On the Application of Machine Learning into Flood Modeling: Data Consideration and Modeling Algorithm

Ali Pourzangbar, Peter Oberle, Andreas Kron and Mário J. Franca

October 13, 2023

ON THE APPLICATION OF MACHINE LEARNING INTO FLOOD MODELING: DATA CONSIDERATION AND MODELING ALGORITHM

Ali Pourzangbar[1], Peter Oberle, Andreas Kron & Mário J. Franca
Institute for Water and River Basin Management, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
ali.pourzangbar@kit.edu

1. KEY WORDS

flood modeling, flood susceptibility mapping, machine learning, conditioning factors, flood catchment

2. ABSTRACT

This article reviews the literature on the application of Machine Learning (ML) to identify flood-prone areas, covering studies published since 2013. The review focuses on data considerations, such as the specifics of the study area and conditioning factors, as well as the ML algorithms used to identify flooding areas. 100 scientific articles were analyzed through a wide scope of geographical areas, ranging from arid to tropical climates and from small catchments to large river basins, to evaluate the influence of geographical features, historical flood occurrences, climatic patterns, urbanization, and data availability on flood susceptibility modeling (FSM). Iran, India, China, and Vietnam are the most frequently studied locations. The slope of the land, topographic wetness index, land use and land cover, rainfall levels and distance to rivers were key conditioning factors in at least 61% of the reviewed articles. Furthermore, the employed ML algorithms can be categorized into various types: statistical, kernel-based, tree-based, Neural Network (NN)-based, ensemble, and hybrid approaches. NN-based models, such as Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs), excel in solving high-dimensional problems but face challenges related to reliability and overfitting. Kernel-based approaches require optimal configuration through a trial-and-error process, while tree-based models offer simplicity and are less prone to overfitting, although they may be less precise. Among these, ensemble and hybrid models generally outperform traditional ML methods, despite their own limitations. These methods primarily focus on event-based historical floods, limiting their ability to make real-time predictions due to the lack of time-series data. Additionally, most models face restrictions given data consistency and validity. They often use inconsistent data, where flood conditions and input parameter values are not aligned in time and space. This discrepancy undermines the models' reliability. Consistent and valid datasets are imperative for accurate model development.

3. INTRODUCTION

Floods cause enormous damage, mortality, and substantial adverse impacts on public health and the economy. Statistics indicate that floods have been a major contributor to global damage from natural disasters, accounting for over half of such damage in the past fifty years [1]. Between 2017 and 2022, floods affected about 244 million people globally; in 2022 alone, there were 7398 flood-related deaths, which is the highest number recorded in the last five years [2]. Available evidence indicates that flood related damage will increase in the future [3-4] both given the intensity and the frequency of flood events [5]. Recent flash floods have had significant socio-economic impacts in various countries. Notable examples include the floods in Germany in July 2021 [6-8]. Thus, accurate prediction of flood-prone areas is crucial for emergency response, life-saving measures, and the development of early warning systems for evacuation strategies.

In the early twentieth century, flood modeling was primarily based on empirical models, e.g. weight of evidence (WOE), analytical hierarchy process (AHP), and frequency ratio (FR). These models were based on observed data and used simple statistical methods to predict flood events [9]. As technology progressed, hydrological models such as the ones produced by the the USACE Hydrologic Engineering Center, SWAT, and HSPF were developed. These models consider various factors like rainfall, soil type, and land use to predict flood events [10]. Afterwards, in the later part of the twentieth century, hydraulic models became popular. These models simulate the flow of water through the river channel networks and floodplains, providing more detailed flood predictions. For example, the widely used HEC-RAS, a software that uses hydraulic principles to model the flow of water in rivers and channels, is a hydraulic-based model numerically models the flood propagation [11]. In the early twentieth-first century, integrated models that combine hydrological and hydraulic

[1] Corresponding author

models were developed. Models such as TELEMAC, DELFT3D, MIKE 11 provide a comprehensive approach to flood modeling [12]. However, using purely physics-based models (hydraulic- or hydrological-based models) necessitated a considerable amount of data, tedious parameterization, and having a detailed comprehension of flood-related parameters, bring extra complecity to set them up and use.

Knowledge-driven approaches (e.g. empirical models) are efficient, but only in areas where enough data is available. For instance, the effectiveness of the FR method is greatly influenced by the size of the sample used. To address these problems, ML models, which are grounded in artificial intelligence (AI), and statistical models are being used increasingly. ML models present certain shortcomings, such as susceptibility to overtraining, variable reliability, limited generalization, and the potential to produce incorrect results. This is often because they heavily depend on the input datasets and specified parameters. Thus, utilizing unrelated or irrelevant data and parameters can lead to inaccurate modeling. Despite these challenges, the simplicity, speed of execution, and reasonable accuracy of these models have garnered significant interest from the research community. They are capable of modeling complex, non-linear phenomena without requiring an understanding of the underlying mechanisms [13]. Nevertheless, they generally lack the ability to integrate the fundamental physics of the problems they are attempting to solve, barring some recent advancements in physics-based ML models, which fall beyond the scope of this article. To mitigate this limitation, it has been demonstrated that the effectiveness of ML models can be significantly enhanced by integrating them with other modeling approaches, such as metaheuristic strategies and optimization algorithms, numerical analyses, and physical models [14-15]. Figure 1 illustrates a simplified evolution of flood modeling.

