (PKR)	Total Estimated GDP Loss (PKR)
Badin	100875	387067	256120	9213917000	110567004
Central Karachi			0	0	0
Dadu	46629	171569	123530	4443991750	53327901
East Karachi	0	0	0	0	0
Ghotki	0	0	0	0	0
Hyderabad	679	2349	70	2518250	30219
Jacobabad	103721	354570	319113	11480090175	137761082
Jamshoro		0	0	0	0
Kambar Shahdad Kot	90072	295506	257090	9248812750	110985753
Kashmore	76598	258373	208356	7495607100	89947285
Khairpur	0	0	0	0	0
Korangi Karachi	0	0	0	0	0
Larkana	106000	358252	326601	11749470975	140993652
Malir Karachi	0	0	0	0	0
Matiari	0	0	0	0	0
Mirpur Khas	0	0	0	0	0
Naushahro Feroze	0	0	0	0	0
Sanghar	0	0	0	0	0
Shaheed Benazir Abad	0	0	0	0	0
Shikarpur	111811	313092	283092	10184234700	122210816
South Karachi	0	0	0	0	0
Sujawal	0	0	0	0	0
Sukkur	0	0	0	0	0
Tando Allahyar	12225	40980	32549	1170950275	14051403
Tando Muhammad Khan	0	0	0	0	0
Tharparkar	0	0	0	0	0
Thatta	60373	234310	53891	1938728725	23264745
Umer Kot	0	0	0	0	0
West Karachi	0	0	0	0	0
Total	**708983**	**2416068**	**1860412**	66928321700	803139860

Normal Year = 2021, Price per m. ton of rice = 35,975 (price data obtained from Rice Paddy Policy Analysis for 2022-2023, Agriculture Policy Institute, Ministry of National Food Security and Research)

9.2.1.3. Sugarcane Crop Damages

Table 9.3 displays the magnitude of the sugarcane crop losses at the district level. The table presents the aggregate extent of sugarcane farming at the district level. The sugarcane farming area in Sindh measures 2,79,694 hectares. According to the data, Ghotki has the most cultivated area for sugarcane, with Thatta and Shaheed Benazirabad closely trailing behind with sizes of 5,78,86, 3,56,15, and 3,42,50 hectares, respectively. The year 2021 serves as the benchmark for this research. According to the data, Ghotki had the highest sugarcane production, totaling 40,73,839 metric tons. In 2021, Shaheed Benazirabad produced 24,54,420 metric tons of sugarcane, while Khairpur produced 15,98,823 metric tons. The model's estimated inundation depths are used to overlay the annual sugarcane yield in order to assess the total level of crop damage resulting from floods. The flood resulted in an estimated total damage of 1,33,66,159 metric tons to the sugarcane crop. Among the affected areas, Ghotki suffered the most significant loss, with a crop damage of 32,25,526 metric tons. The sugarcane crop, which was harmed, had an approximate economic worth of PKR 46568072137, leading to a GDP decline of PKR 325976505. Pakistan's economy suffered a substantial setback due to the flood, resulting in major damage to the sugarcane crop. Approximately 61% of the usual annual production was adversely affected. Qamer et al., 2022 have discovered comparable effects on the sugarcane crop. Figure 9.3 depicts the estimated magnitude of flood-induced harm to the sugarcane harvest. The data is visually depicted in table 9.3.

Table 9. 3: Sugarcane Crop Damages

Districts	Total Crop Cultivated Area (Hectors)	Total Normal Year Production (M. Tons)	Total Estimated Flood Damaged Crop (M. Tons)	Total Estimated Flood Damaged Crop (PKR)	Total Estimated GDP Loss (PKR)
Badin	13095	556576	236972	982367426	6876572
Central Karachi	0	0	0	0	0
Dadu	6095	277129	133021	551438556	3860070
East Karachi			0	0	0
Ghotki	57886	4073839	3225526	13371418033	93599926
Hyderabad	5042	299023	247069	1024224540	7169572
Jacobabad	0	0	0	0	0
Jamshoro	435	18489	15715	65146533	456026
Kambar Shahdad Kot	45	2402	2089	8659950	60620
Kashmore	159	10986	1047	4340339	30382
Khairpur	22782	1598823	271799	1126742755	7887199
Korangi Karachi	0	0	0	0	0
Larkana	689	51146	40916	169617278	1187321
Malir Karachi			0	0	0
Matiari	12454	861702	430851	1786092821	12502650
Mirpur Khas	16525	898368	434755	1802276853	12615938
Naushahro Feroze	22259	1463112	1064547	4413079589	30891557
Sanghar	12951	819282	263856	1093815048	7656705
Shaheed Benazir Abad	34250	2454420	1236188	5124617354	35872321
Shikarpur	70	51200	19498	80828959	565803

South Karachi	0	0	0	0	0
Sujawal	0	0	0	0	0
Sukkur	7385	498779	449656	1864048948	13048343
Tando Allahyar	12480	851165	425582	1764250181	12349751
Tando Muhammad Khan	17789	1142918	1043344	4325182552	30276278
Tharparkar	322	15914	1113	4613942	32298
Thatta	35615	2464230	1667729	6913570570	48394994
Umer Kot	1366	62110	22130	91739915	642179
West Karachi	0	0	0	0	0
Total	**279694**	**18471613**	**11233403**	**46568072137**	**325976505**

Normal Year = 2021, Price per m. ton of sugarcane = 4,145.5 (price data obtained from Sugarcane Policy Analysis for 2022-2023, Agriculture Policy Institute, Ministry of National Food Security and Research)

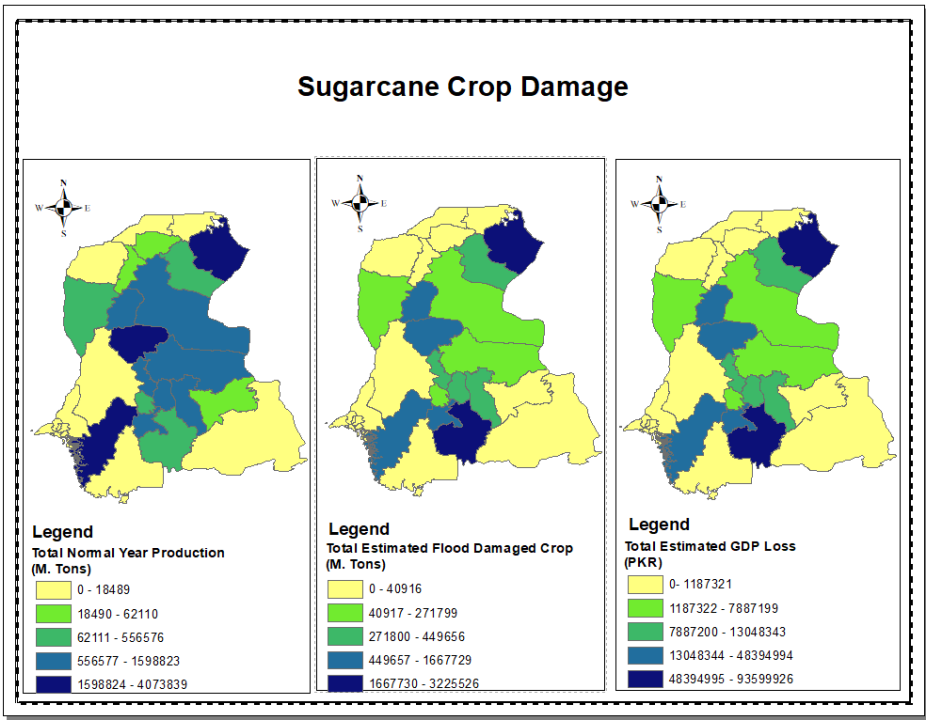


Figure 9. 3: Sugarcane Crop Damages

9.2.2. Industrial Damages

Table 9.4 illustrates the industrial damages due to flood 2022. For the assessment of industrial damages, the total industrial units and the total covered industrial area for each district was obtained from the Sindh Bureau of Statistics, Sindh Planning and Development Board and the land use land cover map, respectively. For this research only 65 large manufacturing units are considered. The statistics thus obtained were overlaid on the estimated model's flood depth and the industrial units' damages were thus obtained. A similar method has been used by Yi et al. (2010), Romali et al. (2015), Huizinga et al. (2017) and Prütz and Månsson (2021) in their studies to estimate flood related monetary industrial damages. Table 9.4 shows that Sindh has total number of 6299 industrial units among which West Karachi possesses the largest number of 1383 units. Among the damaged industrial units, Jamshoro had the highest number of 106 units followed by Kambar Shahdad Kot of 71 units. To assess the

economic value of industrial damages, district-level value of total fixed assets and total loss to GDP at factor cost have been used which were overlaid on the model's estimated inundation depths to calculate the economic losses. In this regard, the economic damages are worth of 362817807 and 240403941, respectively.

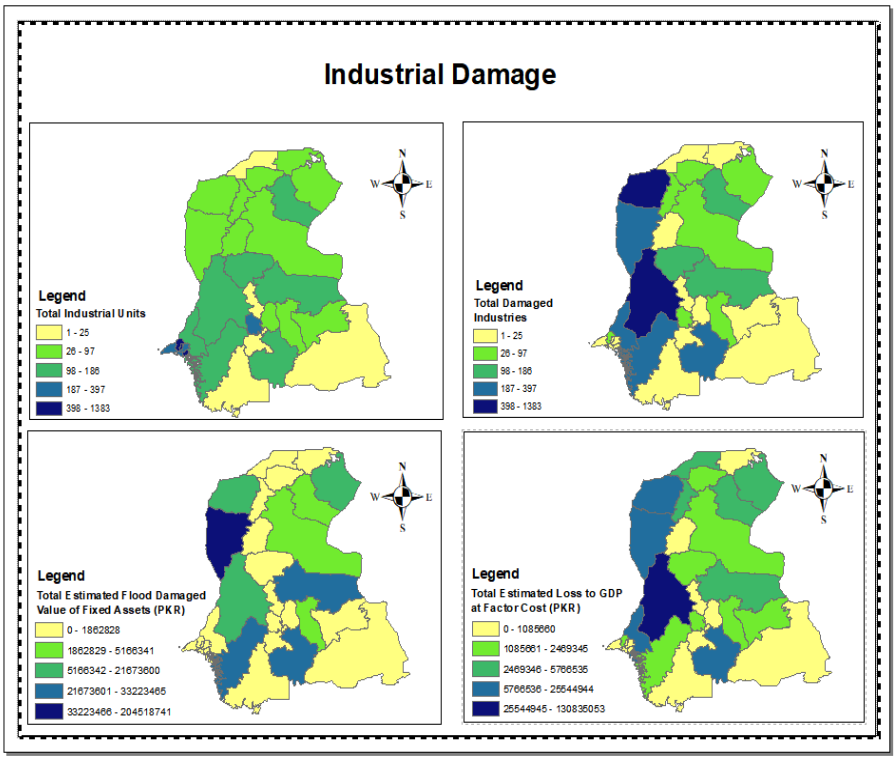


Figure 9. 4: Industrial Damages

Table 9. 4: Industrial Damages

Districts	Total Industrial Units	Total Estimated Affected Industrial Units	Total Estimated Flood Damaged Value of Fixed Assets (PKR)	Total Estimated Loss to GDP at Factor Cost (PKR)
Badin	131	40	29003152	11825180
Central Karachi	1054	2	13945	331138
Dadu	79	38	204518741	25544944
East Karachi	252	2	1095802	662276
Ghotki	60	10	12789303	4423060
Hyderabad	397	14	1107575	2193790
Jacobabad	10	1	39117	5766535
Jamshoro	125	106	21673600	130835053
Kambar Shahdad Kot	82	71	11236944	12289956
Kashmore	42	4	138692	1085660
Khairpur	44	8	2460663	1617846
Korangi Karachi	1123	4	559499	567665
Larkana	97	8	495845	4352102
Malir Karachi	181	43	319920	14014240
Matiari	11	1	7202	188039
Mirpur Khas	53	8	5166341	2086170
Naushahro Feroze	51	5	118874	1021798
Sanghar	138	28	33223465	5548929
Shaheed Benazir Abad	115	16	1862828	2106275
Shikarpur	68	7	131944	1622577

South Karachi	322	5	195717	993414
Sujawal	0	0	0	0
Sukkur	157	17	3280295	3635424
Tando Allahyar	49	3	72069	1002876
Tando Muhammad Khan	21	2	183751	747426
Tharparkar	25	2	103506	175030
Thatta	186	43	31259048	2469345
Umer Kot	43	4	138214	1475930
West Karachi	1383	10	1621755	1821260
Total	**6299**	**499**	**362817807**	**240403941**

9.2.3. Housing Damages

Flood damage has the potential to impact the market valuation of a property. Estimating the construction cost aids in evaluating the effect on property value and providing guidance for decisions related to property repairs, renovations, or sale. To evaluate the extent of residential damage, we gathered data on the total number of residential units and the total area of residential land for each district from the Sindh Bureau of Statistics, Research and Training Wing Planning and Development Board of Sindh, and the land use land cover map. Moreover, the affected residential buildings property cost is estimated through construction cost of the ground floor per unit m^2. Construction cost depicts monetary valuation of the residential building. The method to assess the construction cost and the building materials used for construction of ground floor has been adopted from the study conducted by Aslam et al. (2023). The building materials' costs per m^2 of the area has been obtained from Pakistan Bureau of Statistics. The building materials used for our research are cement concrete (price per unit m^2 = 156.45), bricks and blocks (price per unit m^2 = 720), and sand (price per unit m^2 = 3). The building material prices per unit m^2 were multiplied with the total area of the residential buildings by using the Raster Calculator Tool in Arc GIS and the resultant residential buildings' monetary valuation raster has been utilized for further analysis. The gathered statistics were overlaid on the estimated model's flood depth, resulting in the determination of damages to the residential units. A comparable methodology has been used by Luino et al. (2009), Yi et al. (2010), Middelmann-Fernandes (2010), Huizinga et al. (2017) and Prütz and Månsson (2021), for the estimation of damaged buildings and their cost.

Table 9.5 depicts the residential damages due to flood 2022. The table shows that total number of residential units in Sindh are 5641516. The research has only considered pakka houses. The total estimated flood affected residential units and the total estimated affected property cost (in PKR) were 672202 and 4929725, respectively. The United Nations Office for the Coordination of Humanitarian Affairs, report on released on 05 September, 2022, provides

analogous figures for flood affected houses in Sindh. The figure 9.5 depicts the higher number of houses were damaged in Dadu, Kambar Shahdad Kot, Jamshoro and Malir Karachi.

