WCRG WORLD COMET RESEARCH GROUP

International Journal of Modern Research in Science & Engineering
Vol. 2, No. 2, 2019
Doi: 10.67228/30716357/IJMRSE-2019PII9J4K
PP. 01-14

Original Article

Urban Flood Prediction Models Using GIS and Machine Learning

Dr. Ahmed Hassan[1], Dr. Fatima Noor[2]

[1]*Professor, Cairo University, Egypt.*

[2]*Associate Professor, University of Malaya, Malaysia.*

Received: 09-04-2019 Revised: 09-24-2019 Accepted: 09-29-2019 Published: 10-04-2019

ABSTRACT

Cities flooding is a major hydrological risk in rapidly urbanizing regions, intensified by climate change, extreme rainfall, and inadequate drainage systems. Traditional hydrological models require extensive calibration and high-resolution data, which are often unavailable for many cities. To address this issue, this study proposes a flood prediction framework that integrates Geographic Information Systems (GIS) with Machine Learning (ML). Multi-source geospatial data such as digital elevation models (DEM), land use–land cover (LULC), soil properties, drainage density, rainfall intensity, and historical flood records are used to derive flood conditioning factors in a GIS environment. These factors are then applied as inputs for supervised ML algorithms including Artificial Neural Networks (ANN), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting. Model performance is evaluated using RMSE, MAE, R^2, and ROC–AUC metrics. Results indicate that hybrid GIS-ML models outperform traditional approaches by effectively capturing non-linear relationships between hydrometeorological and urban factors. The framework provides a reliable decision-support tool for urban planners and disaster management authorities to identify flood-prone areas, improve drainage planning, and enhance early warning systems.

KEYWORDS

Urban Flooding, GIS, Machine Learning, Flood Susceptibility Mapping, Random Forest, Artificial Neural Networks, Spatial Analysis.

ISSN: 3071-6357
PP: 01-14.

1. INTRODUCTION

1.1. Background

The effects of heavy rates of rainfall, surface runoff, poor drainage systems and large scale land-use alteration has become one of the most common and destructive natural disasters in cities globally, as a result of the urban flooding phenomenon. Compared to riverine flooding which normally occurs over longer durations and along well-established channel networks, urban flooding is in general characterized with very short response periods as well as high spatial variability. This ensues that even forecasting and control is very complicated as the waters of the flood can pile up quickly in their topographical depressions and in poorly drained routes before they can institute any emergency actions. The issue is also aggravated by fast urbanization that changes the natural landscapes into densely populated areas. The rampant encroachment of vegetated and permeable coverages (asphalt, concrete and rooftops) by impervious surfaces alters the hydrological cycle of the city's urban areas in their most fundamental forms. The surfaces cause significant reductions in infiltration capacity and groundwater recharge and elevated surface runoff that is formed during rainstorm activities. This has led to stormwater systems being overloaded particularly when it happens under short duration high intensity storms that surpass the design capacity. This is further aggravated in most of the cities by aging or poorly maintained drainage networks which limit effective water transportation resulting in waterlogging and flash flooding during the moderate rainfall levels. The effects of urban flooding are not restricted to sheer inundation of the urban setting since its effects are much far-reaching with regard to economic, environmental, and social impacts. The occurrence of floods would affect transportation networks, destroy residential and commercial buildings, pollute water resources, and affect essential infrastructure like power and communication networks. Communities are usually hit disproportionately by vulnerable communities, who suffer displacement, the threat of health, and economic damage in the long term. Such mounting threats explain the need to have more effective flood prediction, assessment and management approaches. It is against such challenges that this research attempts to investigate the emerging and progressive data-driven and spatially explicit methods, capable of enhancing the effectiveness of the urban flood prediction and reinforce the presence of proactive planning and mitigation tools against vulnerabilities to ensure the future development of more resilient and sustainable urban settings.

1.2. Role of GIS in Urban Flood Analysis

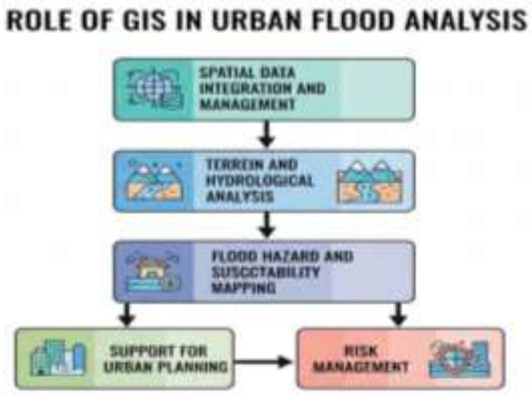

Fig 1 - Role of GIS in Urban Flood Analysis

ISSN: 3071-6357
PP: 01-14.