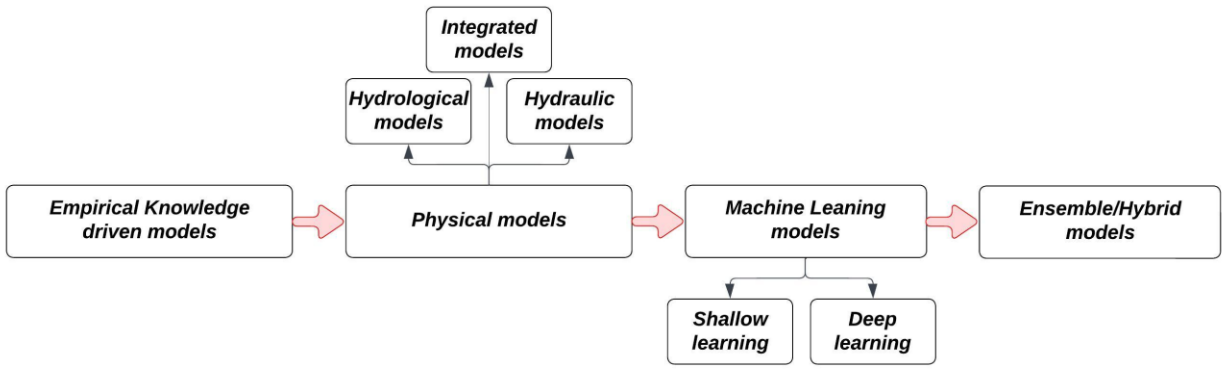

Figure 1: A simplified timeline illustrating the evolution of flood modeling

The current review article discusses the application of ML into flood susceptibility modeling (FSM), focusing on various aspects such as consideration of catchment features, conditioning factors or input parameters, and configuration of the developed ML models. We also discuss the synergy of ML models with metaheuristic algorithms, the development of ensemble models, and the evaluation of these newly developed models. Moreover, this article highlights some topics that have either been overlooked or not sufficiently explored in the existing literature.

4. SCOPE

Between 1970-2023, a total number of 1749 articles have been published and indexed in various databases such as Web of Science, Google Scholar, and Scopus, among others, using keywords like "Machine learning" AND "flood susceptibility mapping" OR "flood susceptibility" OR "flood susceptibility assessment" OR "flood prediction". However, there is a clear trend indicating an increase in the number of published papers employing machine learning methods for flood prediction after 2013, as illustrated in Figure 2 (left pannel). These articles have been published in international journals affiliated with esteemed publishing entities including Elsevier, Springer Nature, MDPI, Taylor & Francis, Wiley, Copernicus Gesellschaft MbH, and Frontiers Media SA. In Figure 2 (right panel), the names of various journals are displayed along with the percentage of papers they publish relative to the total number of articles (1749 articles). The figure only includes journals with a contribution of more than 1 percent. For example, "Natural Hazards" accounts for 105 articles, comprising 5.41% of the total record count, while "Journal of Hydrology" contributed 71 articles, constituting 3.66%.

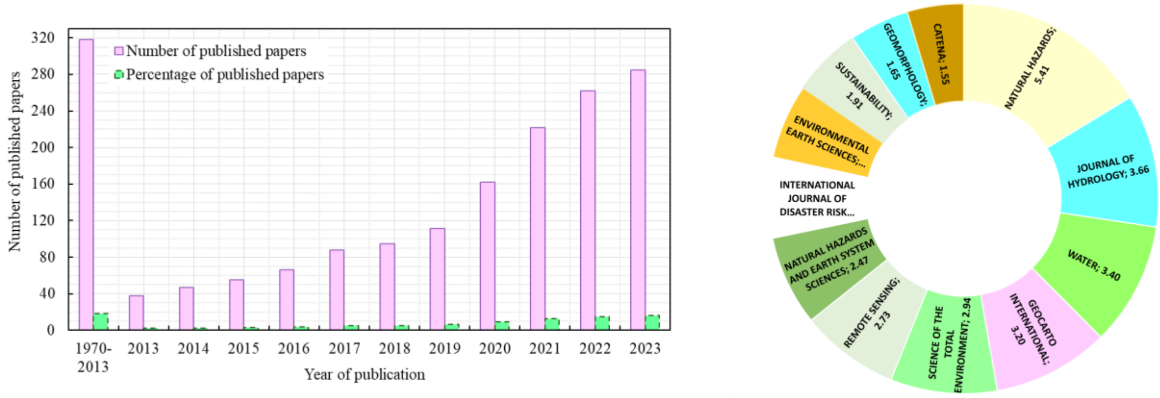

Figure 2: Left Panel: published articles on FSM using ML models in different years. Right Panel: journals with over 1% of published articles on FSM using ML models from 2013 to 2023

A subset of 100 articles published after 2013 was chosen for further investigation, each showcasing innovation in various facets such as data acquisition and preprocessing, input parameters selection, as well as model development and optimization. Figure 3 illustrates the distribution of selected articles for this review across different years. As shown, in the last five years, the number of papers that have studied FSM using ML has significantly increased compared to previous years.

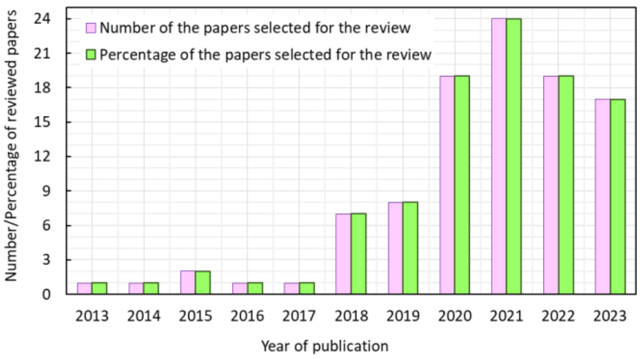

Figure 3: Number/percentage of the articles dealing with FSM using ML models selected in this review

