Scientific Research Publishing

Journal of Geoscience and Environment Protection, 2022, 10, 61-83
https://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Merging GIS and Machine Learning Techniques: A Paper Review

Chikodinaka Vanessa Ekeanyanwu[1]*, Inioluwa Feranmi Obisakin[2]*, Precious Aduwenye[2]*, Nathaniel Dede-Bamfo[1]

[1]Department of Geography and Environmental Studies, Texas State University, San Marcos, TX, USA
[2]Ingram School of Engineering, Texas State University, San Marcos, TX, USA
Email: cve9@txstate.edu

How to cite this paper: Ekeanyanwu, C. V., Obisakin, I. F., Aduwenye, P., & Dede-Bamfo, N. (2022). Merging GIS and Machine Learning Techniques: A Paper Review. *Journal of Geoscience and Environment Protection, 10,* 61-83.
https://doi.org/10.4236/gep.2022.109004

Received: August 18, 2022
Accepted: September 16, 2022
Published: September 19, 2022

Abstract

GIS (Geographic Information Systems) data showcase locations of earth observations or features, their associated attributes and spatial relationships that exist between such observations. Analysis of GIS data varies widely and may include some modeling and predictions which are usually computing-intensive and complicated, especially, when large datasets are involved. With advancement in computing technologies, techniques such as Machine learning (ML) are being suggested as a potential game changer in the analysis of GIS data because of their comparative speed, accuracy, automation, and repeatability. Perhaps, the greatest benefit of using both GIS and ML is the ability to transfer results from one database to another. GIS and ML tools have been used extensively in medicine, urban development, and environmental modeling such as landslide susceptibility prediction (LSP). There is also the problem of data loss during conversion between GIS systems in medicine, while in geotechnical areas such as erosion and flood prediction, lack of data and variability in soil has limited the use of GIS and ML techniques. This paper gives an overview of the current ML methods that have been incorporated into the spatial analysis of data obtained from GIS tools for LSP, health, and urban development. The use of Supervised Machine Learning (SML) algorithms such as decision trees, SVM, KNN, and perceptron including Unsupervised Machine Learning algorithms such as k-means, elbow algorithms, and hierarchal algorithm have been discussed. Their benefits, as well as their shortcomings as studied by several researchers have been elucidated in this review. Finally, this review also discusses future optimization techniques.

Keywords

GIS, Machine Learning, Landslide Susceptibility, Random Forest, Urban Development, Flood Prediction, Health, GeoAI

*Co-first authors.

1. Introduction

The increase in demand for different types of spatial information has necessitated the adoption of advanced techniques such as machine learning (ML). ML has been deemed as having a great potential and solution to several complex spatial analysis problems in Geographic Information Science (Lazar & Shellito, 2005). ML being a subset of Artificial intelligence comprises various models and algorithms that can be applied to geoprocessing tools to solve problems in numerous areas in GIS. In light of this potential, there is a growing need to carefully curate and review key research and approaches in this field as well as to examine previous findings in order to develop best practices.

Spatial data are described as observations with spatial attributes. An observation represents a location in the real world and can be represented as either a polygon, line, or point feature. Observations in spatial data may have various characteristics including latitudes, longitudes, areas (polygon features), perimeters (polygon features), centroids, and lengths (line features). A group of spatial features could also have density, and centrography (point). It must be added though that, each of the three data types can be converted from one to another depending on the project scope. Examples of polygon features are city boundaries, residence blocks, and land-use areas. Roads, pipelines, rivers, and other route networks are often represented as lines. Point features on the other hand include things like fire hydrants and cellphone towers. Observations such as elevation points or spot heights and water table depths are also sometimes represented using points.

Spatial data can integrate with other data types and strengthen complex analysis of the distribution of locations, events, and services (Tohidi & Rustam, 2020). This potential of spatial data provides many opportunities for scientific advancement in predicting various geospatial related issues such as pandemics, drought, landslides, soil erosion and flood.

Similarly, machine learning techniques have caught the attention of scientists and researchers in various fields, as tools for analyzing and managing large amounts of datasets. As a result, ML is considered as one of the most sensational tools that have gained considerable traction in various fields that engage in continuous research and development in recent years and coincidentally, GIS is one of such fields (Lazar & Shellito, 2005). Therefore, the objective of this paper is to review previous scientific research on the application of Machine learning alongside GIS. The remainder of this paper is organized and structured as following. In the next section (Section 2), GIS and ML are defined and broadly explained. Section 3 presents an overview of ML applications in GIS and related works in this area. The last section provides conclusions to the review.

2. Fundamental Concepts

Fundamental Concepts such as GIS and ML must be introduced and discussed in order to review the different types of Machine learning applications in GIS.

The principles and concepts are introduced below.