1.2.1. Spatial Data Integration and Management

Geographic Information Systems (GIS) are of central importance in the process of analysing urban floods as they allow integrating and managing various spatial and non-spatial datasets in a single framework. The urban flood studies demand the concurrent evaluation of numerous factors such as topography, land use- land cover, soil properties, rainfall distribution and drainage systems. GIS will offer potent capabilities to store, operate, and superimpose these heterogeneous datasets in a uniform spatial resolution. GIS allows harmonizing the data provided by various sources and resolutions and thus providing the necessary accuracy in terms of the spatial alignment in the urban landscape and offering the ability to examine the flood-affecting parameters in a detailed manner.

1.2.2. Terrain and Hydrological Analysis

Among the signs that GIS can make an important contribution to flood studies is its capability of extrapolating flood-related and hydrological features using Digital Elevation Models (DEMs). The slope, aspect, curvature, direction of flow as well as flow accumulation maps are generated with the help of GIS-based tools and they map the way water flows on the surface. These derived parameters can be used in determining low modules, natural drainage paths, and runoff convergence zone which may be prone to flooding. In urban areas where the natural drainage patterns may be altered by other constructions, the hydologic analysis performed using GIS offers vital information on the interaction between surface runoff and the manmade and natural drainage infrastructure.

1.2.3. Flood Hazard and Susceptibility Mapping

GIS serves very well to create flood hazard and susceptibility maps that can provide a visual representation of the spatial distribution of risk of floods. Using the multi-criteria assessment techniques or combining the results of hydrological and machine learning models, GIS facilitates categorization of urban regions into various risk levels including low, moderate, and high flood prone regions. Such maps are useful decision support tools enabling the authorities and planners to determine vulnerable areas in a short time and focus on possible measures to mitigate them. The graphic aspect of GIS deliverables also improves the communication between the scientists, policymakers, as well as the population hence facilitating informed and transparent decision-making.

1.2.4. Support for Urban Planning and Risk Management

In addition to analyze and create maps, GIS is a very important tool in correlating the results of flood assessment with strategies of urban planning and managing disasters. The vulnerability of floods can be mapped and superimposed with population density, infrastructure system and land-use plans to determine the exposure and the effects that may occur. With this spatial connection, authorities are able to assess the performance of their current drainage systems and design new infrastructure development schemes as well as controlling development in flood prone regions. Moreover, any GIS-based implementation may be connected to real-time rainfall and sensor observations to facilitate an early warning plan and emergency response planning. With such functions, GIS contributes to not only the technical evaluation provided to the risk of urban floods, but also to its practical use in sustainable urban development and resilience establishment.

ISSN: 3071-6357
PP: 01-14.

1.3. Machine Learning in Hydrological Modeling

Machine learning (ML) has become a highly influential and more and more significant instrument in hydrological models owing to its prospective of depicting multifaceted and nonlinear connections among hydrological variables without any specific physical assumptions. Conventional hydrological models are generally simplistic representations of physical processes and many of them frequently require large numbers of parameters be tuned on with observational data of limited extent. Contrarily, ML methods do not rely on indirect evidence but learn directly on the existing historical data, so they are able to construct complex interaction among rainfall, the overall land surface features, soil properties, and the overall runoff reaction. This information-driven character predisposes ML to highly uncertain and high-variability hydrological systems, including those of urban catchment involved in rapid land-use alteration as well as heterogeneous surface conditions. Over the past few years, there has been a large selection of different ML algorithms utilized in hydrological applications, such as streamflow predictions, rainflood-runoff modeling, prediction of the groundwater level, and flood prone mapping. ANNs have been widely employed to estimate nonlinear rainfall runoff processes, whereas Support Vector Machines (SVMs) have reported high generalization power particularly in instances where data to train the model are small. Ensemble learning techniques like Random Forest and Gradient Boosting are additional techniques that improve predictive accuracy through combination of several weak learners and limit the impact of noise and overfitting. The models can work with high density datasets and determine the most significant predictors which give a rich information on the quantification of the relative significance of environmental and hydrological variables. Although the performance of ML-based hydrological models is good in predicting, there are also some issues that can be associated with it especially in terms of interpretability as well as physical realism. The outputs of many ML algorithms are black boxes, and it is hard to trace the outputs to a particular set of laboratory phenomena. These limitations have however been started to be overcome by recent developments in explainable artificial intelligence (XAI) and hybrid modeling methods, which incorporate physical knowledge besides learning based on the data. In general, machine learning in hydrological models is a meaningful change of methodology, which provides more accuracy, adaptability and scalability. With the further increasing availability of high-resolution spatial and temporal data, ML has a promising future of becoming more central in the areas of flood prediction, water resource management, and climate resilience planning with the assistance of this type of data.