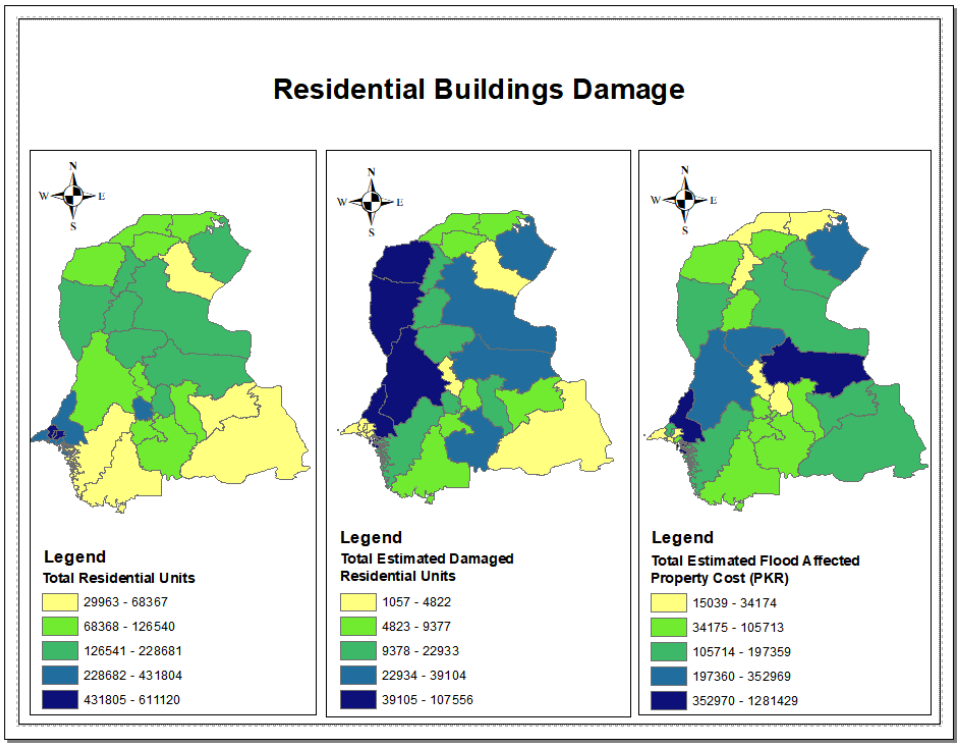


Figure 9. 5: Residential Units Damages

Table 9. 5: Residential Units Damages

Districts	Total Residential Units (Pakka House)	Total Estimated Damaged Residential Units	Total Estimated Flood Affected Property Cost (PKR)
Badin	101243	30677	51248
Central Karachi	528465	1057	147089
Dadu	155899	74832	194332
East Karachi	483758	3870	34174
Ghotki	185646	31560	352969
Hyderabad	388390	13594	105713
Jacobabad	78548	7226	15074
Jamshoro	126388	107556	239586
Kambar Shahdad Kot	112696	97595	96952
Kashmore	84237	8592	16664
Khairpur	228681	39104	171136
Korangi Karachi	431804	1727	58971
Larkana	157592	12607	16490
Malir Karachi	302274	71639	1281429
Matiari	90988	4822	19426
Mirpur Khas	126540	18601	70449
Naushahro Feroze	175386	16837	65095
Sanghar	190406	38843	738700
Shaheed Benazir Abad	167395	22933	247158
Shikarpur	89642	8785	48944
South Karachi	314392	4401	16090
Sujawal	29963	5903	54922
Sukkur	29963	3176	197359
Tando Allahyar	176927	9377	15039

Tando Muhammad Khan	87494	6912	99597
Tharparkar	54618	4042	131856
Thatta	62694	14545	146485
Umer Kot	68367	7110	160673
West Karachi	611120	4278	136105
Total	**5641516**	**672202**	**4929725**

9.2.4. Infrastructural Damages

For estimation of infrastructural damages, this research has only asphalt and paved roads only. Table 9.6 illustrates the damages related to infrastructure. The data on district-level roads in kilometers has been obtained from Sindh Bureau of Statistics and roads shapefile was obtained from Humanitarian Data Exchange. The construction cost of the roads has been utilized as the monetary valuation of the affected roads. In this regard, the method illustrated in the above-mentioned section has been employed and the building materials used for roads are cement concrete and sand. The figure 9.6 depict that the highest infrastructural damage has been caused in the districts of Khairpur, Dadu, Kambar Shahdad Kot, and Jamshoro.

Table 9. 6: Infrastructural Damages

Districts	Total Length of Roads (Kms)	Total Estimated Damaged Roads (kms)	Total Estimated Flood Affected Roads Cost (PKR)
Badin	1453	436	12923535
Central Karachi	357	1	21169
Dadu	1123	539	15981422
East Karachi	357	4	105843
Ghotki	1204	205	6068339
Hyderabad	572	17	508759
Jacobabad	706	64	1883830
Jamshoro	672	571	16934900
Kambar Shahdad Kot	814	708	20996074
Kashmore	612	61	1814454
Khairpur	2796	475	14092256
Korangi Karachi	357	1	42337
Larkana	1174	94	2784534
Malir Karachi	357	86	2540235
Matiari	395	20	585547
Mirpur Khas	1041	156	4629525
Naushahro Feroze	890	89	2638666
Sanghar	1422	284	8431872
Shaheed Benazir Abad	2003	280	8313874
Shikarpur	745	75	2208771
South Karachi	357	4	105843
Sujawal	0	0	0
Sukkur	1047	115	3414553
Tando Allahyar	606	30	898332
Tando Muhammad Khan	427	34	1012773

Tharparkar	1248	87	2590044
Thatta	1535	353	10467203
Umer Kot	1221	122	3620013
West Karachi	357	4	105843
Total	**25848**	**4915**	**145720545**

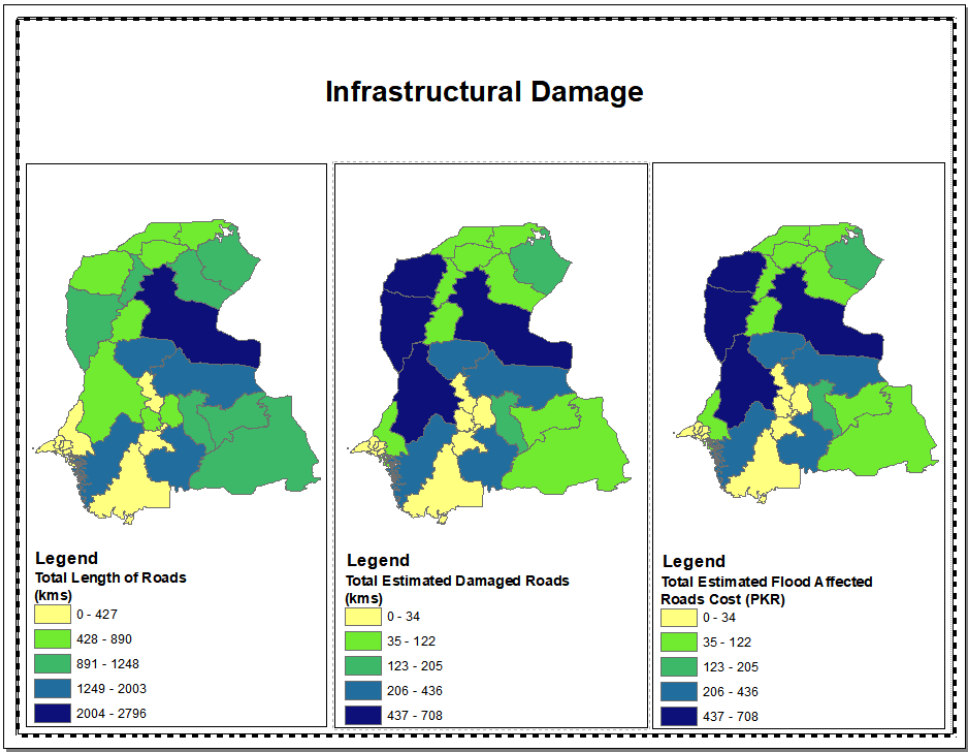


Figure 9. 6: Infrastructural Damages

9.2.5. Educational Units Damages

Table 9.7 illustrates the damage to educational units at district-level. The data on district-level educational institutes has been obtained from Sindh Bureau of Statistics. The research has only considered primary, middle, secondary and higher secondary schools. To estimate the number of affected educational units, the functional school's raster was overlaid on the flood inundation depth raster. The figure 9.7 shows that the largest number of educational institutional damage has been caused in the districts of Kambar Shahdad Kot, Dadu and Malir Karachi.

Table 9. 7: Educational Units Damages

Districts	Total Number of Schools	Total Number of Functional Schools	Total Estimated Flood Damaged Schools
Badin	3127	2517	390
Central Karachi	3036	2629	5
Dadu	2076	1538	617
East Karachi	3036	2629	21
Ghotki	2231	1725	203
Hyderabad	907	818	27

Jacobabad	1555	1165	94
Jamshoro	842	630	282
Kambar Shahdad Kot	1717	1291	859
Kashmore	1681	1225	101
Khairpur	3617	3031	361
Korangi Karachi	3036	2629	10
Larkana	1287	1122	91
Malir Karachi	3036	2629	634
Matiari	971	780	39
Mirpur Khas	2311	1694	142
Naushahro Feroze	2552	2179	142
Sanghar	3350	2476	368
Shaheed Benazir Abad	2743	2194	250
Shikarpur	1374	970	77
South Karachi	3036	2629	37
Sujawal	1829	964	85
Sukkur	1875	1543	179
Tando Allahyar	835	636	33
Tando Muhammad Khan	1166	788	53
Tharparkar	4269	2779	128
Thatta	1607	910	148
Umer Kot	2745	1526	132
West Karachi	3036	2629	19
Total	**64883**	**50277**	**5526**

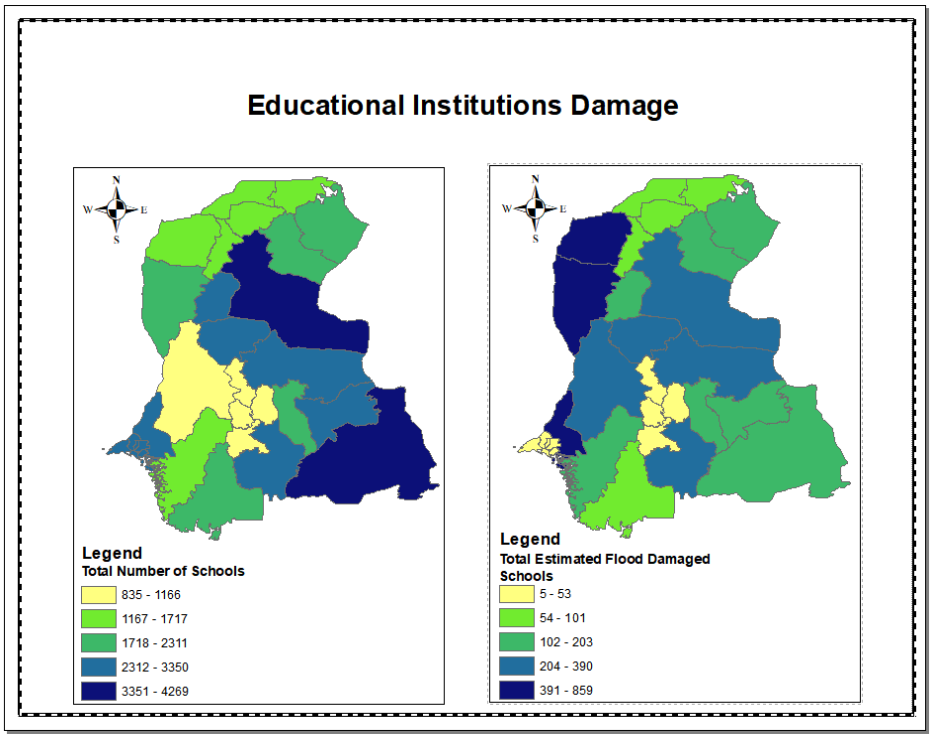


Figure 9. 7: Educational Institutional Damages

9.2.6. Medical Units Damages

Table 9.8 depicts the extent of harm inflicted upon medical facilities at the district level. The information regarding medical institutions at the district level has been acquired from the Sindh Bureau of Statistics. The research has exclusively focused on civil specialized hospitals, dispensaries, rural health units, T.B. clinics, basic health units and maternity care health centers. In order to determine the number of medical units that were impacted, the raster representing the medical units was overlaid on the raster depicting the depth of the flood inundation. According to Figure 9.8, the districts of Kambar Shahdad Kot, Jamshoro and Dadu, have experienced the most significant amount of damage to health institutions.

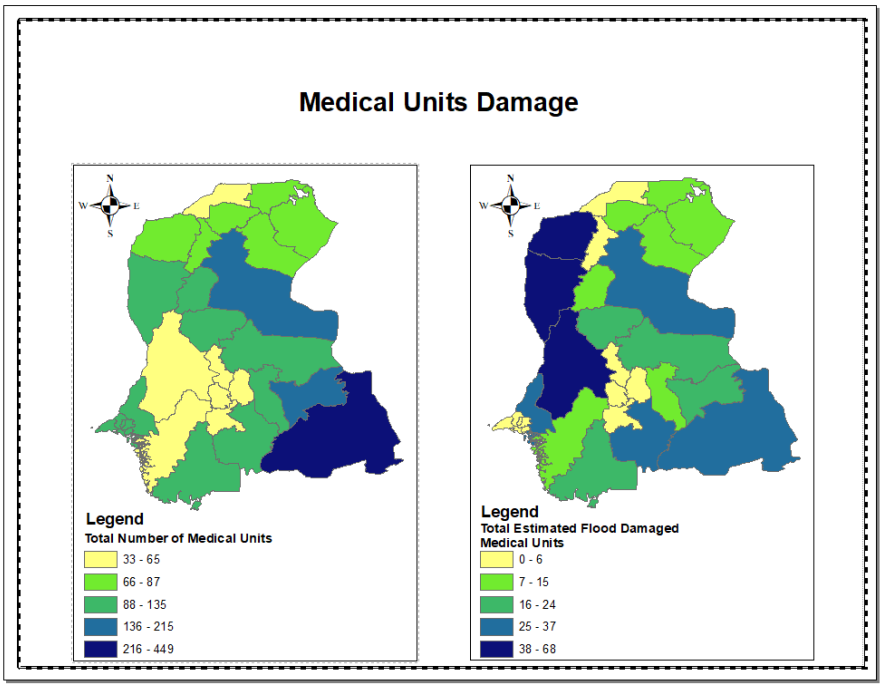


Figure 9. 8: Medical Units Damages

Table 9. 8: Medical Units Damages

Districts	Total Number of Medical Units	Total Estimated Flood Damaged Medical Units
Badin	113	34
Central Karachi	135	0
Dadu	103	49
East Karachi	135	1
Ghotki	70	12
Hyderabad	65	2
Jacobabad	56	5
Jamshoro	59	50
Kambar Shahdad Kot	79	68
Kashmore	80	8
Khairpur	215	37
Korangi Karachi	135	1
Larkana	76	6

Malir Karachi	135	32
Matiari	51	3
Mirpur Khas	105	15
Naushahro Feroze	126	12
Sanghar	106	22
Shaheed Benazir Abad	131	18
Shikarpur	78	8
South Karachi	135	2
Sujawal	122	24
Sukkur	87	9
Tando Allahyar	62	3
Tando Muhammad Khan	33	3
Tharparkar	449	33
Thatta	57	13
Umer Kot	169	18
West Karachi	135	1
Total	**3302**	**489**

9.3. Quantification of Flood Damages for 2032 Flood

The flood extents and depths, acquired from modelling the bag-boost ensemble model (NBT-REPT-RF-LMT), have been used to quantify the economic damages caused by the flood. In order to achieve this objective, the simulated floods and the land use dataset were overlaid in a Geographic Information System (GIS) to identify the assets that were impacted by flooding and to quantify the depths of inundation associated with them. The subsequent sections present an evaluation of the damages caused by the 2032 flood.

9.3.1. District-level Agricultural Losses Estimation

For forecasting purpose, the research focuses on assessing agricultural damages by analyzing three key Kharif crops between the period of June to August as these months are the peak monsoon precipitation and riverine flooding months. Cotton, sugarcane, and rice are being investigated for the quantification and evaluation of damages in this particular setting.