5. STUDY AREA (CATCHMENT) CONSIDERATIONS

The selection of a study area is often a multi-faceted decision, impacted by geographical, climatic, human, geopolitical, historical, environmental conditions and geological factors. Each area has unique features that make it more or less suitable for FSM implementation. The study areas of the reviewed papers seem to cover various climate zones from arid regions in Saudi Arabia and Egypt to tropical climates like Vietnam and Malaysia to temperate zones like Canada and Switzerland. The areas range from small local catchments (e.g., The City of Carlisle, UK with 14.5 km^2) to extensive river basins (e.g., The Brahmaputra River Basin, Bangladesh with 583000 km^2). Some areas are densely populated cities like Seoul, while others are rural districts, e.g., Shangla District in Pakistan. Figure 4 illustrates the study aareas covered in the selected reviewed papers. Iran, India, China, and Vietnam are the most frequently selected locations for case studies on FSM development.

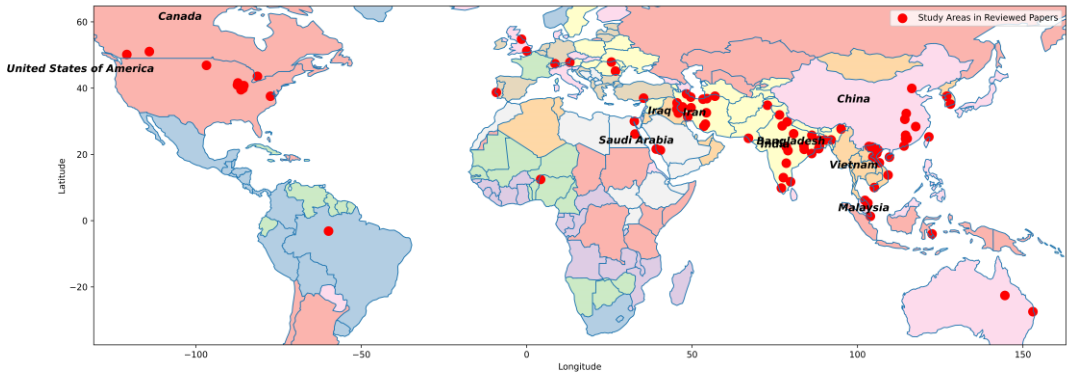

Figure 4: The study areas in the selected papers for the review

By evaluating various study areas worldwide, we can draw insights into the essential parameters and conditions related to the study are that determine the suitability for application of ML into FSM.

- ***Geographical and topological conditions:*** studies often target areas with diverse geography and terrain including variation in elevation, slope, and the presence of different landforms like mountains, rivers, and plains. Such complexities can profoundly affect the flood dynamics and susceptibility of the region. For instance, Maneh and Samalqan City [1] have a diverse terrain, with elevations ranging from 314 to 2,785 m, and the Haraz watershed [16] is characterized by mountains, hills, rivers, and streams, and elevations ranging from 300 m to 5595 m.

- ***Historical Occurrence of Floods or Natural Disasters:*** the selected areas usually have experienced past flooding or other natural disasters. Such historical data are crucial for calibration and validation of flood susceptibility models. Al-Areeq et al. [17] discusses how Jeddah city suffered from flash floods in 2009 and 2011, causing substantial damage. Mahdizadeh Gharakhanlou & Perez [18] specifically looks at the floods of May 2017 in QC and November 2021 in BC, making both watersheds critical for understanding the risk factors leading to such events.

- ***Climatic Patterns and Seasonal Variation:*** locations with a history of variable climatic conditions such as seasonal heavy rainfall or extreme weather events offer important insights into how climate factors into flood risks. Ullah et al. [19] details the heavy rainfall and snowfall in the Hindu Kush region during the summer monsoons, making it an area of interest for studying the impact of climate on flood susceptibility. Quang Binh in Vietnam Faces frequent flooding and storms, notably recorded in 2007, 2010, and 2016, highlighting the need for susceptibility mapping in this climatically vulnerable area [20].

- ***Urbanization and Land Use:*** areas undergoing rapid changes in land use or that have high population densities provide opportunities to explore how human activity correlates with flood risks. Two different regions, Quebec (QC) and British Columbia (BC) in Canada, have been impacted by land use changes, making them suitable for studying the human impact on flood susceptibility [18].

- ***Data Availability:*** adequate and reliable data sources are crucial for modeling and validation of flood susceptibility maps. For instance, in South Korea Coastal Climate [21] 68 weather observatories and 46 tide observatories provide extensive data. Similarly, in Quang Binh, Vietnam [20] data on extreme flood events in 2007, 2010, and 2016 are available from the Quang Binh Centre for Hydrometeorological Forecasting.

6. CONDITIONING FACTORS

The ML models predict flood events are based on the relationship between the dependent variable (flood suceptibility) and independent variables (conditioning factors or input parameters). Each of these criteria was prepared in the form of raster maps with different spatial resolutions, usually the 30 m × 30 m pixel sizes. In the reviewed articles, the principal output is generally flood susceptibility (either binary or probabilistic). Various methodologies exist for generating reference flood maps, such as the interpretation of digital satellite imagery or the utilization of historical flood databases. Though the reviewed papers employed diverse combinations of input parameters for model development, these can be classified into four distinct categories: topographical, hydrological, environmental, and morphological features.

Topographical features like elevation (altitude), and slope relate to the shape, elevation, and layout of the land surface. They serve as initial indicators for predicting the flow direction of water, thereby identifying flood-prone regions. Distances to rivers, streams, and canals are vital, as closer proximity often increases flood risk, while metrics like Digital Terrain Model, curvature, and topographic wetness index provide insights into the terrain water-holding capacity.