2. *LITERATURE SURVEY*

2.1. Conventional Flood Modelling Approaches

Physical Hydrological and Hydraulic models that are based on traditional flooding models have significant modeling dominance and they include the Rational Method, The Soil Conservation Service Curve Number model (SCS-CN) and the Saint-Venant equations. The models strive to model physical processes of the real world e.g. transformation between rainfall and runoff, transformation between rainfall and infiltration and dynamics of the channel flow using mathematical expressions. Their primary advantage is that they can be physically interpreted, such that an engineer and hydrologist can determine the impact of rainfall, land use, or channel geometry adjustments on the behavior of floods. Their usefulness, however, is strongly dependent on the quality and access to input

ISSN: 3071-6357
PP: 01-14.

data concerning the characteristics of catchments, soil properties and channel cross-sections. Practically, these detailed data are usually missing or poorly shaded, particularly in cities or sparsely populated areas. Additionally, these models are also sensitive to the calibration of parameters and assumptions of the boundaries conditions that may result in major errors in predicting floods in highly nonlinear systems with a rapidly changing behaviour with urbanization and climate variation.

2.2. Gis-Based Flood Mapping

The widespread use of Geographic Information Systems (GIS) in flood susceptibility and hazard mapping can be attributed to the possibility of integration, analysis, and visualizations of any spatial data. Various research, such as that of Julian and Stiok (2019) and multi-criteria decision analysis (MCDA) application such as the Analytic Hierarchy Process (AHP), has integrated several factors affecting floods, including slope, elevation, rainfall density, land use, soil type, and soil drainage density. These techniques produce flood hazard zonation maps through the assignment of relative significance (weights) to each factor, and superimposition curves. GIS-based methods are specifically useful in the regional level planning and risk evaluation because they enable the generation of visual outcomes that offer hand-on results to the decision-maker. A major weakness however is that the weighting schemes tend to be subjective because in most cases they are based on the expert judgment. Such subjectivity may bring in bias and lessen reproducibility in various fields of study. Moreover, such approaches, as a rule, do not include complex nonlinear components among the driving forces of floods that may reduce their predictive accuracy in the evolving environment.

2.3. Machine Learning for Flood Prediction

Over the past few years, machine learning (ML) tools have been receiving growing interest in flood prediction and susceptibility evaluation because they can easily model nonlinear relationships that are complex using a large amount of data. Time-series flood forecasting with artificial Neural Networks (ANNs) has shown very promising results with their use in modeling processes such as rain fall-runoff and streamflow dynamics. The Support Vector Machines (SVMs) have been shown to work well with small and medium sized dataset due to the high capacity to generalize and ability to withstand overfitting. Ensemble techniques include random forest (RF) and gradient boosting which have further enhanced accuracy in prediction since they incorporate many decision trees and minimize both variance and bias. The models can be useful especially in the identification of important predictors and handling of high-dimensional data. In more recent years hybrid methods which combine GIS with ML techniques have been suggested, which gives spatial flood susceptibility mapping based on data learning instead of subjective weighting. Although the predictive power and the capability of ML-based models are high, they are inherently black-box, which makes them challenging to interpret and establish physical realism, causing challenges in rolling out these models in engineering and policy-making settings.

2.4. Research Gaps

Although significant progress has been achieved in the traditional modeling, GIS mapping, and machine learning, there are few issues that still require resolution. Model generalization is one of them since most of the suggested approaches can only be effective in the geographical or climatic setting

ISSN: 3071-6357
PP: 01-14.

under which the models have been trained. Interpretability is also an issue, especially in the case of ML models, that are often not easily connected on the basis of physical further flood mechanisms and input variables. Besides, current studies are either inclined toward temporal flood prediction (e.g., rainfall runoff prediction) or spatial flood susceptibility mapping, but there are only few attempts to combine the two viewpoints into a single model. Lack of focus on the integration of flood prediction models in the urban spatial planning and the decision-support systems is also underemphasized. Thus, the complex GISML system to both predictively cover the time and at the same time map spatial susceptibility and applicability of the practice in managing floods in cities is not well-developed yet, which is a clear gap in the research sphere to be filled in the future.

3. METHODOLOGY

3.1. Study Area and Data Collection

The chosen study area is an extremely urbanized area that has been hit by high pluvial floods as a result of high intensity rain levels, high rate of land-use alteration, and poor drainage systems. The preponderance of hard and artificial surfaces like roads, roofs, and pavements decreases the natural infiltration and rises surface runoff therefore leaving the area especially susceptible to high intensity, short period precipitation. The multidimensional relationship between architectural and natural topography implies the use of a comprehensive spatial and hydrometeorological study of the specifics of flood processes. This then led to the cumulative data of geospatial and hydro-meteorological data to facilitate the subjection of flood forecasting and susceptibility mapping.