2.1. GIS as an Interdisciplinary Subject

Geographic Information Systems is seen in one of 3 ways positions 1) GIS as a tool; 2) GIS as tool-making; 3) the science of GIS. GIS as a tool is seen as the use of a software for the management, analyzing and visualization of geographic data in order to advance some specific topic/purpose, with its development and availability mostly independent of its use which is application. GIS as "tool-making" can be seen as the advancement of the tools capabilities by the makers/developers to aid its utilization by practitioners (Wright, Michael, & James, 1997). The authors are also of the opinion that GIS as a science deal with the connection of both the tool and science to address research problems across a variety of disciplines to provide insight that could aid better decision-making and allocation of resources. We see more applications indicative of its acceptance and recognition in many sectors such as politics, engineering, biology, anthropology, etc. with some case studies provided below:

- **GIS in Politics:** Using the 2012 US presidential elections as a case study, the spatial distribution of web pages and twitter messages with election related content, to visually depict spatial patterns and geospatial footprints for specific keywords. Ultimately, the study showed that the changes in the spatial patterns corresponded to certain major campaign events, suggesting a new angle for studying human thought and behavior as well as social activities (Tsou et al., 2013).
- **GIS in Health:** Health care researchers have adopted spatial analytical methods in analyzing the need to improve access to and utilization of health care services. This is aimed at enhancing the planning and evaluation of health care service locations (McLafferty, 2003). The Coronavirus (COVID-19) pandemic for example, presented its own set of challenges as a large volume of data revealed health issues, disease patterns and disparities faced by specific communities. Most of these findings were indicative of racial and location disparities requiring public health intervention towards the proper prevention of the virus and the treatment of affected individuals/communities (Iyanda, Kwadwo, & Yongmei, 2021).
- **GIS in Urban Development:** SafeCity is a GIS based tool developed to support urban development scenarios in the city of Gdansk with the integration of tools for target analysis, possible hazard setting simulations and spatial analytics. The tool is particularly useful in infrastructural vulnerability assessments in as it helps to identify and mitigate against possible risks (Kulawiak & Zbigniew, 2014).

The growing number of published research dedicated to the application of spatial analysis, and mapping in fields such as criminology, real estate, epidemiology, etc., show a shift in acceptance of GIS in the social sciences. This is often attributed to the spread of affordable GIS technology as well as the availability of referenced data (Anselin, 2000).

2.2. Machine Learning

The concept of Machine Learning (ML) in comparison to GIS is new. According to (Wankhede, 2022), machine learning is described as a process whereby a computer program repeatedly learns from various experiences to improve the performance of specific tasks that it has been assigned. The expectation is that, as the computer program gains more and more experience and insight in executing specific tasks, measurable performance in those tasks directly improves. This means that the machine takes assessments and does predictions based on data fed into it (Wankhede, 2022). ML is a division of artificial intelligence that emphasizes the expansion of computer programs that can access data and use it in the process of learning and relearning (Ray, 2019). In reference to the above descriptions, the authors also define ML as a section of artificial intelligence that provides systems with the capacity to instinctively learn and improve their performance from experience without being explicitly programmed. Over the years, Machine Learning has been applied to numerous sectors such as virtual personal assistants (for example Apple's Siri, Amazon's Alexa and Google's voice assistant), computer games, natural language processing, and traffic prediction and transportation analysis. Others include product recommendation (e.g. Amazon recommendation service), stock market prediction, medical diagnosis (e.g. first level cancer diagnosis), online payment fraud prediction, optimization of search engine result and indexing (Burns, Laskowski, & Tucci, 2022).

Furthermore, ML is especially useful in scenarios where using human resources is not time and cost effective. It is also very applicable when one has to consider many variables concurrently. ML uses the prepared data which includes the selected features to train a ML algorithm. A ML model is generated when the algorithm is properly trained on the refined data. Once the training process is concluded, the ML model can be used to make predictions about the future data on its own.

As indicated earlier, ML involves some processes, and these are shown in **Figure 1**. The learning process begins with observations recorded as raw data. The data are then preprocessed, and a part is selected based on some specific features and fed to the ML algorithm. The ML algorithm in turn looks for patterns in the fed data and makes decisions based on provided examples. The prime objective of the entire process is to allow the machine to learn without human involvement and adjust actions accordingly (Tohidi & Rustam, 2020).

Types of Machine Learning:

Machine learning models are often categorized as supervised, unsupervised and reinforcement learning. The type of available training data determines which model to apply in the machine learning algorithm. These model types are described below:

- Supervised Machine Learning (SML) is a method of creating intelligence where the ML algorithm is trained on input data that have been labeled for a particular output (Burns, Laskowski, & Tucci, 2022). The model is trained

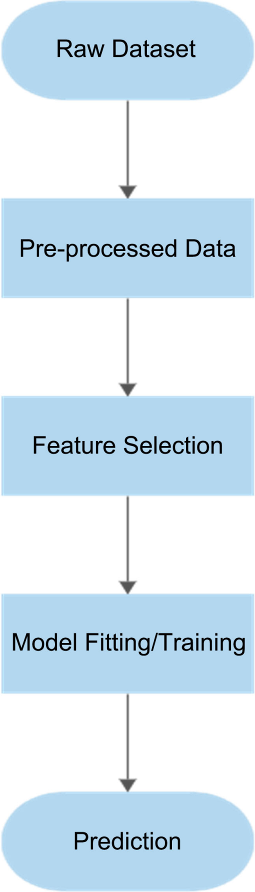

Figure 1. The basic machine learning topology.

until it can distinguish the underlying patterns between the input data and the output labels thus, enabling the ML model to yield accurate results when presented with validation data. Supervised learning is good at classification and problems, such as determining if an incoming email should be classified as spam or a normal mail. In addition, the goal is for the model to understand the data within the context of a particular question. Supervised learning algorithms can use what has been fed to them and learned in the past by applying labeled outputs to predict future events from data that has not been generated yet (Tohidi & Rustam, 2020). Examples of these algorithms include Perceptron, Decision Tree, Random Forest, K-Nearest Neighbour, Logistic Regression, Support Vector Machine (SVM), etc.