Hydrological parameters such as flow rate, flow accumulation, and flow direction reveal the water travel path during heavy rainfall, shaping the understanding of which areas are at imminent risk. The significance of rainfall, precipitation, and soil conditions (like soil hydrological groups, soil depth, and type) can not be overstated, as they define the land ability to absorb water before reaching a saturation point that triggers flooding.

Environmental parameters encompass natural and man-made factors affecting the flood, such as land use, vegetation, and climate. For instance, the land use and urban areas directly affect the permeability of the ground, thereby influencing surface runoff and altering flood patterns. Additionally, natural factors like wind and evapotranspiration rates can either accelerate or decelerate the water cycle inputs and outputs.

Morphological parameters describe the form and structure of channels and basins, like the depth and shape of riverbeds.

These features such as aspect, compactness coefficient, and elongation ratio offer a lens through which the land capability to channel water can be assessed. Geological structures may act as natural barriers or conduits for water flow. Parameters like stream density, drainage density, and bifuration ratio provide insights into how efficiently an area can channel away water. Lastly, land roughness metrics (comprising roughness, texture ratio, and terrain roughness index) help to gauge the friction offered by the land to the flow of water, which can serve as a vital clue for predicting the speed of flood progression.

These parameters collectively provide a robust framework for understanding the intricacies involved in flood modeling, enabling us to develop more accurate and actionable predictive models. Table 1 summarizes all the input parameters used in the reviewed papers to develop ML models and their contribution in the flood susceptibility.

Category	Parameter	Parameter contribution to flooding
Topographical	Slope	Steeper slopes lead to rapid runoff, potentially causing flash floods, while gentler slopes cause water to accumulate [22].
	Elevation	Higher elevations are less prone to flooding, while lower elevations are likely to collect water, making them more susceptible to floods [23].
	Terrain Roughness Index (TRI)	TRI is a measure of the local variability in elevation within a landscape [18].
	Topographic Position Index (TPI)	It is a measure of the relative elevation of a point within a landscape [16].
	Digital Terrain Model (DTM)	It represents the topography of the Earth surface as a grid of elevation values [1-5].
	Curvature	Indicates how flat or curved a land slope is, it influences the speed of flows over the landscape, potentially leading to localized flooding [24].
Hydrological	Topographic Wetness Index (TWI)	It measures the degree to which a landscape is saturated with water [16&25].
	Rainfall	Excessive rainfall in a short period can overwhelm drainage systems and natural barriers, leading to immediate runoff and flooding [21].
	River (stream) Density (Riv-Den)	Higher river density can lead to a more complex network of water paths, potentially accelerating flood spread.
	Drainage Density (Drain-Den)	Increased drainage density facilitates quicker runoff, possibly heightening flood risk.
	Flow Accumulation (FloAcc)	Represents the number of cells that flow into each cell in a grid, helping to identify where water will accumulate and potentially flood [1&17].
	Flow Direction (FlowDir)	Indicates the direction in which water will flow across a landscape, helping to predict the path and extent of potential flooding [26-27].
Environmental	Land Use Land Cover (LULC)	Urban or agricultural land uses can create impermeable surfaces that lead to rapid runoff, thereby increasing flood risk [1].
	Distance to river (DisRiv)	Proximity to a river can signify a higher risk of riverine flooding, especially if the river is prone to overflow its banks [20].
	Lithology	The type of rock or sediment that makes up the ground surface can affect water absorption and runoff [16-17].
	Soil type	Different soil types have varying capacities to absorb water, and less absorbent soils can contribute to faster runoff and flooding [26-27].
	Normalized Difference Vegetation Index (NDVI)	It is a criterion of the health and density of vegetation that can act as a natural barrier against floods by absorbing water [24].
	Distance to the road (DisRoa)	Roads can act as barriers or channels for water flow, affecting how water accumulates and disperses during a flood [26].
	Distance to a fault (DisFau)	Fault lines can sometimes lead to changes in the terrain or water table, indirectly affecting flood patterns [1&19].
Morphological	Stream Power Index (SPI)	SPI relates to the erosion power of floods and has a direct relationship with slope angle and watershed area [18-19].
	Plan Curvature	It affects the horizontal curvature of the land surface, influencing the direction and distribution of water flow during flooding.
	Aspect	It impacts how water flows over it, as well as other factors like solar radiation and evapotranspiration that influence hydrological processes.
	Sediment Transport Index (STI)	Measures how sediment is carried by water flow, which can either reduce or enhance flood risk depending on the landscape [28].

Table 1: The categories and common indicatores of selected input parameters used to FSM in the reviewed papers.

In our review of 100 research articles focusing on machine learning models for flood prediction, several input parameters consistently emerged as crucial for effective modeling. Specifically, the topographic wetness index (TWI), slope of the land (Slope), types of land use and land cover (LULC), rainfall levels, and the distance to the nearest river (DisRiv) were identified as the most commonly used conditioning factors. Remarkably, DisRiv was noted as a significant factor in 61% of the examined articles, while TWI was highlighted in 68%. These findings underscore the prevailing importance of these variables in the field of flood modeling. Figure 5 illustrates the frequency of various parameters that contributed to the development of ML models in the reviewed articles. It is worth noting that only parameters reported here which used more than five times in the reviewed articles (i.e. they are used at least in five papers for the modeling procedure).