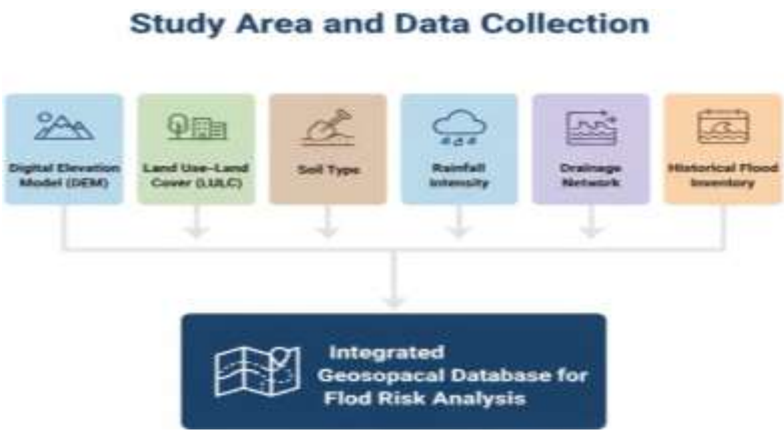

Fig 2 - Study Area and Data Collection

3.1.1. Digital Elevation Model (DEM)

Digital Elevation Model (DEM) is the backbone information of the terrain feature including the elevation, slope, and direction of flow which have potent effects on the pattern of water runoff and water collection on the surface. This paper employed the DEM to compute the topographic factors such as slope gradient, aspect, flow accumulation and watershed boundaries. These derived layers play a vital role in determining lowlands and natural drainage routes on which the floodwaters tend to accumulate thus constitute a vital input to both GIS-based model and machine learning.

ISSN: 3071-6357
PP: 01-14.

3.1.2. Land Use–Land Cover (LULC)

The LULC data give an account of spatial distribution of urban built-up types, vegetation, water bodies and open ground within the area of study. The data is especially valuable in urban flood studies due to the fact that the volume of runoff and peak discharge of the developed areas is quite high as a result of high imperviousness. Temporal variations in LULC are also insurance of patterns of urban expansion, which can increase flood risk by settling in natural drainage systems and floodplains. This paper utilized the LULC layer to assess surface permeability and to determine how the urban development affected the susceptibility to flooding.

3.1.3. Soil Type

The type of soil has a critical influence on the limitation of infiltration capacity, water retention and generation of runoffs. Various soil textures like sandy, loamy, or clay substance soils have different hydraulic conductivities and moisture storage aptitudes. Low permeability soils are likely to enhance the rapid delivery of surface water and hence flood-inducing risks in times of heavy precipitation. The soil map was hence included to reflect the spatial variation in the infiltration behavior to facilitate estimate the response of runoff during the various rainfall conditions.

3.1.4. Rainfall Intensity

The rainfall intensity data is the major attracting variable of pluvial floodings and necessary in the modelling of rainfall-runoff interaction. In this research, rain information was acquired through meteorological stations and complemented as required with those obtained through satellites in form of precipitation product to enhance spatial coverage. Temporal variability was also taken into consideration by taking both short period extreme rain events and long term rainfall statistics. These data have served as major predictors in flood modeling structure to determine the association between precipitation pattern and floods.

3.1.5. Drainage Network

Natural streams and artificial stormwater networks form part of the drainage network layer, and they regulate how run-off flows over the urban space. Features like the density of drainage channels, the connection between the drainage channels and the closeness to drainage lines determine the speed at which the extra water could be moved out of the developed space. Poor or damaged drainage systems tend to locally waterlog and flood the urban areas. Flow pathways were analyzed with the drainage network dataset, which helped detect the regions where drainage congestion is one of the factors that raise the flood risk.

3.1.6. Historical Flood Inventory

The historical flood inventory is a compilation of historical flood occurrences in terms of location and magnitude that is gathered based on field surveys, government reports and remote sensing data. This data will be used to check the outputs of the models and to learn machine-learning models. The inventory facilitates analysis of essential factors influencing floods by establishing connections between the relevant data on flooding and the relevant environmental and hydrological factors to come up with an effective model of flood susceptibility. The use of the historical flood records

ISSN: 3071-6357
PP: 01-14.

also makes it inevitable that the modeling platform captures the actual aspects of the flood behavior in the real world rather than just the normal theoretical assumptions.