- Unsupervised ML (USML) algorithms are used when the training data do not have labeled output (Burns, Laskowski, & Tucci, 2022). Here, the main goal is to understand patterns and clusters within the unlabeled dataset. A user may not specify the appropriate output, but algorithm's ability to discover similarities and differences in the unlabeled data make it the ultimate answer for fact-finding data analysis, cross-selling stratagems, customer breakdown, and image identification. Examples of Unsupervised ML algorithms include Elbow algorithm, K-means, and Hierarchal Algorithm.

- Reinforcement learning algorithms interact with their environment by producing events and receiving reprimands or incentives. The backbone of this type of learning is the system agents that interact with the environment and learn depending on if a reward or reprimand is received at the end of an action. Trial and error search and delayed reward are the most important features of these algorithms. They allow systems agents to automatically determine the ideal behavior in a particular context in order to maximize its performance quality. Simple reward feedback is known as the reinforcement signal. Examples of this type of se learning are Q-learning and Markov Decision Process.

3. GIS and Machine Learning Applications

3.1. Applications in Infrastructure/Urban Development

Several factors can lead to a cost overrun in roadway projects ranging from the project duration, project type, current physical conditions, and traffic volumes etc. A study examined the possible impacts socioeconomic, macroeconomic, and weather factors played in cost overruns considering that they bring in a spatially wide-spread effect on projects, with the use of Random Forest classifier (Yun, Kyeong, & Suyun, 2022). This led to the generation of a stable classification of cost overrun situations as either false positives or false negatives and these classifications were affixed into 2 separate maps to look out for spatial patterns that could imply some underlying spatial dependence or correlation, however no significant residual spatial patterns were observed. Accumulated Local Effects (ALE) were then plotted to show the individual effects the variables had in predicting cost overruns. With these plots, each feature was partitioned into 3 groups of high, moderate, and low. The factors were then spatially overlayed over maps of Florida with overall spatial distribution showing the factors that had a high chance of cost overrun.

Another research was carried out (Effati & Mahyar, 2022) to predict and map unsafe segments of rural multiple lane roads that lead to road crashes based on land-use and road accessibility factors. Land use, accessibility and historical crash data were utilized to serve as the Basic Spatial Units (BSU's) variables needed for a comparison between a Logistic Regression model and a Classification and Regression Tree (CART) model. Results from both models showed that commercial, residential, government, and institutional land uses alongside access roads were most significantly associated with an increase in pedestrian road dangers. The results from both models were also mapped and compared against historical land crash data.

Looking at Environmental Impact Assessment and the problem of noise pollution in highly populated urbanized areas, two different studies have tried to generate traffic noise models in an attempt to aid town planners and traffic modelling experts for traffic noise mapping. In one of the research (Adulaimi et al., 2021b), the authors measured noise levels (New Klang Valley expressway in

Malaysia) and collected traffic flow information (noise samples) from the study area. Three models: 1) Artificial neural network model (ANN), 2) Correlation-based feature selection with ANN, 3) Ensemble random forest with ANN; were applied to all datasets to estimate the continuous noise levels during specific peak hours of the morning, evening and night-time. Results from the model showed that the Ensemble RF-ANN model was the most efficient. Using the design of this model, the parameters were converted into a geodatabase with their spatial positions obtained with a Global Positioning System (GPS). Further conversion was carried out to convert these same parameters into rasters using an Inverse Distance weighted (IDW) interpolation. Based on this, prediction maps for the specific time frames were generated. The second study (Adulaimi et al., 2021a) used a similar methodology and slightly similar set of variables for the same study area but applied four models (using Python) with two being machine learning models (Decision tree and Random tree forest) and the other two being statistical regression models (Linear regression and Support Vector Regression algorithms). The Random Forest proved to be the most effective and successful. Results from the RF model and parameters were converted into a geodatabase to produce noise prediction GIS maps showing that the highest noise levels were concentrated near the expressway and the lowest levels were far from the express and primary road. It is however an interesting comparison between the four models considering that few to no assumptions need to be taken care of with the ML models. They are more flexible whereas statistical models could raise concerns of multi-collinearity and the normal distribution of residuals. The statistical models do also have their own advantages as they help to cut down and streamline the most significant independent variables that influence the overall model thus helping to reduce computational time and efficiency.