9.3.1.1. Cotton Crop Damages

Table 9.9 illustrates the extent of cotton crop damages at the district level. The table displays the total area of cotton cultivation at the district level. The total area of cotton cultivation in Sindh is 5,41,980 hectares. The total cotton crop cultivated area and normal year production for each district has been obtained by simulating historic data and forecasting by utilizing CGREM model. The analysis considers 2031 as the baseline year. The annual cotton

yield is overlaid on the model's projected depths of flooding to determine the overall extent of crop damage caused by the flood. The overall projected damage to the cotton crop due to the flood was 30,52,382 bales, with Sanghar experiencing the highest crop loss of 9,87,458 bales. The cotton crop has an estimated economic worth of PKR 1,58,30,57,05,441. Figure 9.9 illustrates the estimated extent of flood damage to the cotton crop. The graphic representation of the data presented in table 9.9.

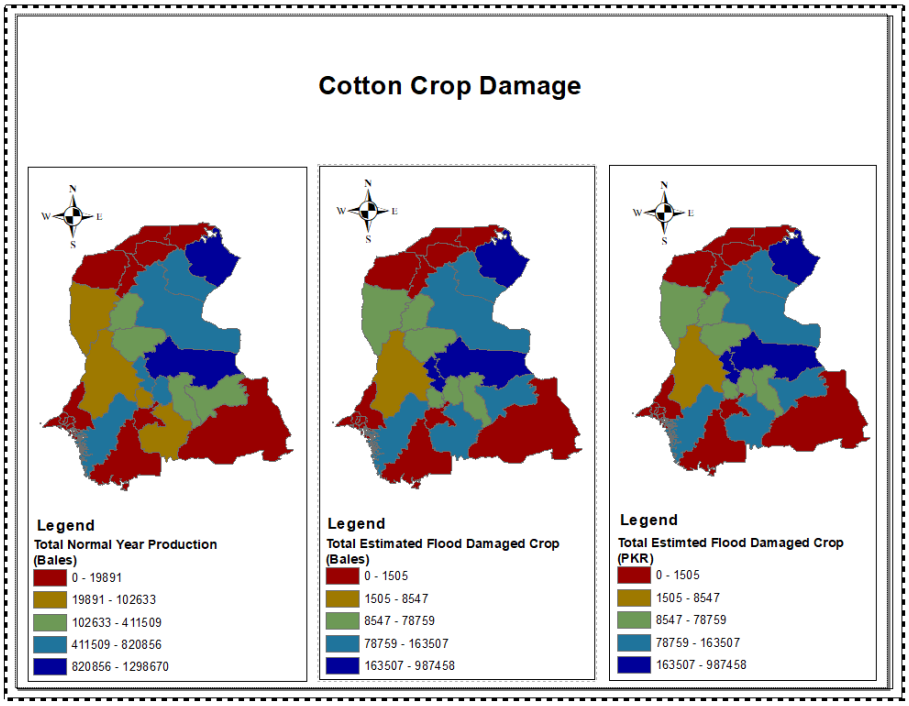


Figure 9. 9: *Cotton Crop Damages*

Table 9. 9: Cotton Crop Damages

Districts	Total Crop Cultivated Area (Hectors)	Total Normal Year Production (Bales)	Total Estimated Flood Damaged Crop (Bales)	Total Estimated Flood Damaged Crop (PKR)
Badin	8940	102633	98019	5083559397
Central Karachi	0	0	0	0
Dadu	23209	84249	78759	4084678017
East Karachi	0	0	0	0
Ghotki	115789	953998	615789	31936664907
Hyderabad	6702	61281	36258	1880448654
Jacobabad		0	0	0
Jamshoro	14307	45793	8547	443273061
Kambar Shahdad Kot	0	0	0	0
Kashmore	0	0	0	0
Khairpur	79879	820856	147464	7647942657
Korangi Karachi	0	0	0	0
Larkana	248	548	76	3955758
Malir Karachi	317	1592	1095	56796038
Matiari	47985	732469	458158	23761448354
Mirpur Khas	22597	341547	47166	2446173181

Naushahro Feroze	27364	312784	37042	1921103392
Sanghar	81647	1298670	987458	51212534254
Shaheed Benazir Abad	69268	411509	64588	3349709051
Shikarpur	87	205	43	2242931
South Karachi	0	0	0	0
Sujawal	0	0	0	0
Sukkur	27438	680723	151019	7832281228
Tando Allahyar	957	556941	35503	1841315286
Tando Muhammad Khan	5487	19891	1505	78074841
Tharparkar	294	827	122	6327708
Thatta	2978	558890	120264	6237228366
Umer Kot	6487	275229	163507	8479948358
West Karachi	0	0	0	0
Total	**541980**	**7260635**	**3052382**	**158305705441**

Normal Year = 2031, Price of 1 bale of cotton = 51,863 (Past prices data obtained from cotton policy analysis, Agriculture Policy Institute, Ministry of National Food Security and Research)

9.3.1.2. Rice Crop Damages

Table 9.10 illustrates the extent of the rice crop damages at district level. The table displays the total area of rice cultivation at the district level. The total area of rice cultivation and total normal year production are forecasted and have been utilized for performing predictions. In this regard, the data has been obtained from Sindh Bureau of Statistics and Ministry of National Food Security and Research. The data indicates that total crop cultivated area, total normal year production, flood damaged crop in m. tons and total monetary value of damaged crop will be 779698, 2707821, 2190115 and 116459389020, respectively. Figure 9.10 illustrates the estimated extent of flood damage to the rice crop. The graphic representation of the data is presented in table 9.10.

Table 9. 10: Rice Crop Damages

Districts	Total Crop Cultivated Area (Hectors)	Total Normal Year Production (M. Tons)	Total Estimated Flood Damaged Crop (M. Tons)	Total Estimated Flood Damaged Crop (PKR)
Badin	120345	627967	578982	30787367850
Central Karachi	0	0	0	0
Dadu	49000	201169	191323	10173600525
East Karachi	0	0	0	0
Ghotki	0	0	0	0
Hyderabad	680	2142	91	4862820
Jacobabad	103147	344973	248972	13239086100
Jamshoro	0	0	0	0
Kambar Shahdad Kot	95726	305566	234879	12489690825
Kashmore	96528	262377	179845	9563257875
Khairpur	0	0	0	0
Korangi Karachi	0	0	0	0

Larkana	123257	349652	301279	16020510825
Malir Karachi	0	0	0	0
Matiari	0	0	0	0
Mirpur Khas	0	0	0	0
Naushahro Feroze	0	0	0	0
Sanghar	0	0	0	0
Shaheed Benazir Abad	0	0	0	0
Shikarpur	115978	333078	201358	10707211650
South Karachi	0	0	0	0
Sujawal	0	0	0	0
Sukkur	0	0	0	0
Tando Allahyar	13458	42987	38798	2063083650
Tando Muhammad Khan	0	0	0	0
Tharparkar	0	0	0	0
Thatta	61579	237910	214588	11410716900
Umer Kot	0	0	0	0
West Karachi	0	0	0	0
Total	**779698**	**2707821**	**2190115**	**116459389020**

Normal Year = 2031, Price per m. ton of rice = 53,175 (Previous prices data obtained from Rice Paddy Policy Analysis, Agriculture Policy Institute, Ministry of National Food Security and Research)

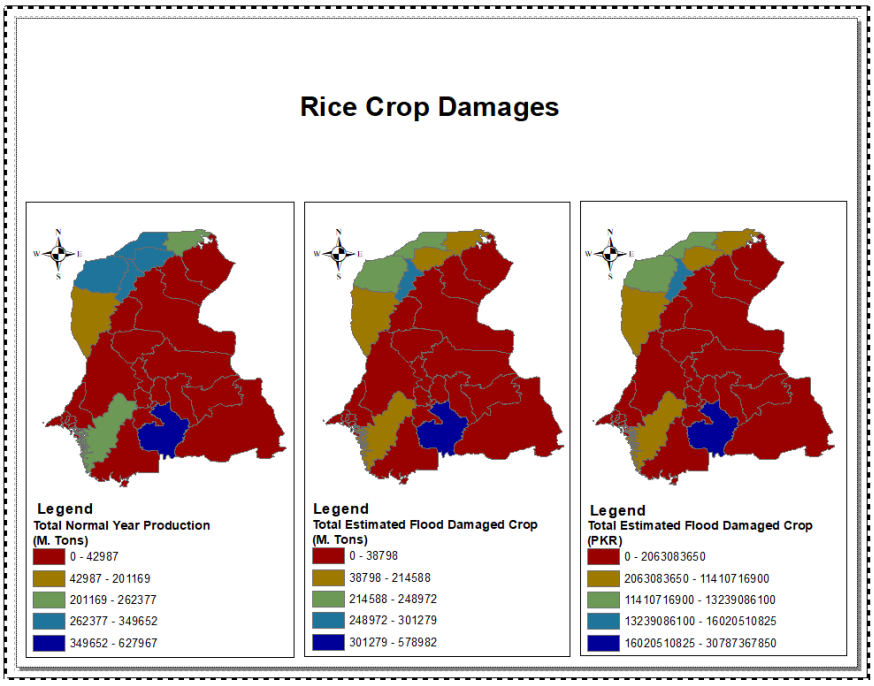


Figure 9. 10: Rice Crop Damages

9.3.1.3. Sugarcane Crop Damages

The table labelled 9.11 illustrates the extent of the sugarcane crop damage at the district level. The table displays the overall scope of sugarcane cultivation at the district level. CGREM model have been used to anticipate the total area of sugarcane cultivation and the total yield in a normal year. Predictions have been made using ten-year data from 2011 to 2021. The data

has been acquired from the Sindh Bureau of Statistics and the Ministry of National Food Security and Research. Figure 9.11 illustrates the estimated extent of damage caused by floods to the sugarcane harvest. The data is graphically represented in table 9.11.

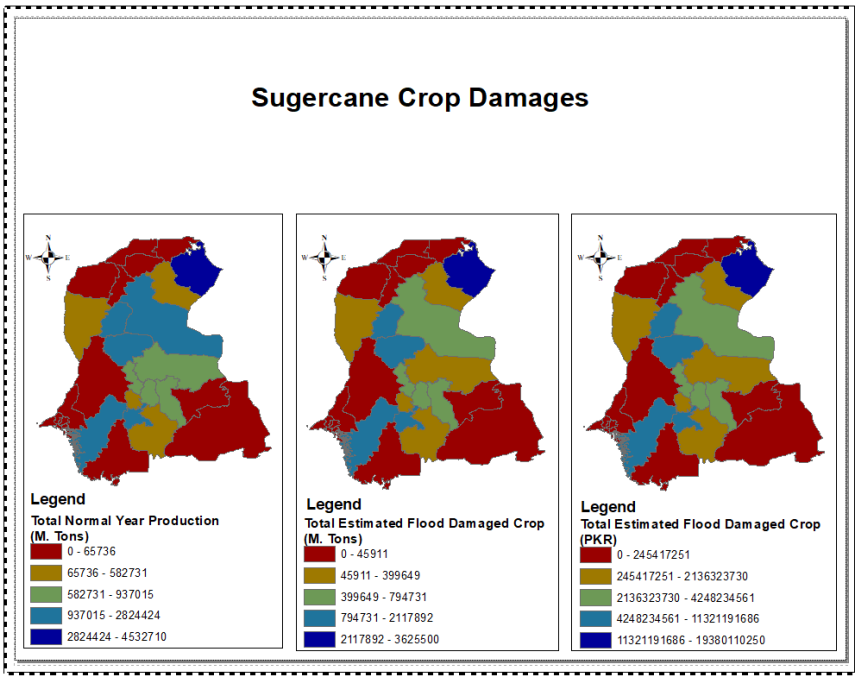


Figure 9. 11: Sugarcane Crop Damages

Table 9. 11: Sugarcane Crop Damages

Districts	Total Crop Cultivated Area (Hectors)	Total Normal Year Production (M. Tons)	Total Estimated Flood Damaged Crop (M. Tons)	Total Estimated Flood Damaged Crop (PKR)
Badin	14238	582731	292479	1563446495
Central Karachi		0	0	0
Dadu	5920	290739	173167	925664199
East Karachi			0	0
Ghotki	58956	4532710	3625500	19380110250
Hyderabad	7805	310914	237726	1270764333
Jacobabad		0	0	0
Jamshoro	6459	19507	12489	66759950
Kambar Shahdad Kot	52	2519	2147	11476789
Kashmore	162	12190	5978	31955399
Khairpur	22983	1812001	471700	2521472350
Korangi Karachi		0	0	0
Larkana	693	58949	45911	245417251
Malir Karachi			0	0
Matiari	14589	889707	703679	3761516095
Mirpur Khas	17896	937015	794731	4248234561
Naushahro Feroze	23147	1563113	1097842	5868514411
Sanghar	12993	829990	333857	1784632594
Shaheed Benazir Abad	35249	2824424	2117892	11321191686
Shikarpur	71	52227	27490	146947795

South Karachi		0	0	0
Sujawal		0	0	0
Sukkur	7412	518770	399649	2136323730
Tando Allahyar	13597	871137	494698	2644408159
Tando Muhammad Khan	19782	1542019	1143317	6111601024
Tharparkar	534	13710	10476	55999458
Thatta	36782	2564190	1447709	7738728460
Umer Kot	1498	65736	42683	228161977
West Karachi		0	0	0
Total	**300818**	**20294298**	**13481120**	**72063326960**

Normal Year = 2031, Price per m. ton of sugarcane = 5345 (Previous prices data obtained from Sugarcane Policy Analysis, Agriculture Policy Institute, Ministry of National Food Security and Research)

Table 9.12 provides details of total estimated GDP loss due to 2022 and 2032 floods.