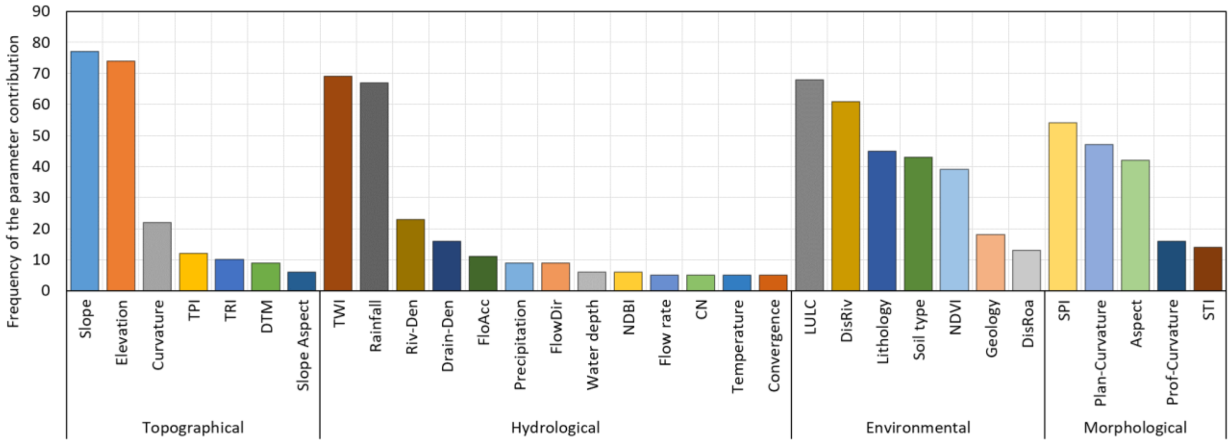

Figure 5: Frequency of the input parameters used in the reviewed paper to model flooding (only parameters are reported here whose contribution in the reviewed articles is more than 5%).

7. ML MODELS EMPLOYED FOR FLOOD SUSCEPTIBILITY MAPPING

The reviewed articles employed a diverse array of machine learning configurations and statistical methods for predicting flood susceptibility. The ML models encompasses: (1) NN-based models that mimic biological neural networks are capable of learning from data for tasks such as classification and regression [29-30]. An example of this is artificial neural networks (ANN); (2) Kernel-based models that utilize kernel functions to map input data into a higher-dimensional space, aiding linear algorithms in solving non-linear problems such as support vector machines (SVMs) [31]; (3) Tree-based models such as M5' model tree and alternating decision tree (ADtree) that employ decision tree structures to make predictions through a series of binary "if-then" decision thresholds, suitable for both regression and classification tasks [32-33]; (4) Ensemble models such as random forest (RF) and rotation forest (ROF) that combine multiple base predictive models to improve overall accuracy and robustness of predictions [34]; (5) Hybrid models such as adaptive fuzzy neural inference systems (ANFIS) that merge different types of models or model architectures to leverage strengths of individual models, enhancing performance and explainability [35]. Statistical models are mathematical models embodying statistical assumptions about data generation to analyze relationships between variables and make inferences about a population based on sample data [36]. The implemented statistical models for flood prediction can be categoried into regression-based such as logistic regression (LR), multi-criteria decision analysis (MCDA) such as FR, bayesian such as Naïve Bayes (NB), instance-based like K-nearest neighbor (KNN), among others. Figure 6 illustrates the diverse ML configurations and statistical approaches utilized in the selected articles for the review.

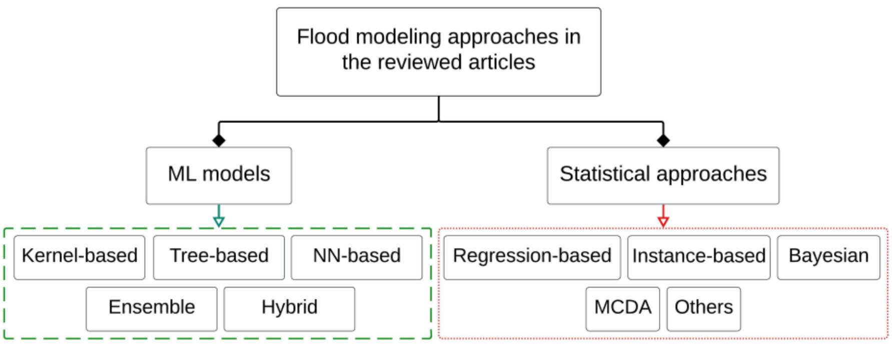

Figure 6: Different ML and statistical models used for FSM in the reviewed literature

Table 2 summarizes various ML and statistical model types, each accompanied by examples; additionally, it reports the frequency with which these configurations and models are cited in the reviewed papers.
Ensemble models are particularly prevalent, appearing 107 times in various papers. These models can be categorized into Bagging-based ensembles (Bagging-Ens), Boosting-based ensembles (Boosting-Ens), and Forest-based ensembles (Forest-Ens). Bagging-Ens take multiple random samples from the data and build a separate model for each before averaging the predictions. Boosting-Ens, including AdaBoost, boosted generalized linear model (GLM), and gradient boosting machine (GBM), build an initial model, identify its errors, and then create another model to correct those errors. Forest-Ens like Random Forest (RF) and Rotation Forest (RoF) are built on forest algorithms. Some models do not neatly fit into these categories and include methods like statistical ensembles and SGD-WOE (Stochastic Gradient Descending-Weights of Evidence).