Another study (Lloyd et al., 2020) explored the use of building footprint polygon obtained from OpenStreetMap (OSM) and OSM-like building attribute data to classify the residential status of urban buildings in both and low- and medium-income countries. Due to the scarcity of available data, and the need to combine building footprint and label datasets from a variety of sources, a GIS workflow was used to merge both datasets into one. This does raise concerns about data loss in the process of merging considering that direct conversion techniques between the 2 data types are unavailable. As such, it has the potential to increase levels and confidence in final results. These datasets along with an OSM highway polyline and impervious cover surface values served as in input into a stacked generalization, ensemble, building classification model that employed the use of the Superlearner package within the R statistical program. The countries in question are the Diplomatic Republic of Congo and Nigeria. Predicted probability values of the final model were then recoded to a classification of 1 for residential and 0 for non-residential.

Water storage and availability serve as some of the major obstacles faced by countries in arid regions. As a result, the location and construction of dams are a top priority for such countries. A hybrid system combining GIS, Analytical Hierarchal Process (AHP) and Machine Learning (ML) was developed to identify the most suitable locations for dam sites construction in the city of Sharjah, UAE (Al-Ruzouq et al., 2019). Data from various data sources were collected and converted to thematic layers and maps, representative of the precipitation, Drainage Streams Density, Slope, Geomorphology, Geology, Total Dissolved soil and Major Fractures of the study area. The weighting process for each layer was based on the AHP and ML (weight determination is based on ground truth data about available groundwater). The machine learning algorithm did however employ the use of the Random Forest, Gradient Boosted Trees and Support Vector Machine techniques. Amongst the ML techniques, the Random Forest (RF) model had the best accuracy of 76.5%. The use of an AHP method although beneficial has its own shortcomings as the values within the Reciprocal matrix are highly dependent on the individual preferences of the experts consulted. It is therefore pertinent to ensure a clear delineation of the purpose of the research project that will aid in the selection of the right experts.

The process of mineral exploration is an intricate one. Accurate prediction of prospective mining sites as well as understanding the effects that mineral exploration has on the environment, are of interest to governments and its allocation of resources. Over time, GIS and Machine learning techniques have proven beneficial in this field as seen with the following research. Using SVM, ANN and RF models, a GIS-based mineral prospectivity system for mapping out Copper sites in the Tongling ore district of China was conducted (Sun et al., 2019). Remote sensing, geological, geophysical as well as geochemical data with respect to Copper mineralization and twelve maps were created representative of key factors to the presence of copper. These 12 predictor maps were merged to create an integral rasterised map of evidential features in the form of cells and these cells were applied in the ML models. The random forest model achieved the greatest predictive accuracy with superior values for sensitivity, negative predictive value, and kappa Index. This RF model was then used as a final predictive model with prospective Copper targets occupying 13.97% of the study area and capturing 80.95% of the known deposits. On another hand, Kopeć et al. (2020) analysed the impact of hard coal mining on the environment; specifically flooding. Hyperspectral imagery and Sentinel-2 imagery of the study area were used to calculate the Normalized Difference Vegetation Index (NDVI) and Modified Normalized Difference Water Index (MNDWI). The NDVI, MNDWI, as well as the terrain displacement data, groundwater depth, geological classification of soil, a Digital Elevation Model (DEM) of the study area, slope and exposure/aspect; were used to conduct a correlation analysis to identify the variables with the highest potential of being descriptive variables for flood occurrence. These same variables except for the MNDWI served as input data in the Random Forest supervised machine learning model for floodplain detection which achieved an accuracy of 75%.

Groundwater yield potential mapping was experimented on with the use of remote sensing and GIS based machine learning techniques was experimented on (Lee et al., 2020). The study utilized Frequency ratio (FR), Boosted Classification Tree (BCT) and an ensemble FR-BCT. The three machine learning models were applied to the training datasets constructed from nine topographic factors, two hydrological factors, forest type, soil material, land use, and two geological factors (all obtained and processed from remotely sensed aerial imagery from KOMPSAT-2 and -3). Overall, all the models showed good performance, but the ensemble FR-BACT model displayed improved accuracy by 6% over the other 2 models and accuracy levels of 87.75% and 81.49% respectively. Another study attempted to model urban growth in the Qingpu-Songjiang area of Shanghai, China with the use of use of 2 models: a Logistic regression Cellular automation (LogCA) and Machine Learning Cellular automata (MachCA) (Feng, Yan, & Michael, 2016). Two Landsat images from 1992 and 2008 as well as a digital topographic map of the study area were georectified and used to produce classified land use maps (non-urban and urban). The MachCA consists of 3 modules: a Least Squares Support Vector Machine (LS-SVM), Land Use Change Decision rules and Map Visualization. Seven variables in total were used to measure urban growth and applied to both models. The MachCA model was finally implemented with the 16th iteration as that was gave the highest accuracy of 81.2%. The MachCA model had a higher overall performance with a lower proportion of missed changes of state from non-urban to urban.