Table 9. 12: Total Flood Damages

Category	**Total Estimated GDP Loss (PKR)**		**Total Estimated Flood Damaged Units**
	2022	**2032**	2022
Agriculture			
Cotton	1949852516	158305705441	
Rice	803139860	116459389020	
Sugarcane	325976505	72063326960	
Industry	240403941		
Houses	4929725		
Infrastructure	145720545		
Education Institutes			5526
Medical Units			489
Total Damage	3470023092	346828421421	

Chapter 10
Conclusion

10.1. Key Findings of the Research

This research is an endeavor to analyze the causation and impacts of floods in the lower Indus basin. Thus, in the research, a broader analysis has been conducted by considering environmental, topographic and human induced factors which may have potential contribution in causing floods in the study area. Following are the key findings of the study;

a) The Particle Swarm Optimization (PSO) algorithm outperformed the other models by selecting the most relevant variables with the fewest iterations. Among the hybrid models of Support Vector Machines (SVM), the findings indicated that the PSO-SVM model with a Radial Basis Function (RBF) kernel achieved the maximum accuracy of 99.3% for geographic datasets that had issues with autocorrelation, heteroscedasticity, and autocorrelation. Therefore, it is recommended to utilize PSO-SVM with RBF kernel for feature selection in spatial datasets. Likewise, PSO-KNN demonstrated a remarkable accuracy of 99.8% when using a k-value of 50. It is determined that in the case of KNN, lower values of k resulted in lower accuracies, followed by reaching a peak value and subsequently dropping. Hence, it is advisable to select a k-value for PSO-KNN that is neither the minimum nor the maximum, but rather a value in the middle. When evaluating the performance of PSO-SVM and PSO-KNN, it was found that PSO-KNN outperformed PSO-SVM. This statement claims that PSO-KNN can be used to choose features from datasets that have significant levels of multicollinearity, heteroscedasticity, and autocorrelation. These datasets specifically include topographic, geo-environmental, and human-induced data. By employing this approach, it is possible to choose a significant subset of variables that are crucial for producing exact flood susceptibility maps, conducting hazard and risk assessments, and calculating economic damages. These refined predictions are particularly valuable for the lower Indus basin.
b) The hybrid bag-boost ensemble model had superior performance across all datasets, while the LMT model exhibited the lowest accuracies and areas under the curve for PSO and ACO variables. Similarly, the RF model had the lowest accuracy and area under the curve for GA variables. When the dataset is partitioned into different subsets, so changing the range of data used for training and validation, the split that allocates 70% of the data for training and 30% for validation yields the optimal outcomes. Therefore, it is advisable to use a 70/30 data split, with 70% of the data used for training and 30% for validation.

c) The primary factors that contributed to the 2022 flood were rainfall, land use and land cover (LULC), temperature, and slope. In contrast, for the projected flood in 2032, the estimated main contributors will be rainfall, LULC, temperature, and distance from the river. During the flood of 2022, the districts that saw the most severe flooding were Dadu, Jamshoro, Kashmore, Khairpur, Larkana, Mitiari, Shaheed Benazirabad, Shikarpur, and Sukkur. The flooded area in these districts exceeded 1000 km^2. Shikarpur was the most severely affected, with a total area of 1900 km^2 devastated. The flood-affected area in the Sindh province covered 22,100 square kilometers. It is anticipated that during the 2032 flood, the districts of Dadu, Ghotki, Jamshoro, Kashmore, Khairpur, Larkana, Mitiari, Shaheed Benazirabad, Shikarpur, Tharparkar, and Sukkur will experience flood inundation areas exceeding 1000 km^2. Kashmore is expected to see the most significant repercussions, with a region covering 2300 square kilometers being affected. The inundated region in the Sindh province is projected to encompass a combined size of 22,500 square kilometers.

d) Alterations in land use and land cover (LULC), particularly in regions situated at a substantial distance from the river, have a substantial impact on riverine floods. Therefore, it may be inferred that alterations in land use and land cover (LULC) have a significant influence on the rise in river floods. Policymakers are anticipated to employ the framework to efficiently reduce flood risk through suitable planning. The research proposes the incorporation of a machine learning ensemble model (NBT-REPT-RF-LMT) with GWR (Geographically Weighted Regression) to conduct relational statistical analysis, notably focusing on spatial nonlinearity. This methodology can help other researchers determine the association between complex and nonlinear spatial variables, including not only changes in land use and land cover but also the possibility for flooding.

e) The lower Indus basin is prone to riverine floods, the flood of 2022 has revealed huge economic losses in this regard. Among the losses, the cotton crop is the most adversely affected crop among the chosen crops for this research. The asset damage figures (residential, infrastructure, medical and education) reveal that Kambar Shahdat Kot, Dadu, Jamshoro, Khairpur, Sukkur and Ghotki are the districts which are the most adversely hit by the flood 2022. The LULC transition in these districts from natural vegetation to barren lands can be a causation factor to this damage.

10.2. Suggestions and Policy Implications

Based on the results, the study suggests the following recommendations and implications;

a) PSO-KNN can be used to choose the most significant flood contributing elements without encountering econometric issues such as multicollinearity, heteroscedasticity, and autocorrelation. The proposed methodology will help policymakers obtain the most pertinent subset of data for mapping and analyzing flood susceptibility, hazard, and risk. The implementation of flood forecasting mechanisms can significantly aid in the planning and mitigation of economic damages and human casualties caused by future floods. The following factors can be utilized for flood predictions and flood susceptibility and hazard mapping in the lower Indus basin: population density, rainfall, stream power index, temperature, Land Use and Land Cover, positive topographic openness, negative topographic openness, distance from the river, slope, lithology, stream density, sediment transport index, topographic wetness index, and topographic Ruggedness Index.
b) The study has proposed a model-based flood susceptibility mapping which produced very close output to the flood images retrieved from Sentinel 1 for the years 2010 and 2022. Thus, the model and the dataset proposed by the research can be utilized for future hazard mapping for any year in the study area, as this area is prone to future floods owing to the predicted climatic and land cover conditions.
c) Khairpur, Sukkur, Ghotki, Kambar Shahdat Kot, Dadu and Jamshoro districts have major land cover transitions from grasslands and shrublands to barren lands which intensified the riverine flood of 2022 in these areas. The correlation between flood and LULC also depicts high correlation between the two in these regions. Thus, these regions may be considered for natural vegetation, grasslands and forests, to minimize the adverse impacts of future floods.
d) Kambar Shahdat Kot, Dadu, Jamshoro, Khairpur, Sukkur and Ghotki districts are the districts that have been largely affected by the flood 2022 in socio-economic terms. There are environmental, climatic, topographic and anthropogenic factors involved in causing floods. The projected precipitation, temperature, population and LULC maps show that these conditions will be intensified in future. Thus, it is suggested that these districts must be considered while formulating environmental and socio-economic policies. In flood analysis, the cost of false negative (missed flood detection) is very high than false positive (predicting a flood when there is no flood) because they can cause severe property damage, loss of life and infrastructure destructions, whereas false positive may only result in temporary disruption and

inconvenience. For this purpose, early warning systems (EWS) and the agent-based models (ABM) can be utilized to mitigate flood destruction. An ABM for flood risk assessment is a simulation approach that models the behaviors and interactions of individual entities—such as residents, emergency responders, and government officials—within a flood-prone environment. These agents operate according to predefined rules and make decisions based on environmental conditions, such as rising water levels or flood warnings. The environment typically includes spatial data like flood maps, land use, and elevation models. ABMs allow for the simulation of complex scenarios, including evacuation behavior, emergency response, and the effectiveness of policy interventions. By capturing individual-level decision-making and interactions, ABMs help researchers and policymakers assess how human behavior influences flood outcomes and evaluate the effectiveness of various risk mitigation strategies. This approach is increasingly applied in urban planning and disaster management studies. Moreover, nature-based solutions (NBS) provide a solution for mitigating flood risks by combining ecosystem-based approaches with responses to societal challenges. NBS for flood prevention and recovery includes strategies like flood plain management, sustainable agricultural practices and water retention initiatives. Hence, these methods can be investigated in future and may be implemented in the flood endangered districts.

10.3. Limitations and Future Research

The study has the following limitations and future research suggestions;

a) The research has only considered flooding in the lower Indus basin, whereas, many regions in the upper Indus basin and the many districts of Baluchistan are prone to flooding. The future research may analyze floods in these areas.
b) The use of geo-spatial data layer maps with high resolutions. However, field-based surveys can offer more accurate and exact datasets for future flood predictions and reducing economic damages.
c) This study has exclusively employed biological metaphor-based metaheuristic algorithms, specifically GA (evolutionary-based), PSO, and ACO (swarm intelligence-based). However, it is worth noting that future studies could explore other variants of metaphor-based algorithms, such as those based on chemistry, mathematics, physics, and so on, as well as non-metaphor-based algorithms.

d) We have exclusively employed only two machine learning algorithms, namely SVM and KNN, for the hybridization of metaheuristics. However, it is worth considering the evaluation of alternative machine learning models in future studies.

e) The dataset used in this research consists of 21 variables that primarily focus on the topographic, environmental, and anthropogenic aspects of the lower Indus basin. These variables may differ in other study regions. Therefore, it is recommended that when doing flood-related feature selection and hydraulic analysis for various study sites, careful consideration should be given to the terrain, environment, and human influence of each individual place.

f) The study has utilized hybrid bag-boost decision trees ensemble model for flood susceptibility mapping, other hybridization techniques such as catBoost, eXtreme Gradient Boost, Light Gradient Boost, adaBoost, etc. can be utilized for future analysis. Furthermore, machine learning models rely on manually engineered features. These models can be applied for raster data files (A raster is a grid-based data format used in geographic information systems (GIS) and remote sensing to represent spatial data. It consists of a matrix of cells or pixels, each with a specific location, value, and resolution) like slope, rainfall, land use, or distance from rivers, and they tend to be more interpretable and faster to train. Whereas, deep learning models consist of multiple layers that can automatically learn complex patterns from raw data like satellite imagery. Thus, it is suggested that Neural Network models of deep learning (DL) and their hybrid and ensemble models can be utilized and compared for future flood and climate change analysis.

g) The study has provided agricultural, residential, industrial and infrastructural damages that are caused due to 2022 flood in the lower Indus basin. The research has explicitly discussed the data availability and sources of district level dataset for quantification of related damages in the concerned sections. The research has a limitation that it has not provided quantified economic damages for educational and medical units as for estimation of construction cost of the buildings, geo-referenced data points/coordinates and the total area covered by the educational and medical institutes at district level are required. For obtaining this data, the international researchers have utilized satellite imageries with fine resolution of 1 m which is unavailable for Pakistan. Thus, due to data limitation the research has not estimated costs for medical and educational institutes damages. In order to overcome this limitation, there needs to be conducted a field survey for mapping geo-referenced coordinates showing the location of each medical and educational units in districts along-with the total area covered by each

institute. This will fill the data limitation for estimation of flood damages for medical and educational units. Thus, this domain can be explored in future research.

h) The research has only provided forecasted agricultural damages for 2032 flood. The method demonstrated in this research can be utilized to forecast flood for any time period and for assessing forecasted asset damage for the rest of the categories in future research.

References

Abbasi, H. U., Baloch, M. A., & Memon, A. G. (2011). Deforestation analysis of riverine forest of sindh using remote sensing techniques. *Mehran University Research Journal of Engineering & Technology*, *30*(3).

Abdullah, A. F. (2015). Flood damage assessment in agricultural area in Selangor river basin. *Jurnal Teknologi*, *76*(15).

Acosta, J. E., De Leon, R. K. L., Hollite, J. R. D., Logronio, R. M., & James, G. R. (2017, April). Flood modeling using GIS and LiDAR of padada river in southeastern Philippines. In *International Conference on Geographical Information Systems Theory, Applications and Management* (Vol. 2, pp. 301-306). SCITEPRESS.

Adnan, M., Khan, F., Rehman, N., Ali, S., Hassan, S. S., Dogar, M. M., & Hasson, S. (2021). Variability and predictability of summer monsoon rainfall over Pakistan. *Asia-Pacific Journal of Atmospheric Sciences*, *57*(1), 89-97.

Afifi A. (2013). Improving the Classification Accuracy using Support Vector Machines (SVMS) with New Kernel. *Journal of Global Research in Computer Science*, *4*(2), 1-7.

Agrawal P., H. F. (2021). Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019). *IEEE*.

Ahmad, A., & Dey, L. (2005). A feature selection technique for classificatory analysis. *Pattern Recognition Letters*, *26*(1), 43-56.

Ahmadlou, M., Karimi, M., Alizadeh, S., Shirzadi, A., Parvinnejhad, D., Shahabi, H., & Panahi, M. (2019). Flood susceptibility assessment using integration of adaptive network-based fuzzy inference system (ANFIS) and biogeography-based optimization (BBO) and BAT algorithms (BA). *Geocarto International*, *34*(11), 1252-1272.

Ahmadisharaf, E., Kalyanapu, A. J., & Chung, E. S. (2015). Evaluating the effects of inundation duration and velocity on selection of flood management alternatives using multi-criteria decision making. *Water Resources Management*, *29*, 2543-2561.

Ahmadisharaf, E., Kalyanapu, A. J., & Chung, E. S. (2016). Spatial probabilistic multi-criteria decision making for assessment of flood management alternatives. *Journal of Hydrology*, *533*, 365-378.

Ajibade S.S, Oyebode OJ, Mejarito CL, Gido NG, Dayupay J, Diaz RD (2022) Feature Selection for Student Prediction Accuracy using Gravitational Search Algorithm. *J Optoelectron Laser,* 41(8):2022.

Aksoy, H., Kirca, V. S. O., Burgan, H. I., & Kellecioglu, D. (2016). Hydrological and hydraulic models for determination of flood-prone and flood inundation areas. *Proceedings of the International Association of Hydrological Sciences*, *373*, 137-141.

Ali, M., Khan, S. J., Aslam, I., & Khan, Z. (2011). Simulation of the impacts of land-use change on surface runoff of Lai Nullah Basin in Islamabad, Pakistan. *Landscape and Urban Planning*, *102*(4), 271-279.

Ali, S. A., Khatun, R., Ahmad, A., & Ahmad, S. N. (2019). Application of GIS-based analytic hierarchy process and frequency ratio model to flood vulnerable mapping and risk area estimation at Sundarban region, India. *Modeling Earth Systems and Environment*, *5*, 1083-1102.

Ali, S., Khalid, B., Kiani, R. S., Babar, R., Nasir, S., Rehman, N., & Goheer, M. A. (2020). Spatio-temporal variability of summer monsoon onset over Pakistan. *Asia-Pacific Journal of Atmospheric Sciences*, *56*, 147-172.

Ali, S. A., Parvin, F., Vojteková, J., Costache, R., Linh, N. T. T., Pham, Q. B., & Ghorbani, M. A. (2021). GIS-based landslide susceptibility modeling: A comparison between fuzzy multi-criteria and machine learning algorithms. *Geoscience Frontiers*, *12*(2), 857-876.

Anshuka, A., van Ogtrop, F. F., Sanderson, D., & Leao, S. Z. (2022). A systematic review of agent-based model for flood risk management and assessment using the ODD protocol. *Natural Hazards*, *112*(3), 2739-2771.

Arabameri A, Rezaei K, Cerda` A, Conoscenti C, Kalantari Z (2019) A comparison of statistical methods and multi-criteria decision making to map flood hazard susceptibility in Northern Iran. *Sci Total Environ, 660*:443–458.

Arabameri, A., Saha, S., Roy, J., Chen, W., Blaschke, T., & Tien Bui, D. (2020). Landslide susceptibility evaluation and management using different machine learning methods in the Gallicash River Watershed, Iran. *Remote Sensing*, *12*(3), 475.

Arabameri A, Chen W, Loche M, Zhao X, Li Y, Lombardo L, Cerda A, Pradhan B, Bui DT (2020a) Comparison of machine learning models for gully erosion susceptibility mapping. *Geosci Front 11*:1609–1620.

Arabameri A, Saha S, Mukherjee K, Blaschke T, Chen W, Ngo PTT, Band SS (2020b) Modeling spatial flood using novel ensemble artificial intelligence approaches in northern Iran. *Remote Sens 12:*1–30.

Araújo L.A, Leite Lopes I., Menali Oliveira R., Godinho Silva S. H., Jarochinski Silva C. S., & Rezende Gomide L. (2022). Simulated Annealing in Feature Selection Approach for Modelling above Ground Carbon Stock at the Transition between Brazilian Savanna and Atlantic Forest Biomes. *Annals of Forest Research (1844-8135)*, *65*(1).

Arora A., Arabameri A., Pandey M., Siddiqui M. A., Shukla U. K., Bui D. T., & Bhardwaj A. (2021). Optimization of State-of-the-Art Fuzzy-Metaheuristic ANFIS-Based Machine Learning Models for Flood Susceptibility Prediction Mapping in the Middle Ganga Plain, India. *Science of the Total Environment*, *750*, 141565.

Arnell, N. W., & Gosling, S. N. (2016). The impacts of climate change on river flood risk at the global scale. *Climatic Change*, *134*, 387-401.