Hybrid models are also frequently used, 102 times in the reviewed articles. These models have several subcategories including Fuzzy Logic-based hybrids (FL-Hyb), Tree-based hybrids (Tree-Hyb), Neural Network-based hybrids (NN-Hyb), Weighted Average-based hybrids (WAve-Hyb), and kernel-based hybrids (Kernel-Hyb). FL-Hyb models rely predominantly on fuzzy logic (FL) but also incorporate other ML methods such as neural networks (e.g., FART (Fuzzy Adaptive Resonance Theory) & FL-NN (Fuzzy Logic-Neural Network)), ensemble methods (e.g., FL-RF (Fuzzy Logic-Random Forest)), and evolutionary algorithms (EA) to improve both accuracy and robustness. Tree-Hyb models primarily use a tree-based model but are integrated with statistical methods like Naïve Bayes Tree (NBT) and evolutionary algorithms such as combined Random Forest and Genetic Algorithm (RF-GA). NN-Hyb models use a neural network as their core component but implement their training algorithms using fuzzy logic, ensemble methods, or evolutionary algorithms like DNN-AO (Deep Neural Network-Aquila Optimizer) and ELM-PSO (Extreme Learning Machine-Particle Swarm Optimization). WAve-Hyb models statistically integrate the results from various ML methods to produce a final outcome, often using weighted averages. For example, in the case of RF-SVM (Random Forest-Support Vector Machine), the results of RF and SVM are separately produced and then combined using weighted averages. Kernel-Hyb models use a kernel-based model like SVM at their core and are integrated with statistical methods such as combined AdaBoost and radial base function (RBF), ensemble methods like Bagging-RBF, and evolutionary algorithms such as support vector regression (SVR) optimized by bat algorithm (BA).

NN-based models (employed 62 times) can be categorized into the following types based on different configurations and applicabilities. Shallow learning NNs (Shallow-NN) models like ANN are generally used for tasks like classification, regression, and pattern recognition. Grid-based neural networks (Grid-NN) like convolutional neural networks (CNN) are specialized for processing grid-structured data like images. Sequential neural networks (Sequential-NN), such as LSTM, a type of RNN, are capable of learning long-term dependencies and are useful for sequential data like time series or natural language processing. Deep neural networks (Deep-NN) have multiple layers between the input and output, enabling the learning of more complex data representations.

Tree based models have been employed 36 times in the selected papers and range from basic to advanced models. Basic models (Basic-DT) are primarily used for classification tasks and include methods like decision trees (DT) and classification and regression tree (CART). Advanced models (Advanced-DT) incorporate specialized, rule-based methods such as Logistic Model Trees (LMT) and reduced error pruning trees (REPT).

Kernel based models have been employed 31 times in the reviewed articles and are divided into three major categories: SVM-based models (Kern-SVM), Kernel-based regression models (Kern-Reg), and Kernel-based classification models (Kern-Class). SVM-based models, e.g. weakly labeled support vector machine (WELLSVM) and SVM-RBF, are used for tasks like classification and regression. They are particularly effective when the dataset is not fully labeled or when it is expensive to obtain such labels.

Statistical models have also been employed frequently, appearing 74 times across different papers. These range from regression models used to predict a numerical value based on one or more variables (e.g., Linear Regression or LR), instance-based models like K-Nearest Neighbors (K-NN) that make predictions based on similarity to known examples, multi-criteria decision analysis (MCDA) such as FR that evaluates and compares alternatives based on multiple criteria to arrive at the most favorable option, and Bayesian methods such as NB that use Bayes' theorem to update probabilities based on new evidence. Some models do not fit neatly into any of these categories, for example, Generalized Linear Models (GLM) or Maximum Entropy models.

Configiration (No. of usage)	Model type	Frequency of utilize	Example(s)
Ensemble (107)	Forest-Ens	38	RF, RoF
	Boosting-Ens	37	AdaBoost, Boosting GBM, BRT
	Others	22	EMca, EMmean, RS-GAM, RS-MARS
	Bagging-Ens	10	Bagging Ensembles
Hybrid (102)	FL-Hyb	34	FART, FL-NN, FL-RF, FL-EA, FL-NN-EA
	Tree-Hyb	21	NBT, RF-GA
	NN-Hyb	17	DNN-AO, ELM-PSO
	WAve-Hyb	17	RF-SVM
	Kernel-Hyb	13	AdaBoost-RBF, Bagging-RBF, SVR-BA
Statistical (74)	Regression	23	LR, MARS
	Instance-based	17	KNN
	MCDA	17	FR, AHP
	Others	11	Maximum Entropy
	Bayesian	6	NB
NN-based (62)	Shallow-NN	34	ANN
	Grid-NN	13	CNN
	Sequential-NN	11	LSTM/RNN
	Deep-NN	4	DNN
Tree-based (36)	Basic-DT	20	Classification Tree, J48 DT, CART
	Advanced-DT	16	LMT, REPT, FT, CDT
Kernel-based (31)	Kern-SVM	26	SVM, SVM-RBF, K-SVM, WELLSVM
	Kern-Reg	4	SVR
	Kern-Class	1	SVC

Table 2: various ML and statistical model types employed in the reviewed papers, each accompanied by examples (the new abbreviation are BRT: boosted regression tree; EMca: ensemble model committee averaging; EMmean: Ensemble Model to estimate the mean; RS-GAM: random subsampling-generalized additive model; RS-MARS: random subsampling-multivariate adaptive regression splines; MARS: multivariate adaptive regression splines; CDT: credal decision tree; K-SVM: Kernel support vector machine; SVC: support vector classification)

8. SUMMARY AND CONCLUSIONS

In the article, 100 papers focusing on flood modeling through machine learning methods are reviewed. In the reviewed papers, various criteria are thoroughly examined. These criteria encompass the selection of the study area, the parameters that contribute to the modeling process, as well as the type and configuration of the models used for flood modeling.