3.2. Data Preprocessing

Preprocessing of data is a very important exercise towards the consistency, reliability and compatibility of multi-source spatial and hydrometeorological data before the models and analysis are done. The all thematic layers of the study such as the topographic, hydrological and land surface variables were initially translated into images in raster format and resampled to a common spatial scale to remove the variability of the scales. This action will help to identify that all the pixels on various datasets are indicating the same amount of land on the ground, and the cell by cell comparison and merging of them can be done correctly in the GIS environment. Moreover, projection of all datasets into similar coordinate reference system was done to prevent positions mismatch, which can occur due to different map projections and datum definitions. This is necessary to provide geometric standardization required to make accurate overlay operations and spatial analysis. In order to enhance the performance of the models and the numerical stability, all the continuous input variables were normalized. As the predictor variables (e.g., elevation, slope, rainfall, and drainage density) have varying unit and value range, raw data may be used directly, making the learning process in machine learning algorithms to be biased to those variables with more numerical magnitude. Normalization is used to convert the initial values into a normalized domain, which is usually between 0 and 1, hence making sure that all variables have proportional contributions to the model training procedure. The procedure increases the convergence rate, decreases the computational complexity and increases prediction accuracy. Also, categorical data sets like land use- land cover and soil type were reclassified to meaningful numerical classes according to hydrological importance and could therefore be used in the model. Data gaps and noise were minimized by looking carefully at the missing or anomalous values and correcting them by methods of spatial interpolation and neighborhood based filtering. The statistical measures was also done to evaluate multicollinearity among conditioning factors and highly correlated variables were eliminated or collapsed to limit redundancy and overfitting. Using these steps of preprocessing, the raw geospatial and hydro-meteorological data were converted into a harmonized, normalized, and modeled input database that can later be analyzed by GIS and be used in machine learning-based flood prediction.

3.3. Feature Extraction In GIS

In a GIS setting, feature extraction is a very vital tool in modifying raw elevation and converting them into significant terrain attributes that determine hydrological and surface runoff dynamics. The Digital Elevation Model (DEM) was used as the main source of obtaining the essential topographic variables in this research, such as slope (S), curvature (C), and flow accumulation (FA). Slope is the rate of change in elevation and has an immediate effect on the velocity of the overland flow; steeper slopes will tend to encourage faster runoff and less infiltration whereas the gentler slopes will tend to encourage water release and ponding. Therefore, slope is a very important aspect that determines regions of surface water retention and flash floods. Curvature explains the form of the surface of the earth and is applied in defining whether there is convergence or divergence of flow on the surface. The profile curvature influences the acceleration and deceleration of the flow, and the plan curvature

ISSN: 3071-6357
PP: 01-14.

influences the lateral dispersion or the concentration of the water. Concave curvy areas usually serve as points of convergence of the water flow and usually experience greater accumulation whereas the convex surfaces are more likely to scattered away the runoffs. The addition of curvature to the analysis thus adds more information to the model that shows the effect of all the minute controls of topography on flood activities that due to elevation alone they were not able to identify. Flow accumulation is a measure of the volume of contributing upstream area that is draining in each grid cell and acts as an indicator of runoff concentration which could occur. Cells that have higher flow accumulation values are an indication of the natural drainage routes or depressions where water is likely to pool during rainfall activities. The parameter is especially crucial in the urban flood research because it aids in the identification of the low lying corridors and blocked drainage paths that might suffer overwaterlogging. These terrain qualities were derived through the standard hydrological models in the GIS platform like the slope analysis and the flow direction and flow accumulation algorithmic methods using the DEM. The derived layers were then normalised and combined with other thematic data to create input feature set of the machine learning models. These GIS-generated features can be treated as independent variables and are used to determine the spatial relationship between topography and flooding occurrence in a particular region and therefore enhance the quality and interpretability of the flood prediction output because they are more physical structure of the landscape represented in quantitative form.

3.4. Machine Learning Model Design

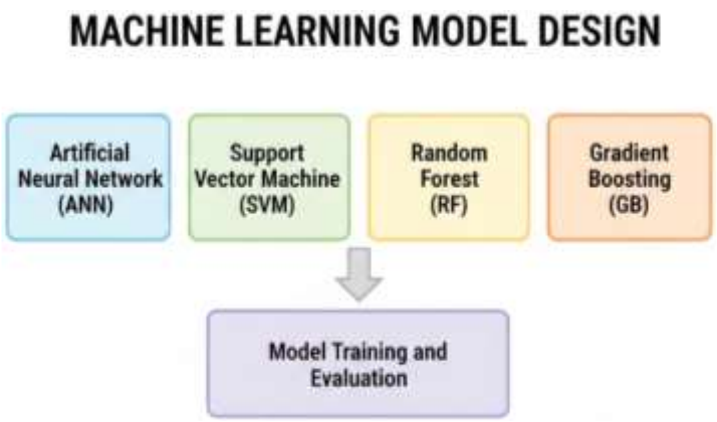

Fig 3 - Machine Learning Model Design

3.4.1. Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) is a model, which aims at the modeling of intricate non-linearity relationships between conditioning factors and flooding. It comprises interwoven networks of neurons, one of them being an input layer that emulates the extracted GIS types, the hidden layers that tend to learn the patterns and an output layer, which predicts vulnerability to flooding. With the aid of a backpropagation process of learning, the ANN modifies the connection weights to reduce the error in prediction. The model is very applicable in the rainfall-runoff and flood prediction work because it has a high ability to be approximated and it can also be modified to suit the major large-scale multi-dimensional data.