3.2. Applications in Health

Location has proven to be an essential component into the better understanding of the health of individuals as well as a population and this has led to the growth of the merge between geography/geographic information science and artificial intelligence; commonly referred to as GeoAI. Here, we see applications of machine learning prediction models and geospatial software in the solving of real-world health issues ranging from public, precision medicine to even smart health cities (Kamel et al., 2019). Applying a multi-criteria decision analysis technique, the risk factors for communicable diseases were examined with an emphasis on vector-borne diseases (for example malaria and dengue fever) (Devarakonda et al., 2021). The variables used for this study were divided into global (temperature and precipitation) and local factors (human mobility, environmental and socio-economic). Shannon's entropy was used to calculate objective weights among the criteria set based on the amount of information within each dataset. It would have been an interesting observation to see the combination of Shannon's entropy with an AHP weighting system or a Weighted Linear Combination (WLC) to designate the weights for each identified variable. The decision criteria after applying the unbiased weights were aggregated using an Artificial Neural Network (ANN) to create risk levels and risk zones were classified using a multilayer perceptron (MLP). The study then analyzed the relationship between population density, education, employment and the final risk classes from the ANN model.

Results showed that the highest risk was along vegetation, wetlands, and water bodies.

The use and combination of remote sensing data with GIS and machine learning models to understand the relationship between the socio-physical conditions in Dak Nong province (Vietnam) and the reported cases of malaria to produce malaria susceptibility distribution maps were explored (Bui et al., 2019). Digital Elevation Model (DEM) data was used to calculate the Slope and Aspect of the study area while Landsat 8 imagery was used to calculate the Normalized Difference Vegetation Index (NDVI). Land use type, distance to roads and residential areas, distance to rivers, temperature, rainfall and elevation were also used as parameters for the machine learning models. For this study, 6 models were used: Artificial Neural Network (ANN), J48, Support Vector Machine (SVM), AdaBoost, Bagging and RandomSubSpace. Each model was evaluated based on the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) and Kappa statistics. The RandomSubSpace model outperformed the other models based on the evaluation parameters and the results were imported into an ArcGIS geodatabase to create the malaria susceptibility maps.

The use of Twitter data for real-time monitoring of events has certainly increased of time with (Allen et al., 2016) implementing a framework capable of monitoring the influenza outbreak in 31 of the most populated cities in the United States with the use of GIS and a Support Vector Machine (SVM) model. To aid the visual exploration of the data obtained from Twitter, a web map was developed by the authors for the display of the search locations and intensity of occurrence related to the keyword's "flu" and "influenza". In order to filter the data and reduce the impact of noise, a Support Vector Machine (SVM) was used, and the model attained a precision score of 0.671, a recall score of 0.949, and a resulting F1 score of 0.786; indicating that the model was able to classify most of the valid tweets correctly. A Pearson coefficient was then run with the results from the model per city and data on national, regional and local influenza-like illness (obtained from the CDC).

Lower respiratory infection (LRI) is one of the major causes of death in the United States of America but little to no research had been done to understand the relationship between the underlying factors and geographic variation. A study conducted (Abolfazl et al., 2020) took into consideration climatic, topographic, demographic, and socio-economic factors across the country at the county level and five machine learning models: logistic regression (LR), Random Forest (FR), Gradient Boosting Decision trees (GBDT), K-nearest neighbours (KNN) and Support Vector Machine (SVM) were used to detect the presence and/or absence of hotspots for elevated age adjusted Lower respiratory infection. A Moran's I and Getis-Ord General G spatial statistic were carried out in the ArcGIS software to examine the extent to which the nearby counties had similar levels of LRI mortality rates. The GBDT was the most accurate with the highest values for both precision and recall with an F1 score of 85%. Hotspots were identified in the earlier years (1980-1985) but they however reduced from 1990 to 2000.

The spread of the Coronavirus (COVID-19) gave way to a lot of research centered on case and death predictions with a particular study (Khan et al., 2021), attempting to predict when the number of reported cases would stop rising in India, to aid policy decision-makers ease lockdown restrictions. A Gaussian Process Regression (GPR), Support Vector Machine (SVM) and a Decision Tree were implemented for this. The performance of the models was evaluated based on the values of the coefficient of determination (R2), mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE). Results from the models showed that the GPR model was the most optimal with an RMSE score of 124.38 and the highest coefficient of determination value (R2 = 0.95). Based on the forecasted attributes for each model, a cumulative score of the region was calculated using a Multi-criteria decision-making technique and served as the criticality index for the classification of regions into low-risk clusters, moderate risk, and high risk. Projection maps based on the prediction results of the GPR were created for all 5 parameters of Daily positive cases, population, population density, deceased cases and recovered cases for the 4 districts in the country. Criticality index maps for the districts were also made for the visual depiction of the high-risk, moderate-risk and low-risk clusters.

3.3. Applications in Flood Prediction

Floods are one of the most destructive natural disasters and they trigger massive destruction to human life, agriculture, buildings, and infrastructure. Therefore, due to these adverse effects, there is a need to be very preemptive about developing precise mapping of areas susceptible to floods and further plan for sustainable flood risk management focusing on prevention, protection, and preparedness (Stefanidis & Stathis, 2013). Flood prediction models are significant for hazard assessment and extreme event management. Robust and accurate prediction contribute highly to water recourse management strategies, policy suggestions and analysis, and further evacuation modeling (Meyer, Scheuer, & Haase, 2009). Recent studies have even sought out ways to automate the process of flood susceptibility modeling in a way that the selected criteria/variables serve as input into a single tool in an attempt to reduce computational time (Ekeanyanwu et al., 2022).