In the selection of catchment areas in which flood modeling using machine learning is possible, it is crucial to choose areas that offer a wide range of values for influential parameters. For example, should an area with limited variability in factors like elevation or slope be chosen, the machine learning algorithm could incorrectly treat these parameters as constants. This would result in an physically unsound model that neglects the contributions of certain parameters, affecting its final robustness and reliability. Selecting expansive enough area ot only improves the model accuracy but also broadens its applicability to various future scenarios. Additionally, it is important for the chosen area to have a reliable dataset for model development. Incomplete or erroneous datasets can lead to inaccurate models. Finally, to ensure the models account for extreme events, it is advisable to select areas that have a history of severe flooding. This will make the models more applicable to future extreme events.

An analysis of the input parameters for the ML revealed that the most relevant conditioning factors used in the develoment of ML models are slope and elevation representing topographical parameters, TWI and rainfall representing hydrological features, land use/cover and distance to the nearest river representing environmental parameters, and finally stream power index and plane curvature representing morphological parameters .

The trend in using ML models for flood modeling has significantly increased recently, particularly focusing on ensemble and hybrid models. The reviewed papers featured a wide range of ML approaches, from neural network-based models like ANNs to kernel-based solutions like SVMs, as well as tree-based and ensemble methods like random forests. Statistical models also varied, covering regression-based, multi-criteria, Bayesian, and instance-based techniques such as logistic regression and K-nearest neighbor. This diversity in ML and statistical methods highlights the evolving complexity and capability in flood susceptibility prediction.

The discussion in this article will be further developed to encompass the types of ML implemented models, the development flowchart for ML models, methods for data collection and preprocessing (selection of train, validation and test datasets), and input parameters selection using statistical and ML methods. Furthere, the performance of these models to guide the selection of the most appropriate approaches, and ultimately focus on optimizing model hyperparameters for improved accuracy and efficiency, will also be assessed in the future.

9. REFERENCES

[1] Eslaminezhad, S.A., Eftekhari, M., Azma, A., Kiyanfar, R., & Akbari, M. (2022). Assessment of flood susceptibility prediction based on optimized tree-based machine learning models. Journal of Water and Climate Change, 13. https://doi.org/10.2166/wcc.2022.435

[2] Ritchie, H., Rosado, P., & Roser, M. (2022). Natural Disasters. https://ourworldindata.org/natural-disasters (Accessed on 11.09.2023).

[3] Tellman, B., Sullivan, J.A., Kuhn, C., et al. (2021). Satellite imaging reveals increased proportion of population exposed to floods. Nature, 596(7870), 80-86. https://doi.org/10.1038/s41586-021-03695-w

[4] Simpson, N.P., Mach, K.J., Constable, A., et al. (2021). A framework for complex climate change risk assessment. One Earth, 4(4), 489-501. https://doi.org/10.1016/j.oneear.2021.03.005

[5] Deroliya, P., Ghosh, M., Mohanty, M.P., Ghosh, S., Durga Rao, K.H.V., & Karmakar, S. (2022). A novel flood risk mapping approach with machine learning considering geomorphic and socio-economic vulnerability dimensions. Science of The Total Environment, 851(1), 158002. https://doi.org/10.1016/j.scitotenv.2022.158002

[6] Mohr, S., Ehret, U., Kunz, M., Ludwig, P., Caldas-Alvarez, A., Daniell, J.E., Ehmele, F., et al. (2022). A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe. Part 1: Event description and analysis. Natural Hazards and Earth System Sciences Discussions, 2022, 1-44. https://doi.org/10.5194/nhess-23-525-2023

[7] Ludwig, P., Ehmele, F., Franca, M.J., Mohr, S., Caldas-Alvarez, A., Daniell, J.E., Ehret, U., et al. (2023). A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe–Part 2: Historical context and relation to climate change. Natural Hazards and Earth System Sciences, 23(4), 1287-1311. https://doi.org/https://doi.org/10.5194/nhess-23-1287-2023

[8] Piper, D., Kunz, M., Ehmele, F., Mohr, S., Mühr, B., Kron, A., & Daniell, J. (2016). Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany – Part 1: Meteorological background. Natural Hazards and Earth System Sciences, 16, 2835-xxxx. https://doi.org/10.5194/nhess-16-2835-2016

[9] Chow, V. T., Maidment, D. R., & Mays, L. W. (1988). Applied Hydrology. McGraw-Hill.

[10] Crawford, N. H., & Linsley, R. K. (1966). Digital simulation in hydrology: Stanford Watershed Model IV. Technical report, Stanford University.

[11] Brunner, G. W. (2016). HEC-RAS, River Analysis System Hydraulic Reference Manual. US Army Corps of Engineers.

[12] DHI Water & Environment. (2009). MIKE 11 - A modeling system for rivers and channels. DHI Water & Environment.

[13] Rajaee, T. (2011). Wavelet and ANN combination model for prediction of daily suspended sediment load in rivers. Science of The Total Environment, 409, 2917-2928. https://doi.org/10.1016/j.scitotenv.2010.11.028

[14] Mosavi, A., Ozturk, P., & Chau, K.W. (2018). Flood prediction using machine learning models: Literature review. Water (Switzerland), 10(11). https://doi.org/10.3390/w10111536

[15] Razavi Termeh, S.V., Kornejady, A., Pourghasemi, H.R., & Keesstra, S. (2018). Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. Science of The Total Environment, 615, 438-451. https://doi.org/10.1016/J.SCITOTENV.2017.09.262

[16] 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. https://doi.org/10.1016/j.envsoft.2017.06.012

[17] Al-Areeq, A.M., Abba, S.I., Yassin, M.A., Benaafi, M., Ghaleb, M., & Aljundi, I.H. (2022). Computational Machine Learning Approach for Flood Susceptibility Assessment Integrated with Remote Sensing and GIS Techniques from Jeddah, Saudi Arabia. Remote Sensing, 14, 5515. https://doi.org/10.3390/rs14215515