Arrighi, C., Rossi, L., Trasforini, E., Rudari, R., Ferraris, L., Brugioni, M., ... & Castelli, F. (2018). Quantification of flood risk mitigation benefits: A building-scale damage assessment through the RASOR platform. *Journal of environmental management*, *207*, 92-104.

Arunyanart, N., Limsiri, C., & Uchaipichat, A. (2017). Flood hazards in the Chi River Basin, Thailand: impact management of climate change. *Applied Ecology & Environmental Research*, *15*(4).

Ashwini Venkatasubramaniam, A. V., Wolfson, J., Mitchell, N., Barnes, T., JaKa, M., & French, S. (2017). Decision trees in epidemiological research.

Aslam, B., Maqsoom, A., Inam, H., Basharat, M. U., & Ullah, F. (2023). Forecasting Construction Cost Index through Artificial Intelligence. *Societies*, *13*(10), 219.

Ay Ş., Ekinci E., & Garip Z. (2023). A Comparative Analysis of Meta-Heuristic Optimization Algorithms for Feature Selection on ML-Based Classification of Heart-Related Diseases. *The Journal of Supercomputing*, 1-30.

Avand, M., & Moradi, H. (2021). Spatial modeling of flood probability using geo-environmental variables and machine learning models, case study: Tajan watershed, Iran. *Advances in Space Research*, *67*(10), 3169-3186.

Aziz, F. (2022). Pakistan floods: What role did climate change play? *The Conversation*. Retrieved from https://theconversation.com/pakistan-floods-what-role-did-climate-change-play-189833

Balakrishnan K, Dhanalakshmi R, Khaire UM (2021) Improved Salp Swarm Algorithm Based on the Levy Flight for Feature Selection. *J Supercomput*, 77(11):12399–12419.

Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: simple definition and calculation of accuracy, sensitivity and specificity.

Basset A.M. (2018). Metaheuristic Algorithms: A Comprehensive Review. *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications, Elsevier*.

Beasley, T. M., & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational statistics & data analysis*, *42*(4), 569-593.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24-31.

Beven KJ, Kirkby MJ. (1979). A Physically Based, Variable Contributing Area Model of Basin Hydrology. *Hydrol Sci Bull*, 24:43–69.

Bhargavi, P., & Jyothi, S. (2009). Applying naive bayes data mining technique for classification of agricultural land soils. *International journal of computer science and network security*, *9*(8), 117-122.

Bhutto (2022). *The west is ignoring Pakistan's super-floods*. Heed this warning: tomorrow it will be you. The Guardian. Retrieved from https://www.theguardian.com/commentisfree/2022/sep/08/pakistan-floods-climate-crisis

Bing, L., Shao, Q., & Liu, J. (2011, June). Runoff characteristic in flood and dry seasons based on wavelet analysis in the source regions of Yangtze and Yellow River. In *2011 International Conference on Remote Sensing, Environment and Transportation Engineering* (pp. 705-710). IEEE.

Biswajeet, P., & Mardiana, S. (2009). Flood hazrad assessment for cloud prone rainy areas in a typical tropical environment. *Disaster Advances*, *2*(2), 7-15.

Biswajeet Pradhan, B. P. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS.

Blistanova M., Zeleňáková M., Blistan P., & Ferencz V. (2016). Assessment of Flood Vulnerability in Bodva River Basin, Slovakia. *Acta Montanistica Slovaca*, *21*(1).

Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information fusion*, *52*, 1-12.

Bonabeau E., Dorigo M., & Théraulaz G. (1999). *Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press*

Borga, M., Anagnostou, E. N., Blöschl, G., & Creutin, J. D. (2011). Flash flood forecasting, warning and risk management: the HYDRATE project. *Environmental Science & Policy*, *14*(7), 834-844.

Bormudoi, A., Huy, H. Q., Hazarika, M. K., & Samarakoon, L. (2013). Integration of remote sensing data with a numerical model to prepare accurate flood hazard maps for effective flood management in the mekong delta. In *34th Asian conference on remote sensing* (pp. 3637-3645).

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Brown, P., Daigneault, A., & Gawith, D. (2017). Climate change and the economic impacts of flooding on Fiji. *Climate and Development*, *9*(6), 493-504.

Bui, D. T., Lofman, O., Revhaug, I., & Dick, O. (2011). Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural hazards*, *59*, 1413-1444.

Bui, D. T., Pradhan, B., Nampak, H., Bui, Q. T., Tran, Q. A., & Nguyen, Q. P. (2016). Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic

optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. *Journal of Hydrology*, *540*, 317-330.

Bui D. T., Panahi M., Shahabi H., Singh V. P., Shirzadi A., Chapi K., & Ahmad B. B. (2018). Novel Hybrid Evolutionary Algorithms for Spatial Prediction of Floods. *Scientific reports*, *8*(1), 15364.

Bui, D. T., Tsangaratos, P., Ngo, P. T. T., Pham, T. D., & Pham, B. T. (2019). Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods. *Science of the total environment*, *668*, 1038-1054.

Cao, C., Xu, P., Wang, Y., Chen, J., Zheng, L., & Niu, C. (2016). Flash flood hazard susceptibility mapping using frequency ratio and statistical index methods in coalmine subsidence areas. *Sustainability*, *8*(9), 948.

Cham, T. C., & Mitani, Y. (2015). Flood control and loss estimation for paddy field at midstream of Chao Phraya River Basin, Thailand. In *IOP conference series: earth and environmental science* (Vol. 26, No. 1, p. 012022). IOP Publishing.

Chapi, K., Singh, V. P., Shirzadi, A., Shahabi, H., Bui, D. T., Pham, B. T., & Khosravi, K. (2017). A novel hybrid artificial intelligence approach for flood susceptibility assessment. *Environmental modelling & software*, *95*, 229-245.

Chen, H. L., Yang, B., Wang, G., Wang, S. J., Liu, J., & Liu, D. Y. (2012). Support vector machine based diagnostic system for breast cancer using swarm intelligence. *Journal of medical systems*, *36*, 2505-2519.

Chen, J., Yu, Z., Zhu, Y., & Yang, C. (2011). Relationship between land use and evapotranspiration-a case study of the Wudaogou Area in Huaihe River basin. *Procedia Environmental Sciences*, *10*, 491-498.

Chen, W., Zhang, S., Li, R., & Shahabi, H. (2018). Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Science of the total environment*, *644*, 1006-1018.

Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., & Ahmad, B. B. (2020). Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. *Science of The Total Environment*, *701*, 134979.

Chen, W., Shahabi, H., Shirzadi, A., Li, T., Guo, C., Hong, H., & Bin Ahmad, B. (2018). A novel ensemble approach of bivariate statistical-based logistic model tree classifier for landslide susceptibility assessment. *Geocarto International*, *33*(12), 1398-1420.

Chen, W., Pradhan, B., Li, S., Shahabi, H., Rizeei, H. M., Hou, E., & Wang, S. (2019). Novel hybrid integration approach of bagging-based fisher's linear discriminant function for groundwater potential analysis. *Natural Resources Research*, *28*, 1239-1258.

Chen, W., Zhao, X., Shahabi, H., Shirzadi, A., Khosravi, K., Chai, H., & Li, R. (2019). Spatial prediction of landslide susceptibility by combining evidential belief function, logistic regression and logistic model tree. *Geocarto International*, *34*(11), 1177-1201.

Chen, X., Quan, Q., Zhang, K., & Wei, J. (2021). Spatiotemporal characteristics and attribution of dry/wet conditions in the Weihe River Basin within a typical monsoon transition zone of East Asia over the recent 547 years. *Environmental Modelling & Software*, *143*, 105116.

Chen, Y., Xu, Y., & Yin, Y. (2009). Impacts of land use change scenarios on storm-runoff generation in Xitiaoxi basin, China. *Quaternary International*, *208*(1-2), 121-128.

Chen, Z., Liu, Z., Yin, L., & Zheng, W. (2022). Statistical analysis of regional air temperature characteristics before and after dam construction. *Urban Climate*, *41*, 101085.

Cheng, D., Zhang, S., Deng, Z., Zhu, Y., & Zong, M. (2014). k NN algorithm with data-driven k value. In *Advanced Data Mining and Applications: 10th International Conference, ADMA 2014, Guilin, China, December 19-21, 2014. Proceedings 10* (pp. 499-512). Springer International Publishing.

Chowdhuri I., Pal S.C., Chakrabortty R. (2020). Flood Susceptibility Mapping by Ensemble Evidential Belief Function and Binomial Logistic Regression Model on River Basin of Eastern India. *Adv. Space Res. 65 (5),* 1466–1489

Christidis N., Stott P. A., Scaife A. A., Arribas A., Jones G. S., Copsey D. (2013). A new HadGEM3-A-based system for attribution of weather- and climate-related extreme events. *Journal of Climate*, **26**(9), 2756–2783.

Chung, C. J. F., & Fabbri, A. G. (2003). Validation of spatial prediction models for landslide hazard mapping. *Natural Hazards*, *30*, 451-472.

Crippen, R. E. (1990). Calculating the vegetation index faster. *Remote sensing of Environment*, *34*(1), 71-73.

Costache, R., Pham, Q. B., Avand, M., Linh, N. T. T., Vojtek, M., Vojteková, J., ... & Dung, T. D. (2020). Novel hybrid models between bivariate statistics, artificial neural networks and boosting algorithms for flood susceptibility assessment. *Journal of Environmental Management*, *265*, 110485.

Costea, G. (2013). Deforestation process consequences upon surface runoff coefficients. catchment level case study from the Apuseni Mountains, Romania. *Geographia Technica*, *8*(1), 28-33.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Csatáriné Szabó Z., Mikita T., Négyesi G., Varga O. G., Burai P., Takács-Szilágyi L., & Szabó S. (2020). Uncertainty and Overfitting in Fluvial Landform Classification using Laser Scanned Data and Machine Learning: A Comparison of Pixel and Object-Based Approaches. *Remote Sensing, 12(21),* 3652.

Dai, J., Feng, H., Shi, K., Ma, X., Yan, Y., Ye, L., & Xia, Y. (2022). Electrochemical degradation of antibiotic enoxacin using a novel PbO2 electrode with a graphene nanoplatelets inter-layer: characteristics, efficiency and mechanism. *Chemosphere*, *307*, 135833.

Dang, A. T. N., & Kumar, L. (2017). Application of remote sensing and GIS-based hydrological modelling for flood risk analysis: a case study of District 8, Ho Chi Minh city, Vietnam. *Geomatics, Natural Hazards and Risk*, *8*(2), 1792-1811.

De Rosa, P., Fredduzzi, A., & Cencetti, C. (2019). Stream power determination in gis: an index to evaluate the most'Sensitive'points of a river. *Water*, *11*(6), 1145.

Deep K. (2022). A Random Walk Grey Wolf Optimizer Based on Dispersion Factor for Feature Selection on Chronic Disease Prediction. *Expert Syst Appl*, 206:117864.

Deka P. C. (2014). Support Vector Machine Applications in the Field of Hydrology: A Review. *Applied soft computing*, *19*, 372-386.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, *7*, 1-30.

Dhiman G, Oliva D, Kaur A, Singh KK, Vimal S, Sharma A, Cengiz K (2021). BEPO: A Novel Binary Emperor Penguin Optimizer for Automatic Feature Selection. *Knowl-Based Syst*, 211:106560.

Di Capua, G., Sparrow, S., Kornhuber, K., Rousi, E., Osprey, S., Wallom, D. (2021). Drivers behind the summer 2010 wave train leading to Russian heatwave and Pakistan flooding. *NPJ Climate and Atmospheric Science*, **4**(1), 1–14.

Dibike Y. B., Velickov S., Solomatine D., & Abbott M. B. (2001). Model Induction with Support Vector Machines: Introduction and Applications. *Journal of Computing in Civil Engineering*, *15*(3), 208-216.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, *10*(7), 1895-1923.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, *40*, 139-157.

Dodangeh E., Panahi M., Rezaie F., Lee S., Bui D. T., Lee C. W., & Pradhan B. (2020). Novel Hybrid Intelligence Models for Flood-Susceptibility Prediction: Meta Optimization of the GMDH and SVR Models with the Genetic Algorithm and Harmony Search. *Journal of Hydrology*, *590*, 125423.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241-258.

Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. *IEEE transactions on systems, man, and cybernetics, part b (cybernetics)*, *26*(1), 29-41.

Dutta, D., Herath, S., & Musiake, K. (2003). A mathematical model for flood loss estimation. *Journal of hydrology*, *277*(1-2), 24-49.

Eberhart, R., & Kennedy, J. (1995, October). A new optimizer using particle swarm theory. In *MHS'95. Proceedings of the sixth international symposium on micro machine and human science* (pp. 39-43). Ieee.

Eisavi, V., Homayouni, S., Yazdi, A. M., & Alimohammadi, A. (2015). Land cover mapping based on random forest classification of multitemporal spectral and thermal images. *Environmental monitoring and assessment*, *187*, 1-14.

Ettinger, S., Mounaud, L., Magill, C., Yao-Lafourcade, A. F., Thouret, J. C., Manville, V., & Llerena, N. M. (2016). Building vulnerability to hydro-geomorphic hazards: Estimating damage probability from qualitative vulnerability assessment using logistic regression. *Journal of Hydrology*, *541*, 563-581.

Eskandari S, Seifaddini M (2022). Online and Offline Streaming Feature Selection Methods with Bat Algorithm for Redundancy Analysis. *Pattern Recognit*, 133:109007.

Eslamian, S. (Ed.). (2014). *Handbook of engineering hydrology: modeling, climate change, and variability*. CRC Press.

Ethem, A. (2014). Introduction to Machine Learning: Massachusetts Institute of Technology.

Faghih, M., Mirzaei, M., Adamowski, J. (2017). Uncertainty estimation in food inundation mapping: an application of non-parametric bootstrapping. *River Res Appl 33*:611–619.

Forsati, R., Moayedikia, A., Jensen, R., Shamsfard, M., & Meybodi, M. R. (2014). Enriched ant colony optimization and its application in feature selection. *Neurocomputing*, *142*, 354-371.

Foudi, S., Osés-Eraso, N., & Tamayo, I. (2015). Integrated spatial flood risk assessment: The case of Zaragoza. *Land Use Policy*, *42*, 278-292.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, *32*(200), 675-701.

Galathiya, A. S., Ganatra, A. P., & Bhensdadia, C. K. (2012). Improved decision tree induction algorithm with feature selection, cross validation, model complexity and reduced error pruning. *International Journal of Computer Science and Information Technologies*, *3*(2), 3427-3431.

Gan, B. R., Liu, X. N., Yang, X. G., Wang, X. K., & Zhou, J. W. (2018). The impact of human activities on the occurrence of mountain flood hazards: lessons from the 17 August 2015 flash

flood/debris flow event in Xuyong County, south-western China. *Geomatics, Natural Hazards and Risk, 9*(1), 816-840.

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information sciences, 180*(10), 2044-2064.

Gauhar N., Das S., & Moury K. S. (2021). Prediction of Flood in Bangladesh using K-Nearest Neighbors Algorithm. In *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 357-361). *IEEE.*

Gayen, A., Pourghasemi, H. R., Saha, S., Keesstra, S., & Bai, S. (2019). Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. *Science of the total environment, 668*, 124-138.

Gergel'ová M, Kuzevičová Ž, Kuzevič Š, Sabolová J (2013) Hydrodynamic modeling and GIS tools applied in urban areas. Acta Montan Slovaca 18:226–233

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. *Pattern recognition letters, 27*(4), 294-300.