Amongst the various research on flood prediction using machine learning algorithms, a study used random forest (RF) and Bayesian generalized linear model to determine the spatial patterns of flood susceptibility in the present and future for the Tajan watershed (Avand, Moradi, & Ramazanzadehlasboyee, 2021). Twelve geophysical and anthropogenic factors that affect flood risk were used to determine areas that are more likely to flood. Using the Area under the curve (AUC) metric, evaluations showed that the RF model is more accurate and can be used to accurately determine areas susceptible to flood in the location considered for the research. In the bid to create a flood risk index and hotspots analysis in the city of Lisbon, a flood prediction system was developed using a combination of Machine Learning classifiers namely: Support vector machines, Random Forest and Logistic Regression (Motta, Neto, & Sarmento, 2021). The ML model was combined with GIS technique for Hotspot Analysis. The implementation was designed to be used as an effective tool for urban management and resilience planning. Using several metrics, the best performing Machine Learning model was the Random Forest. To further augment the capabilities of the Machine Learning model, a GIS model was developed to find areas with higher likelihood of being flooded under critical weather conditions. Therefore, hotspots were defined for the entire city of Lisbon given the observed flood history. The results acquired from the RF model and the Hotspot analysis were then combined to create a flood risk index. An interesting perspective for further exploration in addition to existing research would be the use of satellite imagery and application of a cellular automata model to flood risk exposure in line with the specific set of pre-determined criteria and compare results to determine the accuracy and level of dependence on these results.

The use of ensemble machine-learning-based geospatial approach for flood risk assessment using multisensory remote-sensing data and GIS has also been explored where the ensemble method was used to create flood probability indices in Malaysia (Mojaddadi et al., 2017). In order to achieve this, Frequency ratio and Support Vector Machine ML models were combined in an ensemble format to produce a flood hazard map. The results showed that ensemble learning is an effective way for flood risk management.

Due to the significant effects of flash floods worldwide, a study assessed Flash-Flood Susceptibility using Multi-Criteria Decision Making and ML supported by Remote Sensing and GIS Techniques (Costache et al., 2019). The main purpose of the study was to gauge the efficiency of the Analytical Hierarchy Process (AHP), k-Nearest Neighbors (kNN), K-Star (KS) algorithms and their ensembles in flash-flood susceptibility mapping. The two stand-alone models and their ensembles were trained separately using data from the areas affected in the past by torrential phenomena which were identified using remote sensing techniques. Receiver Operating Characteristics (ROC) Curve was the main metric employed in the validation of results of the standalone models and their ensembles. The highest performance, in terms of success rate, was reached by the kNN-AHP ensemble model. This study provided results which are applicable for improving the flashflood forecast and warning activities. In another study (Costache et al., 2020), Machine Learning was used to highlight the Correlation between the Land-Use/Land-Cover Changes and Flash-Flood Potential changes in Zăbala catchment (Romania) between 1989 and 2019. The assessment of potential correlation was carried out with a multilayer perceptron (MLP) neural network. By ensuring that the land-use/land-cover change indicator, as well as the relative evolution of the flash flood potential index, was included in a geographically weighted regression (GWR), the study was able to prove that the land-use/land-cover changes were highly correlated with the changes that occurred in flash-flood potential. From extended research, shows the need for the implementation of more Geocompuational techniques for flood modeling considering that GIS has development so much over time to handle modeling solution rather than serving the purpose of just map-making. In as much as Geocomputation is an emerging branch of GIS, it promises great results in the world of modeling and computation.

3.4. Applications in Groundwater Detection and Contamination

Groundwater is the main source of water in many parts of the world (Kurwadkar, Kanel, & Nakarmi, 2020). Therefore, detecting and preserving groundwater quality is of critical concern. In line with being able to model groundwater potential, a study (Arabameri et al., 2021) mapped the groundwater potential (GWP) with a new hybrid model combining random subspace (RS) with the multilayer perception (MLP), naïve Bayes tree (NBTree), and classification and regression tree (CART) algorithms. This novel ensemble learning was introduced with goal of determining the possible distribution of groundwater without the need for more involved modeling efforts. In such a research case where the aim was to involve less modeling techniques, the use of a statistical regression also seems appropriate to take into consideration when comparing the above-mentioned ML techniques. The hybrid MLP-RS model achieved high validation scores and indicated that slope, elevation, TRI and HAND are the most important predictors of groundwater presence. A novel method was suggested in another study (Al-Mayahi, Al-Abadi, & Fryar, 2021) for the spatial delineation of groundwater contamination in aquifers specifically focusing on the Dammam Formation in the southern and western deserts of Iraq. Three machine learning classifiers; backpropagation multi-layer perceptron artificial neural networks (ANN), support vector machine with radial basis function (SVM-radial), and random forest (RF) with GIS, were used to map the probability of contamination in this aquifer. The three models had excellent goodness-of-fit with being over 90%, however, the ANN outperformed the other two models therefore proving that Deep learning models can be used to create guides for drilling uncontaminated wells of groundwater.