ISSN: 3071-6357
PP: 01-14.

3.4.2. Support Vector Machine (SVM)

Support Vector machine (SVM) is put into practice as a supervised learning classifier, which divides flood and non-floods through building an ideal decision boundary in a high dimensional feature space. SVM is able to take nonlinear correlation between terrain features and occurrence of floods by employing kernel functions including radial basis function (RBF). SVM is also notorious with its high levels of resiliency with sparse training sparse data and capability of avoiding overfitting with enhancement of margin maximization, it is highly applicable in spatial flood susceptibility modelling.

3.4.3. Random Forest (RF)

Random Forest is an ensemble method based on the outlook of learning where numerous decision trees are created after applying random selections of the training samples and predictor variables. The last prediction is the sum of the predictions of the individual trees which enhances better accuracy and less variance. RF specifically appears to work well with high-dimensional data and discovery of the relative significance of input variables including slope, flow accumulation, and land use. It is a stable model of flood prediction application because it is not susceptible to overfitting and can handle noisy data sets.

3.4.4. Gradient Boosting (GB)

My other technique that is an ensemble one is Gradient Boosting, which assembles a sequence of weak models of prediction which are usually decision trees in a serial fashion. A new model is trained to reduce a given loss function to remove the mistakes of the last one. Gradient Boosting implements this iterative type of learning and thus attains a high degree of predictive accuracy and quality generalization. The algorithm has been found useful in flood susceptibility modeling to capture subtle nonlinear interactions between conditioning factors and is found to be particularly effective in enhancing classification performance in urban environments of complexity.

3.5. Model Evaluation

The model evaluation is a vital procedural requirement of determining the stability and persona forecast of the suggested machine learning models in flood susceptibility mapping. In this work, a mixture of the statistical and classification-based measures of accuracy are used in modeling performance on the basis of the comparison of forecasted flood-prone regions with the history of floods. The data set is separated into training data and testing data to avoid biased evaluation of the case of generalizing the model. The typical metrics of evaluation used include accuracy, precision, recall (sensitivity) and F1-score which are used to measure the classification performance of any given algorithm. Accuracy evaluates the total percentage of the total samples that were rightly classified as floods and non-floods, whereas accuracy evaluates the percentage of correctly predicted cases of floods among all the cases that are predicted to be floods. Recall can measure how well the model detects real flood events, but this is especially relevant in flood risks research because a false negative (i.e. missing flood events) may have significant impact. Besides these, the area under the curve (AUC) and the Receiver Operating Characteristic (ROC) curve are applied in determining the discriminatory power of the models. ROC curve indicates true positive rate verses false positive rate across various threshold values, giving insight in to model robustness to be expected in various classification conditions. With

ISSN: 3071-6357
PP: 01-14.

the increase in AUC, this indicates enhanced all-round performance and increased capacity to separate the flood and non-flood prone land. Moreover, the confusion matrix analysis is applied to illustrate the classification outcomes as well as to gain a clearer idea of how the correct and incorrect predictions were distributed.

4. RESULTS AND DISCUSSION

4.1. Model Performance Comparison

4.1.1. Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) model shows good predictive power, whereby, the RMSE value amounts to 0.42 and MAE amounts to 0.31, which implies a moderate prediction error. The coefficient of determination [R 2 (= 0.86) indicates that the model can follow a big percentage change in the flood occurrence. The AUC of 0.89 is also another indication that ANN is highly discriminatory as regards flood-prone and flood-free areas. Such findings emphasize the capability of ANN to represent nonlinear interdependencies among the factors used in the conditioning of floods; but its relatively low error rates in comparison with the ensemble models reveal their sensitivity to parameter and training data distribution.

4.1.2. Support Vector Machine (SVM)

The SVM model presents a better performance compared to ANN with lower RMSE (0.39) and MAE (0.28) which leads to a better predictive power. R 2 = 0.88 is more explanatory since it represents the better representation of the variability of the flood patterns. The value of AUC is 0.91 that provides good classification and accurate separation of flood and non-flood classes. The noise conditions and sample size characteristics of the techniques and its capability to optimize its margin attribute to the high strength of the SVMs which is capable of finding high-quality solutions to non-linear decision boundaries.