Guan, D., Yuan, W., Lee, Y. K., Najeebullah, K., & Rasel, M. K. (2014). A review of ensemble learning based feature selection. *IETE Technical Review, 31*(3), 190-198.

Gunavathi C., & Premalatha K. (2014). Performance Analysis of Genetic Algorithm with KNN and SVM for Feature Selection in Tumor Classification. *International Journal of Computer and Information Engineering*, *8*(8), 1490-1497.

Guo, Y., Yang, Y., Kong, Z., & He, J. (2022). Development of similar materials for liquid-solid coupling and its application in water outburst and mud outburst model test of deep tunnel. *Geofluids, 2022*.

Gusyev, M. A., Kwak, Y., Khairul, M. I., Arifuzzaman, M. B., Magome, J., Sawano, H., & Takeuchi, K. (2015). Effectiveness of water infrastructure for river flood management–Part 1: Flood hazard assessment using hydrological models in Bangladesh. *Proceedings of the International Association of Hydrological Sciences, 370*(370), 75-81.

Han, D., Kwong, T., & Li, S. (2007). Uncertainties in real-time flood forecasting with neural networks. *Hydrological Processes: An International Journal, 21*(2), 223-228.

Hanbay K. (2022). A New Standard Error Based Artificial Bee Colony Algorithm and its Applications in Feature Selection. *J King Saud Univ Comput Inf Sci*, 34(7):4554–4567.

Hashemi A, Joodaki M, Joodaki NZ, Dowlatshahi MB (2022). Ant Colony Optimization Equipped with an Ensemble of Heuristics through Multi-Criteria Decision Making: A Case Study in Ensemble Feature Selection. *Appl Soft Comput*, 124:109046.

Hashmi H. N., Siddiqui Q. T. M., Ghumman A. R., Kamal M. A., & Mughal H. U. R. (2012). A critical analysis of 2010 floods in Pakistan. *African Journal of Agricultural Research, 7*(7), 1054-1067.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. *The elements of statistical learning: data mining, inference, and prediction*, 337-387.

Hirabayashi Y., Mahendran R., Koirala S., Konoshima L., Yamazaki D., Watanabe S., (2013). Global flood risk under climate change. *Nature Climate Change*, **3**(9), 816–821.

Ho, J. C. (2009). Coastal flood risk assessment and coastal zone management: A case study of Seberang Perai and Kuantan Pekan in Malaysia.

Hoang N. D., & Tran X. L. (2021). Remote Sensing–Based Urban Green Space Detection using Marine Predators Algorithm Optimized Machine Learning Approach. *Mathematical Problems in Engineering, 2021*, 1-22.

Hoque N., Bhattacharyya D. K., & Kalita J. K. (2014). MIFS-ND: A Mutual Information-Based Feature Selection Method. *Expert Systems with Applications*, *41*(14), 6371-6385

Hong H., Tsangaratos P., Ilia I., Liu J., Zhu A. X., & Chen W. (2018). Application of Fuzzy Weight of Evidence and Data Mining Techniques in Construction of Flood Susceptibility Map of Poyang County, China. *Science of the Total Environment*, *625*, 575-588.

Hong, C. C., Hsu, H. H., Lin, N. H., & Chiu, H. (2011). Roles of European blocking and tropical-extratropical interaction in the 2010 Pakistan flooding. *Geophysical Research Letters*, 38(13), L13806.

Hong H, Liu J, Bui DT, Pradhan B, Acharya TD, Pham BT, Zhu AX, Chen W, Ahmad B, Bin, (2018) Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *CATENA 163*:399–413.

Hossain, M. N., Uddin, M. N., Rokanuzzaman, M., Miah, M. A., & Alauddin, M. (2015). Effects of flooding on socio-economic status of two integrated char lands of Jamuna River, Bangladesh. *Journal of Environmental Science and Natural Resources*, *6*(2), 37-41.

Huang, Y., Bárdossy, A., & Zhang, K. (2019). Sensitivity of hydrological models to temporal and spatial resolutions of rainfall data. *Hydrology and Earth System Sciences*, *23*(6), 2647-2663.

Huizinga, J., De Moel, H., & Szewczyk, W. (2017). *Global flood depth-damage functions: Methodology and the database with guidelines* (No. JRC105688). Joint Research Centre (Seville site).

Hussain E., Ural S., Malik A., Shan J. (2011). Mapping Pakistan 2010 floods using remote sensing data. In Proceedings of the American Society for Photogrammetry and Remote Sensing Annual Conference, Milwaukee, WI, USA, 15–222.

Islam, K.M.N. (2000). Impact of floods in Bangladesh, in Floods. *Routledge Hazards and Disaster* Series, 1, 156–171.

Islam, M. M., Yao, X., Nirjon, S. S., Islam, M. A., & Murase, K. (2008). Bagging and boosting negatively correlated neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *38*(3), 771-784.

Islam, M., Shehzad, F., Ray, S., & Abbas, M. W. (2023). Forecasting the population growth and wheat crop production in Pakistan with non-linear growth and ARIMA models. *Population and Economics*, *7*(3), 172-187.

James, L. D., & Lee, R. R. (1971). Economic of water resources planning MC Graw-Hill. *New Delhi*, 20.

James, L. D., & Hall, B. (1986). Risk information for floodplain management. *Journal of Water Resources Planning and Management*, *112*(4), 485-499.

Javidan N., Kavian A., Pourghasemi H. R., Conoscenti C., & Jafarian Z. (2020). Data Mining Technique (Maximum Entropy Model) for Mapping Gully Erosion Susceptibility in the Gorganrood Watershed, Iran. *Gully Erosion Studies from India and Surrounding Regions*, 427-448.

Jensen R. and Shen Q. (2005). Fuzzy-rough Data Reduction with Ant Colony Optimization. *Fuzzy Sets and Systems*, 149, 5–20.

Jin, Y., Liu, X., Chen, Y., & Liang, X. (2018). Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: A case study of central Shandong. *International journal of remote sensing*, *39*(23), 8703-8723.

Jones, B. (2022). *Pakistan flooding: How melting glaciers fueled the disaster*. Vox. Retrieved from https://www.vox.com/science-and-health/2022/8/30/23327341/pakistan-flooding-monsoon-melting-glaciers-climate-change

Jović A., Brkić K., & Bogunović N. (2015). A Review of Feature Selection Methods with Applications. In *2015 38th International Convention on Information And Communication Technology, Electronics And Microelectronics (MIPRO)* (pp. 1200-1205). *IEEE*.

Kancherla D., Bodapati J.D., & Veeranjaneyulu N. (2019). Effect of Different Kernels on the Performance of an SVM-Based Classification. *International Journal of Recent Technology and Engineering*, *7*.

Karamouz, M., Zahmatkesh, Z., Goharian, E., & Nazif, S. (2015). Combined impact of inland and coastal floods: Mapping knowledge base for development of planning strategies. *Journal of Water Resources Planning and Management*, *141*(8), 04014098.

Katipoğlu, O. M., & Sarıgöl, M. (2023). Prediction of flood routing results in the Central Anatolian region of Türkiye with various machine learning models. *Stochastic Environmental Research and Risk Assessment*, *37*(6), 2205-2224.

katipoğlu, O. M., & Sarıgöl, M. (2023). Boosting flood routing prediction performance through a hybrid approach using empirical mode decomposition and neural networks: a case study of the Mera River in Ankara. *Water Supply*, *23*(11), 4403-4415.

Kecman V. (2001). Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. *MIT Press*.

Kennedy, J., & Eberhart, R. C. (1997, October). A discrete binary version of the particle swarm algorithm. In *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation* (Vol. 5, pp. 4104-4108). IEEE.

Kiefer, J. C., & Willett, J. S. (1996). *Analysis of nonresidential content value and depth-damage data for flood damage reduction studies* (p. 0115). US Army Corps of Engineers, Water Resources Support Center, Institute for Water Resources.

Khosravi, K., Nohani, E., Maroufinia, E., & Pourghasemi, H. R. (2016). A GIS-based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique. *Natural hazards*, *83*, 947-987.

Khosravi, K., Pourghasemi, H. R., Chapi, K., & Bahri, M. (2016). Flash flood susceptibility analysis and its mapping using different bivariate models in Iran: a comparison between Shannon's entropy, statistical index, and weighting factor models. *Environmental monitoring and assessment*, *188*, 1-21.

Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., & Bui, D. T. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science of the Total Environment*, *627*, 744-755.

Khosravi, K., Melesse, A. M., Shahabi, H., Shirzadi, A., Chapi, K., & Hong, H. (2019). Flood susceptibility mapping at Ningdu catchment, China using bivariate and data mining techniques. In Extreme hydrology and climate variability. *Elsevier*, 419-434.

Khwaja, A. S., Anpalagan, A., Naeem, M., & Venkatesh, B. (2020). Joint bagged-boosted artificial neural networks: Using ensemble machine learning to improve short-term electricity load forecasting. *Electric Power Systems Research*, *179*, 106080.

Klein, A. G., Gerhard, C., Büchner, R. D., Diestel, S., & Schermelleh-Engel, K. (2016). The detection of heteroscedasticity in regression models for psychological data. Psychological Test and Assessment Modeling, 58(4), 567.

Kobayashi, K., Takara, K., Sano, H., Tsumori, H., & Sekii, K. (2016). A high-resolution large-scale flood hazard and economic risk model for the property loss insurance in Japan. *Journal of Flood Risk Management*, *9*(2), 136-153.

Kohavi, R. (1996, August). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd* (Vol. 96, pp. 202-207).

Komolafe, A. A., Herath, S., & Avtar, R. (2018). Methodology to assess potential flood damages in urban areas under the influence of climate change. *Natural Hazards Review*, *19*(2), 05018001.

Komolafe, A. A., Herath, S., & Avtar, R. (2018). Development of generalized loss functions for rapid estimation of flood damages: a case study in Kelani River basin, Sri Lanka. *Applied Geomatics*, *10*, 13-30.

Komolafe, A. A., Herath, S., Avtar, R., & Vuillaume, J. F. (2019). Comparative analyses of flood damage models in three Asian countries: towards a regional flood risk modelling. *Environment systems and decisions*, *39*, 229-246.

Kuiper, E. (1971). Water resources project economics.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine learning*, *59*, 161-205.

Lawrence, C. B., Pindilli, E. J., & Hogan, D. M. (2019). Valuation of the flood attenuation ecosystem service in Difficult Run, VA, USA. *Journal of environmental management*, *231*, 1056-1064.

Lee, S., & Oh, H. J. (2012). Ensemble-based landslide susceptibility maps in Jinbu area, Korea. Terrigenous Mass Movements: Detection, Modelling, Early Warning and Mitigation Using Geoinformation Technology, 193-220.

Lee, W. K., & Mohamad, I. N. (2014). Flood economy appraisal: an overview of the Malaysian scenario. In *InCIEC 2013: Proceedings of the International Civil and Infrastructure Engineering Conference 2013* (pp. 263-274). Springer Singapore.

Lekuthai, A., & Vongvisessomjai, S. (2001). Intangible flood damage quantification. *Water Resources Management*, *15*, 343-362.

Lieskovský, J., Kaim, D., Balázs, P., Boltižiar, M., Chmiel, M., Grabska, E., ... & Radeloff, V. C. (2018). Historical land use dataset of the Carpathian region (1819–1980). *Journal of Maps*, *14*(2), 644-651.

Li Y., Khan M. Y. A., Jiang Y., Tian F., Liao W., Fu S., & He C. (2019). CART and PSO+ KNN Algorithms to Estimate the Impact of Water Level Change on Water Quality in Poyang Lake, China. *Arabian Journal of Geosciences*, *12*, 1-12.

Lin L., & Gen M. (2009). Auto-Tuning Strategy for Evolutionary Algorithms: Balancing between Exploration and Exploitation. *Soft Computing*, *13*, 157-168.

Liu H., & Yu L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on knowledge and data engineering*, *17*(4), 491-502.

Liu, Y., Zhang, Z., Liu, X., Wang, L., & Xia, X. (2021). Efficient image segmentation based on deep learning for mineral image classification. *Advanced Powder Technology*, *32*(10), 3885-3903.

Liu, E., Chen, S., Yan, D., Deng, Y., Wang, H., Jing, Z., & Pan, S. (2022). Detrital zircon geochronology and heavy mineral composition constraints on provenance evolution in the western Pearl River Mouth basin, northern south China sea: A source to sink approach. *Marine and Petroleum Geology*, *145*, 105884.

Luino, F., Cirio, C. G., Biddoccu, M., Agangi, A., Giulietto, W., Godone, F., & Nigrelli, G. (2009). Application of a model to the evaluation of flood damage. *Geoinformatica*, *13*, 339-353.

Mahmood, S., Rahman, A. U., & Shaw, R. (2019). Spatial appraisal of flood risk assessment and evaluation using integrated hydro-probabilistic approach in Panjkora River Basin, Pakistan. *Environmental Monitoring and Assessment*, *191*, 1-15.

Mallapaty, S. (2022). Why are Pakistan's floods so extreme this year? *Nature*.

Mao, D., & Cherkauer, K. A. (2009). Impacts of land-use change on hydrologic responses in the Great Lakes region. *Journal of Hydrology*, *374*(1-2), 71-82.

McBean, E. A., Gorrie, J., Fortin, M., Ding, J., & Moulton, R. (1988). Flood Depth—Damage Curves by Interview Survey. *Journal of Water Resources Planning and Management*, *114*(6), 613-634.

Merz, B., Hall, J., Disse, M., & Schumann, A. (2010). Fluvial flood risk management in a changing world. *Natural Hazards and Earth System Sciences*, *10*(3), 509-527.

Merz, B., Kreibich, H., Schwarze, R., & Thieken, A. (2010). Review article" Assessment of economic flood damage". *Natural Hazards and Earth System Sciences*, *10*(8), 1697-1724.

Merz, B., Hall, J., Disse, M., & Schumann, A. (2010). Fluvial flood risk management in a changing world. *Natural Hazards and Earth System Sciences*, *10*(3), 509-527.

Micheletti, N., Foresti, L., Kanevski, M., Pedrazzini, A., & Jaboyedoff, M. (2011). Landslide susceptibility mapping using adaptive support vector machines and feature selection. *Geophysical Research Abstracts, EGU*, *13*.

Middelmann-Fernandes, M. H. (2010). Flood damage estimation beyond stage–damage functions: an Australian example. *Journal of Flood Risk Management*, *3*(1), 88-96.

Mihu-Pintilie, A., Cîmpianu, C. I., Stoleriu, C. C., Pérez, M. N., & Paveluc, L. E. (2019). Using high-density LiDAR data and 2D streamflow hydraulic modeling to improve urban flood hazard maps: A HEC-RAS multi-scenario approach. *Water*, *11*(9), 1832.

Minea, G. (2013). Assessment of the flash flood potential of Basca river catchment (Romania) based on physiographic factors. *Open Geosciences*, *5*(3), 344-353.

Mirjalili S. M. S. M., Mirjalili S. M., & Lewis A. (2014). Grey Wolf Optimizer Adv *Eng Softw* 69: 46–61.

MOC, (1996). Flood Damage Statistics in Japan, Technical Report, River Engineering Bureau, Ministry of Construction, Japan.

Mohamed, W. N. H. W., Salleh, M. N. M., & Omar, A. H. (2012, November). A comparative study of reduced error pruning method in decision tree algorithms. In *2012 IEEE International conference on control system, computing and engineering* (pp. 392-397). IEEE.

Mohammadi, S. A., Nazariha, M., & Mehrdadi, N. (2014). Flood damage estimate (quantity), using HEC-FDA model. Case study: the Neka river. *Procedia Engineering*, *70*, 1173-1182.