[18] Mahdizadeh Gharakhanlou, N., & Perez, L. (2022). Spatial Prediction of Current and Future Flood Susceptibility: Examining the Implications of Changing Climates on Flood Susceptibility Using Machine Learning Models. Entropy, 24, 1630. https://doi.org/10.3390/e24111630

[19] Ullah, K., Wang, Y., Fang, Z., Wang, L., & Rahman, M. (2022). Multi-hazard susceptibility mapping based on Convolutional Neural Networks. Geoscience Frontiers, 13(5), 101425. https://doi.org/10.1016/j.gsf.2022.101425

[20] Luu, C., Bui, Q.D., Costache, R., Nguyen, L.T., Nguyen, T.T., Phong, T.V., Le, H.V., & Pham, B.T. (2021). Flood-prone area mapping using machine learning techniques: a case study of Quang Binh province, Vietnam. Natural Hazards, 108, 3229–3251. https://doi.org/10.1007/s11069-021-04821-7

[21] Park, S-J., & Lee, D-K. (2020). Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms. Environmental Research Letters, 15(9), 094052. https://doi.org/10.1088/1748-9326/aba5b3

[22] Abu El-Magd, S., Maged, A., & Farhat, H. I. (2022). Hybrid-based Bayesian algorithm and hydrologic indices for flash flood vulnerability assessment in coastal regions: machine learning, risk prediction, and environmental impact. Environmental Science and Pollution Research, 29, 57345–57356. https://doi.org/10.1007/s11356-022-19903-7

[23] Luu, C., Von Meding, J., & Kanjanabootra, S. (2018). Assessing flood hazard using flood marks and analytic hierarchy process approach: a case study for the 2013 flood event in Quang Nam, Vietnam. Natural Hazards, 90, 1031–1050. https://doi.org/10.1007/s11069-017-3083-0

[24] Ha, H., Bui, Q.D., Nguyen, H.D., Pham, B.T., Lai, T.D., & Luu, C. (2023). A practical approach to flood hazard, vulnerability, and risk assessing and mapping for Quang Binh province, Vietnam. Environment, Development and Sustainability, 25, 1101–1130. https://doi.org/10.1007/s10668-021-02041-4

[25] Bui, Q.-T., Nguyen, Q.-H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., & Pham, V.-M. (2020). Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. Journal of Hydrology, 581, 124379. https://doi.org/10.1016/j.jhydrol.2019.124379

[26] Nguyen, H.D., Van, C.P., & Do, A.D. (2023). Application of hybrid model-based deep learning and swarm-based optimizers for flood susceptibility prediction in Binh Dinh province, Vietnam. Earth Science Informatics, 16, 1173–1193. https://doi.org/10.1007/s12145-023-00954-4

[27] Aldiansyah, S., & Wardani, F. (2023). Evaluation of flood susceptibility prediction based on a resampling method using machine learning. Journal of Water and Climate Change, 14(3), 937–961. https://doi.org/10.2166/wcc.2023.494

[28] Wang, Y., Hong, H., Chen, W., Li, S., Panahi, M., Khosravi, K., Shirzadi, A., Shahabi, H., Panahi, S., & Costache, R. (2019). Flood susceptibility mapping in Dingnan County (China) using adaptive neuro-fuzzy inference system with biogeography based optimization and imperialistic competitive algorithm. Journal of Environmental Management, 247, 712–729. https://doi.org/10.1016/j.jenvman.2019.06.102

[29] Pourzangbar, A., Losada, M.A., Saber, A., Rasoul Ahari, L., Larroudé, P., Vaezi, M., & Brocchini, M. (2017). Prediction of non-breaking wave induced scour depth at the trunk section of breakwaters using Genetic Programming and Artificial Neural Networks. Coastal Engineering, 121, 107–118. https://doi.org/10.1016/j.coastaleng.2016.12.008

[30] Pourzangbar, A., Saber, A., Yeganeh-Bakhtiary, A., & Rasoul Ahari, L. (2017). Predicting scour depth at seawalls using GP and ANNs. Journal of Hydroinformatics, 19(3), 349–363. https://doi.org/10.2166/hydro.2017.125

[31] Pourzangbar, A., Brocchini, M., Saber, A., Mahjoobi, J., Mirzaaghasi, M., & Barzegar, M. (2017). Prediction of scour depth at breakwaters due to non-breaking waves using machine learning approaches. Applied Ocean Research, 63, 120–128. https://doi.org/10.1016/j.apor.2017.01.012

[32] Ong, Y.J., Baracaldo, N., & Zhou, Y. (2022). Tree-Based Models for Federated Learning Systems. In: Ludwig, H., Baracaldo, N. (eds) Federated Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-96896-0_2

[33] Afsarian, F., Saber, A., Pourzangbar, A., Olabi, A.G., & Khanmohammadi, M.A. (2018). Analysis of recycled aggregates effect on energy conservation using M5′ model tree algorithm. Energy, 156, 264–277. https://doi.org/10.1016/j.energy.2018.05.099

[34] Obregon, J., & Jung, J.-Y. (2022). Explanation of ensemble models. In: Human-Centered Artificial Intelligence: Research and Applications (pp. 51-72). https://doi.org/10.1016/B978-0-323-85648-5.00011-6

[35] Kurz, S., De Gersem, H., Galetzka, A., et al. (2022). Hybrid modeling: towards the next level of scientific computing in engineering. J. Math. Industry, 12, 8. https://doi.org/10.1186/s13362-022-00123-0

[36] Adèr, H. J. (2008). Modelling. In: Adèr, H. J.; Mellenbergh, G. J. (eds), Advising on Research Methods: A consultant's companion (pp. 271–304). Johannes van Kessel Publishing, Huizen, The Netherlands.