Another method for predicting the vulnerability of groundwater contamination using the GIS DRASTIC method and machine learning classifiers was introduced (Khan, Liaqat, & Mohamed, 2022). The extracted point values from a grid in the Al Khatim study area of United Arab Emirates were classified based on nitrate concentration at a particular threshold and divided into classes. Using four machine learning Algorithms which were Random Forest, Support Vector Machine, Naive Bayes and C4. Five ML models were trained and developed, using several features which includes depth to water (D), recharge (R), aquifer media (A). Accuracy showed the model developed by Random Forest gained highest accuracy. Groundwater vulnerability maps were developed from machine learning classifiers and were compared with base method of DRASTIC index. Comparison proved that machine learning is an efficient tool to access, analyze and map groundwater vulnerability.

3.5. Applications in Erosion Modeling and Prediction

Erosion is a disturbing occurrence which affects many places in the world today. A number of studies have been undertaken to study this process and to predict how several places are susceptible to erosion and also how these methodologies can be applied to other places in a broader worldwide effort to reduce the negative effects of erosion on communities.

A study was carried out to assess the performance of ML models while using different accuracy measures in determining susceptibility to gully erosion (Garosi et al., 2019). It involved using four ML models; Generalized additive model (GAM), support vector machine (SVM), Naïve Bayes (NB), and Random Forest (RF) models to create a gully erosion susceptibility map (GESM) in Hamedan, Iran. The functional relationships between gully erosion and controlling factors were evaluated using these models and several metrics such as; 10-fold cross-validation based on efficiency, Kappa coefficient, receiver operating characteristic curve (ROC), mean absolute error (MAE), and root mean square error (RMSE) in order to determine the best model. At the end of the study Random Forest model showed the highest predictive performance and thus proving that ML models can be used to build stable and accurate GESM depending on how they are calibrated and validated. Another study was conducted (Anh Nguyen & Chen, 2021) to analyze soil erosion depth using ML and GIS techniques. In order to achieve this fit, the soil erosion depth of a typical watershed in Taiwan was studied and modeled. Feature selection was performed using the Boruta algorithm and then the machine learning models, including the random forest (RF) and gradient boosting machine (GBM), were used to create prediction models validated by erosion pin measurements. The results show that GBM achieved the best result using the root mean square error (RMSE) and Nash-Sutcliffe efficiency (NSE) metrics. At the end of the study, the maps of soil erosion depth were created for conservation planning and mitigating future soil erosion. GIS-Based ML models have been used for erosion susceptibility mapping in a selected region in Iran (Lei et al., 2020). The gully erosion susceptibility assessment was performed using four ML techniques: credal decision trees (CDTree), kernel logistic regression (KLR), random forest (RF), and best-first decision tree (BFTree). Twelve gully erosion conditioning factors, including topographic, geomorphological, environmental, and hydrologic factors, were selected to estimate gully erosion susceptibility. The area under the ROC curve (AUC) was used to estimate the performance of the models and this showed that the RF model had the best performance. Therefore, further proving that ML models such as RF and SVM can be used to accurately map gully erosion susceptibility in other prone areas, hence ensuring their reproducibility.

3.6. GIS and Machine Learning Applications in Landslide Susceptibility Prediction (LSP)

Landslide is one of the many natural disasters plaguing the earth. It is a complex natural phenomenon particularly common in mountainous and hilly areas (Lee et al., 2004; Yilmaz, 2010; Pham et al., 2016). Landslide can be sudden, irreversible, and disastrous. Statistics have shown that it has caused more damage to life within the last decade. Several researchers and agencies have sought ways to curtail its suddenness through adequate management plans (Guzzetti et al., 2012) with much research still ongoing (Chen et al., 2018; Rabby, Hossain, & Abedin, 2020). It is believed that if landslide occurrence can be predicted, we can better account for it to prevent loss to humans, properties, and the environment. In addition, it will help with urban development and planning (Guzzetti et al., 2012). The technique for predicting the possibility of landslide occurrence in a geographic location is known as Landslide Susceptibility Prediction (LSP) (Chang et al., 2020; Rabby et al., 2020).

LSP is a very complicated process that works with past data to unveil future possibilities of landslide occurrence (van Westen, van Asch, & Soeters, 2006). Past data required are usually inventory maps acquired with GIS and site-related landslide causal factors. These causal factors could be internal or external related. Internal factors include lithography while external factors are related to human activities that could cause a landslide. Methods for LSP can be qualitative or quantitative depending on analysis (Juliev et al., 2019). Quantitative methods comprise statistical and deterministic theories while qualitative techniques are usually formed off expert opinion such as Boolean and fuzzy logic (Abella & Van Westen, 2007; Carrara et al., 1999). The heuristic method can be classified as a semi-quantitative method which has also been used in LSP. However, researchers have highlighted several difficulties with these methods especially due to their high subjectivity and simple linear approach in a non-linear area such as LSP.