Mojaddadi Rizeei, H. (2018). *Flood Risk Assessment using Multi-Sensor Remote Sensing, Geographic Information System, 2D Hydraulic And Machine Learning Based Models* (Doctoral dissertation).

Morita, M. (2014). Flood risk impact factor for comparatively evaluating the main causes that contribute to flood risk in urban drainage areas. *Water*, *6*(2), 253-270.

Mosavi A., Golshan M., Janizadeh S., Choubin B., Melesse A.M., Dineva A.A. (2020). Ensemble Models of GLM, FDA, MARS, and RF for Flood and Erosion Susceptibility Mapping: A Priority Assessment of Sub-basins. *Geocarto Int.,* 1–20.

Mukherjee F., Singh D. (2020). Detecting Flood Prone Areas in Harris County: A GIS-Based Analysis. *Geo Journal, 85*, 647–663.

Munteanu, C., Kuemmerle, T., Boltiziar, M., Lieskovský, J., Mojses, M., Kaim, D., ... & Radeloff, V. C. (2017). Nineteenth-century land-use legacies affect contemporary land abandonment in the Carpathians. *Regional environmental change*, *17*, 2209-2222.

Nanditha J. S., Kushwaha A. P., Singh R., Malik I., Solanki H., Chuphal D. S., & Mishra V. (2023). The Pakistan Flood of August 2022: Causes and Implications. *Earth's Future*, *11*(3), e2022EF003230.

Neubert, M., Naumann, T., Hennersdorf, J., & Nikolowski, J. (2016). The geographic information system-based flood damage simulation model HOWAD. *Journal of Flood Risk Management*, *9*(1), 36-49.

NDMA. (2022). NDMA monsoon 2022 daily situation report No 093. Retrieved from http://cms.ndma.gov.pk/storage/app/public/situationreports/September2022/Erjx3YwEjYMYLBiOeOx7.pdf

Nga, P. H., Takara, K., & Van, N. C. (2018). Integrated approach to analyze the total flood risk for agriculture: The significance of intangible damages–A case study in Central Vietnam. *International Journal of Disaster Risk Reduction*, *31*, 862-872.

Nguyen, Q. H., Ly, H., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., … & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 2021, 1-15.

Nieto P. G., García-Gonzalo E., Fernández J. A., & Muñiz C. D. (2014). Hybrid PSO–SVM Based Method for Long-Term Forecasting of Turbidity in the Nalón River Basin: A Case Study in Northern Spain. *Ecological Engineering*, *73*, 192-200.

Olorunda O., & Engelbrecht A. P. (2008). Measuring Exploration/Exploitation in Particle Swarms using Swarm Diversity. *IEEE.*

Otto, F. E., Zachariah, M., Saeed, F., Siddiqi, A., Shahzad, K., Mushtaq, H., ... & van Aalst, M. (2022). Climate change likely increased extreme monsoon rainfall, flooding highly vulnerable communities in Pakistan: World Weather Attribution Scientific Report.

Owojori A, Xie H (2005) Landsat image-based LULC changes of San Antonio,Texas using advanced atmospheric correction and object-oriented image analysis approaches. Paper presented at the 5th international symposium on remote sensing of urban areas, Tempe, AZ.

Pan J-S, Liu N, Chu S-C (2022) A Competitive Mechanism Based Multi-Objective Differential Evolution Algorithm and its Application in Feature Selection. *Knowl-Based Syst*, 245:108582.

Panahi M., Dodangeh E., Rezaie F., Khosravi K., Van Le H., Lee M. J., & Pham B. T. (2021). Flood Spatial Prediction Modeling using a Hybrid of Meta-Optimization and Support Vector Regression Modeling. *Catena*, *199*, 105114.

Papaioannou, G., Vasiliades, L., & Loukas, A. (2015). Multi-criteria analysis framework for potential flood prone areas mapping. *Water resources management*, *29*, 399-418.

Paul G. C., Saha S., & Hembram T. K. (2019). Application of the GIS-Based Probabilistic Models for Mapping the Flood Susceptibility in Bansloi Sub-Basin of Ganga-Bhagirathi River and their Comparison. *Remote Sensing in Earth Systems Sciences*, *2*, 120-146.

Parker, D. J., Green, C. H., & Thompson, P. M. (1987). Urban flood protection benefits: A project appraisal guide.

Parker, D. J. (1992). The assessment of the economic and social impacts of natural hazards. In *international conference on Preparedness and Mitigation for Natural Disasters*, 92, 28-29.

Parra, F., González, J., Chacón, M., & Marín, M. (2023). Modeling and evaluation of the susceptibility to landslide events using machine learning algorithms in the province of Chañaral, Atacama region, Chile. *Sustainability*, *15*(24), 16806.

Pathak, P., Bhandari, M., Kalra, A., & Ahmad, S. (2016). Modeling floodplain inundation for monument creek, Colorado. In *World Environmental and Water Resources Congress 2016* (pp. 131-140).

Pedergnana, M., Marpu, P. R., Dalla Mura, M., Benediktsson, J. A., & Bruzzone, L. (2013). A novel technique for optimal feature selection in attribute profiles based on genetic algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(6), 3514-3528.

Penning-Rowsell, E. C., & Chatterton, J. B. (1978). The benefits of flood alleviation. A manual of assessment techniques.

Pham, B. T., Tien Bui, D., Dholakia, M. B., Prakash, I., & Pham, H. V. (2016). A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area. *Geotechnical and Geological Engineering*, *34*, 1807-1824.

Pham, B. T., Pradhan, B., Bui, D. T., Prakash, I., & Dholakia, M. B. (2016). A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environmental Modelling & Software*, *84*, 240-250.

Pham, B. T., Tien Bui, D., Prakash, I., & Dholakia, M. B. (2016). Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. *Natural Hazards*, *83*, 97-127.

Pham, B. T., Bui, D. T., Dholakia, M. B., Prakash, I., Pham, H. V., Mehmood, K., & Le, H. Q. (2017). A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS. *Geomatics, Natural Hazards and Risk*, *8*(2), 649-671.

Pham, B. T., Bui, D. T., Prakash, I., & Dholakia, M. B. (2017). Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *Catena, 149*, 52-63.

Pham, B. T., Tien Bui, D., Pourghasemi, H. R., Indra, P., & Dholakia, M. B. (2017). Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. *Theoretical and Applied Climatology*, *128*, 255-273.

Pham, B. T., Prakash, I., & Bui, D. T. (2018). Spatial prediction of landslides using a hybrid machine learning approach based on random subspace and classification and regression trees. *Geomorphology, 303*, 256-270.

Phinzi, K., Abriha, D., Bertalan, L., Holb, I., & Szabó, S. (2020). Machine learning for gully feature extraction based on a pan-sharpened multispectral image: Multiclass vs. Binary approach. *ISPRS International Journal of Geo-Information*, *9*(4), 252.

Pike RJ (1988) The geometric signature: quantifying landslide-terrain types from digital elevation models. Math Geol 20:491–511.

PMD. (2022). Pakistan meteorological department. Retrieved from https://www.pmd.gov.pk/en/

Pourghasemi H. R., Razavi-Termeh S. V., Kariminejad N., Hong H., & Chen W. (2020). An Assessment of Metaheuristic Approaches for Flood Assessment. *Journal of Hydrology*, *582*, 124536.

Pradhan, B. (2009). Groundwater potential zonation for basaltic watersheds using satellite remote sensing data and GIS techniques. *Central European Journal of Geosciences*, *1*, 120-129.

Pradhan, B. (2010). Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing. *Journal of Spatial Hydrology*, *9*(2).

Pradhan, B., & Youssef, A. M. (2011). A 100-year maximum flood susceptibility mapping using integrated hydrological and hydrodynamic models: Kelantan River Corridor, Malaysia. *Journal of Flood Risk Management*, *4*(3), 189-202.

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers & Geosciences, 51*, 350–365.

Pradhan, B., Hagemann, U., Tehrany, M. S., & Prechtel, N. (2014). An easy to use ArcMap based texture analysis program for extraction of flooded areas from TerraSAR-X satellite image. *Computers & geosciences*, *63*, 34-43.

Pham BT, Pradhan BT, Bui D, Prakash I, Dholakia MB (2016) A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). *Environ Modell Softw, 84*:240–250.

Pham BT, Tien Bui D, Prakash I, Dholakia MB (2017) Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA 149*:52–63.

Pham, B. T., Prakash, I., Jaafari, A., & Bui, D. T. (2018). Spatial prediction of rainfall-induced landslides using aggregating one-dependence estimators classifier. *Journal of the Indian Society of Remote Sensing*, *46*, 1457-1470.

Pielke, R. A., & Avissar, R. (1990). Influence of landscape structure on local and regional climate. *Landscape Ecology*, *4*, 133-155.

Polo, J. L., Berzal, F., & Cubero, J. C. (2008). Class-oriented reduction of decision tree complexity. In *Foundations of Intelligent Systems: 17th International Symposium, ISMIS 2008 Toronto, Canada, May 20-23, 2008 Proceedings 17* (pp. 48-57). Springer Berlin Heidelberg.

Prütz, R., & Månsson, P. (2021). A GIS-based approach to compare economic damages of fluvial flooding in the Neckar River basin under current conditions and future scenarios. *Natural Hazards*, *108*(2), 1807-1834.

Qamer, F. M., Abbas, S., Ahmad, B., Hussain, A., Salman, A., Muhammad, S., ... & Thapa, S. (2023). A framework for multi-sensor satellite data to evaluate crop production losses: the case study of 2022 Pakistan floods. *Scientific Reports*, *13*(1), 4240.

Qiao, C., Huang, Q., Chen, T., & Li, Z. (2018). Key algorithms and its realization about snowmelt flood disaster model based on remote sensing and GIS. In *E3S Web of Conferences* (Vol. 53, p. 03058). EDP Sciences.

Qiao, C., Huang, Q. Y., Chen, T., & Chen, Y. M. (2019). Study on snowmelt flood disaster model based on remote sensing and GIS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 709-713.

Quan, Q., Gao, S., Shang, Y., & Wang, B. (2021). Assessment of the sustainability of Gymnocypris eckloni habitat under river damming in the source region of the Yellow River. *Science of the Total Environment*, *778*, 146312.

Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, *27*(3), 221-234.

Quinlan, J. R. (1993). C4. 5: Programming for machine learning. *Morgan Kauffmann*, *38*(48), 49.

Rahman M, Chen N, Elbeltagi A, Islam MM, Alam M, Pourghasemi HR, Tao W, Zhang J, Shufeng T, Faiz H, Baig MA, Dewan A (2021) Application of stacking hybrid machine learning algorithms in delineating multi-type flooding in Bangladesh. J Environ Manage 295:113086

Rahmati, O., Pourghasemi, H. R., & Zeinivand, H. (2016). Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan Province, Iran. *Geocarto International*, *31*(1), 42-70.

Rahmati O., Darabi H., Panahi M., Kalantari Z., Naghibi S. A., Ferreira C. S. S., & Haghighi A. T. (2020). Development of Novel Hybridized Models for Urban Flood Susceptibility Mapping. *Scientific reports*, *10*(1), 12937.

Rai M. K., & Sharma P. (2021). Classification of Malware using Av Labels Technique with Various Approaches. *Indian Journal of Computer Science and Engineering*, 2231-3850.

Rajkumar K. V., & Subrahmanyam K. (2021). A Hybrid ACO-CS Based Optimized KNN Classifier Algorithm for Rainfall Detection & Prediction. *Journal of Theoretical and Applied Information Technology*, *99*(13).

Rasid, H., & Paul, B. K. (1987). Flood problems in Bangladesh: Is there an indigenous solution? *Environmental Management*, *11*, 155-173.

Razavi-Termeh, S. V., Sadeghi-Niaraki, A., & Choi, S. M. (2019). Groundwater potential mapping using an integrated ensemble of three bivariate statistical models with random forest and logistic model tree models. *Water*, *11*(8), 1596.

Ren H., Pang B., Bai P., Zhao G., Liu S., Liu Y., & Li M. (2024). Flood Susceptibility Assessment with Random Sampling Strategy in Ensemble Learning (RF and XGBoost). *Remote Sensing*, *16*(2), 320.

Rodriguez-Galiano, V. F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P. M., & Jeganathan, C. J. R. S. E. (2012). Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, *121*, 93-107.

Rogger, M., Agnoletti, M., Alaoui, A., Bathurst, J. C., Bodner, G., Borga, M., & Blöschl, G. (2017). Land use change impacts on floods at the catchment scale: Challenges and opportunities for future research. *Water resources research*, *53*(7), 5209-5219.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., ... & Hsiang, S. (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, *12*(1), 4392.

Romali, N. S., Sulaiman, M. S. A. K., Yusop, Z., & Ismail, Z. (2015). Flood damage assessment: A review of flood stage–damage function curve. In *ISFRAM 2014: Proceedings of the International Symposium on Flood Research and Management* (pp. 147-159). Springer Singapore.

Ronco, P., Gallina, V., Torresan, S., Zabeo, A., Semenzin, E., Critto, A., & Marcomini, A. (2014). The KULTURisk Regional Risk Assessment methodology for water-related natural hazards–Part 1: Physical–environmental assessment. *Hydrology and Earth System Sciences*, *18*(12), 5399-5414.

Sagi O., & Rokach L. (2018). Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *8*(4), e1249.

Saini, S. S., Kaushik, S. P., & Jangra, R. (2016). Flood-risk assessment in urban environment by geospatial approach: a case study of Ambala City, India. *Applied Geomatics*, *8*, 163-190.

Samanta, S., Pal, D. K., & Palsamanta, B. (2018). Flood susceptibility analysis through remote sensing, GIS and frequency ratio model. *Applied Water Science*, *8*(2), 66.

Samantaray S., Sahoo A., & Agnihotri A. (2023). Prediction of Flood Discharge Using Hybrid PSO-SVM Algorithm in Barak River Basin. *MethodsX, 10*, 102060.

Sajithra N., Ranyachitra D. (2021). Comparative Aanlysis of Various Tree Classifier Algorithms for Disease Datasets. *International Journal of Engineering Trends and Technology, 69*(6), 8-13.

Schmid-Breton, A., Kutschera, G., Botterhuis, T., & ICPR Expert Group 'Flood Risk Analysis'(EG HIRI). (2018). A novel method for evaluation of flood risk reduction strategies: explanation of ICPR FloRiAn GIS-tool and its first application to the rhine river Basin. *Geosciences*, *8*(10), 371.

Schlosser, A. D., Szabó, G., Bertalan, L., Varga, Z., Enyedi, P., & Szabó, S. (2020). Building extraction using orthophotos and dense point cloud derived from visual band aerial imagery based on machine learning and segmentation. *Remote Sensing*, *12*(15), 2397.

Scorzini, A. R., Radice, A., & Molinari, D. (2018). A new tool to estimate inundation depths by spatial interpolation (RAPIDE): Design, application and impact on quantitative assessment of flood damages. *Water*, *10*(12), 1805.

Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., & Shirzadi, A. (2018). Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *Journal of environmental management*, *217*, 1-11.

Shanmugam S, Preethi J (2019). Improved Feature Selection and Classification for Rheumatoid Arthritis Disease using a Weighted Decision Tree Approach. *J. Super Comput*, 75(8):5507–5519.

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39-42.

Smith, K., & Ward, R. (1998). *Floods: physical processes and human impacts* (pp. xii+-382).