4.1.3. Random Forest (RF)

Random Forest (RF) has the highest overall performance as indicated by the lowest RMSE (0.34) and MAE (0.24), which indicates a minimum prediction error. The best value of R 2 (0.91) is observed to be more competent in describing the variation of flood vulnerability. Moreover, the value of AUC (0.94) is great discriminating power with great ability to generalize. We can attribute the good performance of the RF parameter to its ensemble structure, which minimizes overfitting as it incorporates a series of decision trees in addition to effectively encompassing even complicated interactions between the parameters of topography, hydrology and land-use.

4.1.4. Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) also demonstrates good predictive accuracy whereby RMSE and MAE are 0.36 and 0.26 respectively. The fact that its R 2 value is 0.90 indicates that it is well capable of depicting underlying relationships between drivers of floods and the observed floods. The AUC value is 0.93, which means that the classification is excellent, although less than that of the Random Forest but better than ANN and SVM. GBM has been strong in its performance due to its strategy of sequential learning with every new model correcting all the flaws made by the previous model

ISSN: 3071-6357
PP: 01-14.

enabling it to identify subtle non-linear trends. In general, RF and GBM decide these two models though in this case, RF proves to be a more credible algorithm in predicting urban flood.

4.2. Spatial Flood Susceptibility Mapping

The trained machine learning models are applied to all the conditioning factors in the whole study area to create spatial flood susceptibility maps. The value of each grid cell is determined to be the flood susceptibility measure, according to the relationships that have been learned between the topographic, hydrological, and landscape surface variables and the occurrence of floods in the past. The resulting continuous susceptibility is then categorized into the qualitative risk groups, including very low, low, moderate, high and very high susceptibility so as to be easily interpreted and presented in a form of a visual representation. Such maps are also useful as they offer a spatial picture of zones with flood risks, and it is used as the effective tool to identify the priority areas in flood risk management and the public plan. The resultant patterns of flood susceptibility indicate that the areas that are at high risk are mostly located at the low terrain area where the natural drainage lines meet with where surface water gathers. Regions with low runoff velocity and high runoff accumulation values indicate the high inclinations of occurrence of flood occurrence because of low run off velocity with large water contents. Moreover, areas where there is a lot of coverage of impervious surface i.e. dense residential areas, commercial and industrial areas have a higher susceptibility level. These areas do not have permeable areas reducing the capacity of infiltration and amplifying the surface runoff greatly during flood events, thus exaggerating flood hazards. Thickness and lack of interconnection of the drainage networks also contribute more to the susceptibility toward floods especially in areas where the stormwater channels are too small or blocked. This is also facilitated by the fact that most built-up areas are located near the natural streams and artificial drainage lines so that when a storm occurs, any physical surplus will quickly flow into the drain lines. On the other hand, the regions with elevation, slopes, and densely vegetated covers tend to be less susceptible because they have better conveyance of the runoffs and their infiltration is greater. In general, the flood susceptibility maps created point to the interplay of the terrain morphology, land use, and drainage features in determining the flood hazard distribution. Not only are these spatial outputs confirming the predictive ability of the machine learning models but also offer a useful decision-making framework to help estimate the vulnerable neighborhoods, such as the optimization of the draining structures upgrading and the provision of strategic planning of sustainable urban development options to minimize the impact of floods in the future.

4.3. Discussion

Findings of this paper show that the Random Forest (RF) model is the best of the identified machine learning models in terms of predicting the likelihood of urban floods due to having the lowest error rates, greater explanatory and classification power. It is the ensemble learning format of RF that allows the high performance of the species to be explained first of all by the ability of RF to combine the results of multiple decision trees to come up with a more stable and robust result in terms of prediction. With many trees being trained with various subsets of the data and predictor variables RF helps in reducing variance and lessening the chances of overfitting. This feature is of key significance especially in flood modeling where the input factors (slope, Land use, rainfall and drainage density)

ISSN: 3071-6357
PP: 01-14.

are highly correlated. RF is reliable because it allows multicollinearity to be managed without making strong data independence assumptions, which contributes to its suitability in complex urban settings in terms of both surface and hydrological heterogeneity. When machine learning is combined with GIS, the analytical framework is reinforced further as a spatially explicit model result could be interpreted. Although machine learning algorithms have good predictive accuracy, they are much more useful when geospatially mapped. GIS has enabled visualization of the patterns of susceptibility and enables direct correlations between the predicted flood risk and the physical locations, elements of infrastructure, and classes of land-use. This spatial interpretability converts abstract numerical projections into useful information that can be taken by urban planners, disaster management authorities, and policymakers.

5. CONCLUSION

This paper has constructed and presented a hybrid system based on Geographic Information Systems (GIS) and machine learning methods to predict urban floods and the vulnerability of an area. The proposed method is efficient because it integrates the physical and the spatial causes of urban flooding using spatial datasets like topography, land use-land cover, soil characteristics, rain level, as well as drainage system. Results made it very clear that data-based machine learning models can be compared far better than traditional hydrological and hydraulic models in their level of prediction and their generalizational capacities especially within the complex urban contexts where nonlinear interplay of variables prevails. The algorithms that were tested, in particular, ensemble-based, with a significant impact on the methods of random forests, demonstrated better results due to resistant to noise, multicollinearity, and the possibility to create the complex relationships without strict assumptions regarding the data distributions. Incorporation of machine learning entries into a GIS context can be seen as an important breakthrough in the flood risk analysis as it allows visualizing and interpreting model forecasts. The resulting flood susceptibility maps give a clear and realistic form of flood prone areas in the study area that is easy to identify sensitive areas that may be affected by pluvial flooding. The maps would be of great use to the urban planners, engineers and customers of disaster management as they indicate where drainage capacity is minimal, where the area is covered with excessively solid ground or where the natural course of the drainage or flow is blocked. In turn, the framework stimulates evidence-based planning and allows prioritizing the mitigation efforts, e.g., improving the drainage infrastructure, retention and detention units, and establishing land-use regulations that not only deter the development of the high-risk area. Also, the results demonstrate why data-driven methods of developed urban flood management plans should be applied, especially in swiftly urbanizing areas where the physically based model traditionally relies on the limited availability of data and a large degree of uncertainty in the parameters. Machine learning methods are flexible enough to enable continuous improvement as new data is available hence maximizing the flexibility of the models in the context of changing climatic and urbanization conditions. Even though the current framework has been proved efficient, it can be revised and modified even more. The new studies that need to be conducted in the future should concentrate on the integration of real-time and near-real-time sources of data including the rainfall radar, Internet of Things (IoT) sensors, and crowd-sourced flood reports to make it possible to implement dynamic flood forecasting and early warning

ISSN: 3071-6357
PP: 01-14.

systems. Moreover, the use of deep learning engines, such as convolutional and recurrent neural networks has great opportunities to enhance the spatiotemporal pattern recognition and representation of rich dependencies in large datasets. On the whole, the suggested GIS-machine learning framework offers a scalable and transferable approach to the issue of urban flood predictions and also offers a solid base to the creation of more resilient and adaptable flood risk management frameworks.

REFERENCES

[1] Chow, V. T., Maidment, D. R., & Mays, L. W. (1988). *Applied Hydrology*. McGraw-Hill, New York.
[2] Beven, K. J. (2001). *Rainfall–Runoff Modelling: The Primer*. John Wiley & Sons, Chichester.
[3] HEC-RAS River Analysis System User's Manual. (2016). Hydrologic Engineering Center, U.S. Army Corps of Engineers.
[4] Te Chowdhury, R., & Ward, N. (2004). Hydrometeorological variability and flood prediction in the Upper Blue Nile Basin. *Journal of Hydrology*, 291(3–4), 291–300.
[5] Pradhan, B. (2010). Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing. *Journal of Spatial Hydrology*, 9(2), 1–18.
[6] Kazakis, N., Kougias, I., & Patsialis, T. (2015). Assessment of flood hazard areas using GIS-based multi-criteria analysis. *Water Resources Management*, 29(10), 3681–3697.
[7] Gigović, L., Pamučar, D., Bajić, Z., & Drobnjak, S. (2017). GIS-based flood hazard mapping using MCDA: A case study of Belgrade, Serbia. *Water*, 9(3), 1–19.
[8] Jain, A., & Srinivasulu, S. (2004). Development of effective and efficient rainfall–runoff models using artificial neural networks. *Journal of Hydrologic Engineering*, 9(6), 404–415.
[9] Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. *Progress in Physical Geography*, 25(1), 80–108.
[10] Yu, P. S., Chen, S. T., & Chang, I. F. (2006). Support vector regression for real-time flood stage forecasting. *Journal of Hydrology*, 328(3–4), 704–716.
[11] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
[12] Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water*, 10(11), 1–40.
[13] Rahmati, O., et al. (2016). Flood susceptibility mapping using frequency ratio and weights-of-evidence models. *Geomorphology*, 268, 66–81.
[14] Khosravi, K., et al. (2018). Flood susceptibility modeling using machine learning methods. *Journal of Hydrology*, 561, 940–955.
[15] Merz, B., Vorogushyn, S., Uhlemann, S., Delgado, J., & Hundecha, Y. (2015). More efforts and scientific rigour are needed to attribute trends in flood time series. *Hydrology and Earth System Sciences*, 16(5), 1379–1387.