Smith, D.I. (1981). Assessment of urban flood damage, In Proceedings of Flood Plain Management Conference, *Australian Water Council, Canberra, Australia*, 145–180.

Smith, D. I. (1994). Flood damage estimation-A review of urban stage-damage curves and loss functions. *Water Sa*, *20*(3), 231-238.

Sohail, M., & Muhammad, A. (2023, July). Assessment Of The 2022 Flood Disaster in Pakistan's Lower Indus Plain Using Sar And Optical Remote Sensing. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium* (pp. 2173-2176). IEEE.

Soliman, M. M., El Tahan, A. H. M., Taher, A. H., & Khadr, W. M. (2015). Hydrological analysis and flood mitigation at Wadi Hadramawt, Yemen. *Arabian Journal of Geosciences*, *8*(11), 10169-10180.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.

Tagar, H. K., & Shah, A. R. A. (2015). Sindh Forestry Resources: Causes of Deforestation and Policy Guideline for Its Conservation (A Case Study of Lower Indus Valley Sindh-Pakistan). *International Journal of Innovative Research and Development*.

Tarigan, A. P. M., Zevri, A., Iskandar, R., & Indrawan, I. (2017). A study on the estimation of flood damage in Medan city. In *MATEC web of conferences* (Vol. 138, p. 06010). EDP Sciences.

Tarigan, A. P. M., Hanie, M. Z., Khair, H., & Iskandar, R. (2018, March). Flood prediction, its risk and mitigation for the Babura River with GIS. In *IOP conference series: earth and environmental science* (Vol. 126, No. 1, p. 012119). IOP Publishing.

Tariq, A., Riaz, I., Ahmad, Z., Yang, B., Amin, M., Kausar, R., ... & Rafiq, M. (2020). Land surface temperature relation with normalized satellite indices for the estimation of spatio-temporal trends in temperature among various land use land cover classes of an arid Potohar region using Landsat data. *Environmental Earth Sciences*, *79*, 1-15.

Tariq, A., Mumtaz, F., Zeng, X., Baloch, M. Y. J., & Moazzam, M. F. U. (2022). Spatio-temporal variation of seasonal heat islands mapping of Pakistan during 2000–2019, using day-time and night-time land surface temperatures MODIS and meteorological stations data. *Remote sensing applications: Society and Environment*, *27*, 100779.

Tehrany M.S., Pradhan B., Jebur M.N. (2015). Flood Susceptibility Analysis and its Verification Using a Novel Ensemble Support Vector Machine and Frequency Ratio Method. *Stoch. Environ. Res. Risk Assess*, 29:1149–1165.

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2013). Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. *Journal of hydrology*, *504*, 69-79.

Tehrany, M. S., Lee, M. J., Pradhan, B., Jebur, M. N., & Lee, S. (2014). Flood susceptibility mapping using integrated bivariate and multivariate statistical models. *Environmental earth sciences*, *72*, 4001-4015.

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2014). Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *Journal of hydrology*, *512*, 332-343.

Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena*, *125*, 91-101.

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2015). Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stochastic environmental research and risk assessment*, *29*, 1149-1165.

Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena*, *125*, 91-101.

Termeh S.V.R., Kornejady A., Pourghasemi H.R., Keesstra S. (2018). Flood Susceptibility Mapping Using Novel Ensembles of Adaptive Neuro-Fuzzy Inference System and Metaheuristic Algorithms. *Sci. Total Environ.*, 615:438–451.

Tama B.A. and Rhee K.H. (2015). A Combination of PSO-based Feature Selection and Tree-Based Classifiers Ensemble for Intrusion Detection Systems. *Advances in Computer Science and Ubiquitous Computing*, vol.373, pp.489–495.

Tang J., Alelyani S., & Liu H. (2014). Feature Selection for Classification: A Review. *Data Classification: Algorithms and Applications*, 37.

Tien Bui D., Pradhan B., Nampak H., Bui Q.T., Tran Q.A., Nguyen Q.P. (2016a). Hybrid Artificial Intelligence Approach Based on Neural Fuzzy Inference Model and Metaheuristic Optimization for Flood Susceptibility Modeling in a High-Frequency Tropical Cyclone Area using GIS. *J Hydrol* 540:317–330.

Thomas, V. (2017). *Climate change and natural disasters: Transforming economies and policies for a sustainable future* (p. 158). Taylor & Francis.

Thompson, C. M., & Frazier, T. G. (2014). Deterministic and probabilistic flood modeling for contemporary and future coastal and inland precipitation inundation. *Applied Geography*, *50*, 1-14.

Tarigan, A. P. M., Zevri, A., Iskandar, R., & Indrawan, I. (2017). A study on the estimation of flood damage in Medan city. In *MATEC web of conferences* (Vol. 138, p. 06010). EDP Sciences.

Tarigan, A. P. M., Hanie, M. Z., Khair, H., & Iskandar, R. (2018, March). Flood prediction, its risk and mitigation for the Babura River with GIS. In *IOP conference series: earth and environmental science* (Vol. 126, No. 1, p. 012119). IOP Publishing.

Tien Bui, D., Pradhan, B., Lofman, O., & Revhaug, I. (2012). Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and Naive Bayes Models. *Mathematical problems in Engineering*, *2012*.

Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., & Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, *13*, 361-378.

Tien Bui, D., Ho, T. C., Pradhan, B., Pham, B. T., Nhu, V. H., & Revhaug, I. (2016). GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. *Environmental Earth Sciences*, *75*, 1-22.

Tien Bui, D., Pham, B. T., Nguyen, Q. P., & Hoang, N. D. (2016). Spatial prediction of rainfall-induced shallow landslides using hybrid integration approach of Least-Squares Support Vector

Machines and differential evolution optimization: a case study in Central Vietnam. *International Journal of Digital Earth*, *9*(11), 1077-1097.

Too J., Mirjalili S. (2021) A Hyper Learning Binary Dragonfly Algorithm for Feature Selection: A COVID-19 Case Study. *Knowl Based Syst,* 212:106553.

Trovato, M. R., & Giuffrida, S. (2018). The monetary measurement of flood damage and the valuation of the proactive policies in Sicily. *Geosciences*, *8*(4), 141.

Ullah, A., Suhail, M., & Ilyas, M. (2017). Modified method for choosing ridge parameter. *Journal of Statistics*, *24*(1), 20.

UNDP. (2022). Scaling-up of Glacial Lake Outburst Flood (GLOF) risk reduction in northern Pakistan United Nations development programme. Retrieved from https://www.undp.org/pakistan/projects/scaling-glacial-lake-outburst-flood-glof-risk-reduction-northern-pakistan

UNICEF. (2022). Devastating floods in Pakistan claim lives of more than 500 children. Retrieved from https://www.unicef.org/press-releases/devastating-floods-pakistan-claim-lives-more-500-children.

UNOCHA. (2022). In *Pakistan: 2022 monsoon floods humanitarian response snapshot (as of 13 September 2022)—Pakistan ReliefWeb*. Retrieved from https://reliefweb.int/report/pakistan/pakistan-2022-monsoon-floods-humanitarian-response-snapshot-13-september-2022.

Union, I. (2014). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. *Brussels. http://ww. w. xploit-eu. com/pdfs/Europe*, *202020*, 20.

UNSW, Evaluation Methodology of Flood Damage in Australia, Technical Project Report, 1981.

USACE. (1988). National Economic Development Procedures Manual, USA.

Ureta J.C., Zurqani H.A., Post C.J., Ureta J., Motallebi M. (2020). Application of Nonhydraulic Delineation Method of Flood Hazard Areas using Lidar-Based Data. *Geosciences 10 (9),* 338.

Vojtek, M., & Vojteková, J. (2016). GIS-based approach to estimate surface runoff in small catchments: A case study. *Quaestiones Geographicae*, *35*(3), 97-116.

VoPham, T., Hart, J. E., Laden, F., & Chiang, Y. Y. (2018). Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. *Environmental Health*, *17*, 1-6.

Vozinaki, A. E. K., Karatzas, G. P., Sibetheros, I. A., & Varouchakis, E. A. (2015). An agricultural flash flood loss estimation methodology: the case study of the Koiliaris basin (Greece), February 2003 flood. *Natural Hazards*, *79*, 899-920.

Wahla, S. S., Kazmi, J. H., Sharifi, A., Shirazi, S. A., Tariq, A., & Joyell Smith, H. (2022). Assessing spatio-temporal mapping and monitoring of climatic variability using SPEI and RF machine learning models. *Geocarto International*, *37*(27), 14963-14982.

Wang, S., Yan, Y., Yan, M., & Zhao, X. (2012). Quantitative estimation of the impact of precipitation and human activities on runoff change of the Huangfuchuan River Basin. *Journal of Geographical Sciences, 22*, 906-918.

Wang, S., Zhang, K., Chao, L., Li, D., Tian, X., Bao, H., ... & Xia, Y. (2021). Exploring the utility of radar and satellite-sensed precipitation and their dynamic bias correction for integrated prediction of flood and landslide hazards. *Journal of Hydrology*, *603*, 126964.

Wang, S., Jiang, L., & Li, C. (2015). Adapting naive Bayes tree for text classification. *Knowledge and Information Systems*, *44*, 77-89.

Wang, W. C., Chau, K. W., Xu, D. M., & Chen, X. Y. (2015). Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. *Water Resources Management*, *29*, 2655-2675.

Wang Y., Fang Z., Hong H., Costache R., & Tang X. (2021). Flood Susceptibility Mapping by Integrating Frequency Ratio and Index of Entropy with Multilayer Perceptron and Classification and Regression Tree. *Journal of Environmental Management*, *289*, 112449.

Waghwala, R. K., & Agnihotri, P. G. (2019). Assessing the impact index of urbanization index on urban flood risk. *International Journal of Recent Technology and Engineering*, *8*(2), 509-512.

Waqas, H., Lu, L., Tariq, A., Li, Q., Baqa, M. F., Xing, J., & Sajjad, A. (2021). Flash flood susceptibility assessment and zonation using an integrating analytic hierarchy process and frequency ratio model for the Chitral District, Khyber Pakhtunkhwa, Pakistan. *Water*, *13*(12), 1650.

Wu J., Liu H., Wei G., Song T., Zhang C., & Zhou H. (2019). Flash Flood Forecasting using Support Vector Regression Model in a Small Mountainous Catchment. *Water*, *11*(7), 1327.

Xie W, Wang L, Kun Yu, Shi T, Li W (2023) Improved Multi-Layer Binary Firefly Algorithm for Optimizing Feature Selection and Classification of Microarray Data. *Biomed Signal Process Control*, 79:104080.

Xu H., Yu S., Chen J., & Zuo X. (2018). An Improved Firefly Algorithm for Feature Selection in Classification. *Wireless Personal Communications*, *102*, 2823-2834.

Yang T., Asanjan A. A., Welles E., Gao X., Sorooshian S., & Liu X. (2017). Developing Reservoir Monthly Inflow Forecasts using Artificial Intelligence and Climate Phenomenon Information. *Water Resources Research*, *53*(4), 2786-2812.

Yaseen, A., Lu, J., & Chen, X. (2022). Flood susceptibility mapping in an arid region of Pakistan through ensemble machine learning model. *Stochastic Environmental Research and Risk Assessment*, *36*(10), 3041-3061.

Yokoyama R., Shirasawa M., & Pike R. J. (2002). Visualizing Topography by Openness: A New Application of Image Processing to Digital Elevation Models. *Photogrammetric Engineering and Remote Sensing*, *68*(3), 257-266.

Yu X., Liong S. Y., & Babovic V. (2004). EC-SVM Approach for Real-Time Hydrologic Forecasting. *Journal of Hydroinformatics*, *6*(3), 209-223.

Yamamuro, A. M., Rosi-Marshall, E. J., Lamberti, G. A., & Cordova, J. M. (2022). Quantity, controls and functions of large woody debris in midwestern USA streams.

Yue, Z., Zhou, W., & Li, T. (2021). Impact of the Indian Ocean dipole on evolution of the subsequent ENSO: Relative roles of dynamic and thermodynamic processes. *Journal of Climate*, *34*(9), 3591-3607.

Yi, C. S., Lee, J. H., & Shim, M. P. (2010). GIS-based distributed technique for assessing economic loss from flood damage: pre-feasibility study for the Anyang Stream Basin in Korea. *Natural hazards*, *55*, 251-272.

Yin, L., Wang, L., Keim, B. D., Konsoer, K., & Zheng, W. (2022). Wavelet analysis of dam injection and discharge in three gorges dam and reservoir with precipitation and river discharge. *Water*, *14*(4), 567.

Yin, L., Wang, L., Zheng, W., Ge, L., Tian, J., Liu, Y., ... & Liu, S. (2022). Evaluation of empirical atmospheric models using Swarm-C satellite data. *Atmosphere*, *13*(2), 294.

Youssef A.M., Pradhan B., Hassan A.M. (2011). Flash Flood Risk Estimation along the St. Katherine Road, Southern Sinai, Egypt Using GIS Based Morphometry and Satellite Imagery. *Environ. Earth Sci.*, 62:611–623.

Youssef, A. M., Pradhan, B., & Sefry, S. A. (2015). Remote sensing-based studies coupled with field data reveal urgent solutions to avert the risk of flash floods in the Wadi Qus (east of Jeddah) Kingdom of Saudi Arabia. *Natural Hazards*, *75*(2), 1465-1488.

Yu, C., Hall, J. W., Cheng, X., & Evans, E. P. (2013). Broad scale quantified flood risk analysis in the Taihu Basin, China. *Journal of Flood Risk Management*, *6*(1), 57-68.

Zhan, C., Dai, Z., Samper, J., Yin, S., Ershadnia, R., Zhang, X., ... & Soltanian, M. R. (2022). An integrated inversion framework for heterogeneous aquifer structure identification with single-sample generative adversarial network. *Journal of Hydrology*, *610*, 127844.

Zhang, K., Wang, S., Bao, H., & Zhao, X. (2019). Characteristics and influencing factors of rainfall-induced landslide and debris flow hazards in Shaanxi Province, China. *Natural hazards and earth system sciences*, *19*(1), 93-105.

Zhang, X., Ma, F., Yin, S., Wallace, C. D., Soltanian, M. R., Dai, Z., ... & Lü, X. (2021). Application of upscaling methods for fluid flow and mass transport in multi-scale heterogeneous media: A critical review. *Applied energy*, *303*, 117603.

Zhao, X., Li, D., Yang, B., Ma, C., Zhu, Y., & Chen, H. (2014). Feature selection based on improved ant colony optimization for online detection of foreign fiber in cotton. *Applied Soft Computing*, *24*, 585-596.

Zhou J., & Hua Z. (2022). A Correlation Guided Genetic Algorithm and its Application to Feature Selection. *Applied Soft Computing*, *123*, 108964.

Zhou, L., & Kang, L. (2023). A comparative analysis of multiple machine learning methods for flood routing in the Yangtze River. *Water*, *15*(8), 1556.

Zhu, Z., Wu, Y., & Liang, Z. (2022). Mining-induced stress and ground pressure behavior characteristics in mining a thick coal seam with hard roofs. *Frontiers in Earth Science*, *10*, 843191.

Zhu, J., & Pierskalla Jr, W. P. (2016). Applying a weighted random forests method to extract karst sinkholes from LiDAR data. *Journal of Hydrology*, *533*, 343-352.

Zúñiga, E., & Novelo-Casanova, D. A. (2019). Hydrological hazard estimation for the municipality of Yautepec de Zaragoza, Morelos, Mexico. *Hydrology*, *6*(3), 77.