Another study (Xiao et al., 2019) compared a ML model (Random Forest) with 3 statistical models: frequency ratio, certainty factor, and index of entropy (IOE) for Chongqing, China. The landslide inventory map used was an aggregate of previous detailed geotechnical investigation reports, field surveys, and aerial images. Correlation analysis was used to determine the most important causal factors in the study area. The factors selected include aspect, slope, topography wetness index (TWI) and stream power index (SPI); 15 variables in total. Results obtained showed that the ML model had superior performance to the statistical models. This agrees with recent ideas that statistical models are also subjective when compared to ML techniques.

Over the past years, LSP has evolved high-speed computing power and easy access to data have favored the use of ML and data mining techniques. The use of GIS and Remote sensing can help generate better data and ML models can consider the nonlinear nature of landslides. In addition, it can give mathematically verifiable and accurate results following some standard guidelines. Several ML models such as KNN, SVM, LRM, ANN, DT, XGBoost, RF, and NBT have been used for LSP.

Using 222 susceptible landslide areas (70% training and 30% validation), (Chen et al., 2018) compared four ML models namely RF, BN, RBF classifier, and LMT to assess the best for LSP modelling in a study area in Chongren county, China. The Information Gain technique was used for selecting the landslide conditioning factors which include lithology, SPI, distance to river, distance to roads, rainfall, sediment transport index (STI), normalized difference vegetation index (NDVI), aspect, and TWI. These models were compared based on statistical measures and receiver operating characteristics (ROC). LMT and RBF classifier had better results when the AUC was considered but had considerable variation in statistical measure. However, RF had an overall better result based on statistical measures and ROC. This agrees with a similar study carried out by researchers in another study (Bai, Liu, & Liu, 2021) in the Chongqing Area of China using four algorithms: RF, SVM, multilayer perceptron, and logistic regression. The study area consists of 581 landslide points divided into datasets of 70% training and 30% validation data. ROC and AUC are used as performance evaluation tools for comparing the algorithms. Based on the study area and selected conditioning factors, SVM, RF, and Multilayer perceptron gave good results. However, RF gave a better result with 0.848 and 0.822 for the training and test data respectively.

In contrast, a different study (Nsengiyumva & Valentino, 2020) favored NBT as a better algorithm than RF. The study compared NBT, RF, and LMT algorithms using the upper Nyabarongo Catchment of Rwanda as a case study. Using 15 conditioning factors selected by information gain (IG) technique with a detailed inventory map made from several sources including actual field surveys (for about 11 months), government agencies, and websites, they arrived at the conclusion. The average length of the landslide used was about 67 m, with an average extension of about 473 m^2. The data set was randomly split into 75% training and 25% validation data. The NBT had superior measurements of 82.4% for AUC values with 0.799, 0.745, and 0.301 as accuracy, precision, and RMSE values respectively.

Similarly, in a study of three Upazila areas (Rangamati Sadar, Kaptai, and Kawkhali) of Bangladesh, (Rabby, Hossain, & Abedin, 2020) showed that XGBoost was a superior ML model to KNN and RF. He compared the three models based on the highest area under the curve (AUC) with XGBoost having 95.27% and 90.63% in success and prediction rates respectively. Although there were some shortcomings recorded due to the lack of use of critical factors such as moisture content of the soil and its permeability, their study produced a similar result as with other research done in that area (Sifa et al., 2020). New grounds were explored by comparing the performance of SML and USML in LSP with data gotten through GIS and remote sensing (Chang et al., 2020). The SML used were SVM and CHAID, while K-means and Kohonen were the representative USML. The study area consisting of 446 landslide points was Ningdu county in China. After careful analysis of aerial images, government reports, and digital terrain models, 12 conditioning factors were selected including NDBI, slope aspect, TWI, and lithology. SVM algorithm had the best performance of the 4 ML models considered. In addition, SML was noted to have an inherent advantage due to their use of training data sets. Nonetheless, their accuracy and efficiency in prediction could be hindered by small training data. USML has the advantage of a simple modelling approach and good scalability which favors its use.

The shortcomings in conventional GIS-based ML algorithms have been acknowledged (Zhu et al., 2022). The researcher introduced the sparse feature extraction network (SFE+) for use in landslide prediction. This has the advantage of preventing overfitting while mining nonlinear features and can utilize lifetime sparsity to improve sparse details. The use of SFE+ has recorded positive effects in other fields including face recognition and image classification. The selected traditional algorithms (SVM, LR, and SGD) were also compared with the SFE+ modified algorithm (SFE-SVM, SFE-LR, and SFE-SGD). The SFE models outperformed others with the SFE-SVM showing the best performance.

ML in combination with GIS and remote sensing has been the backbone in LSP in recent times. ML models have achieved better results especially due to increased computing power and data mining. Several ML models can give reasonable results in the susceptibility prediction of landslides but the selection of ML models to use for a particular study area has been highly subjective and left largely to the decision of researchers. However, some ML models seem to be better performers than others. There is no consensus on the best model to use yet nor the best conditioning factors because each study area is unique. Landslide is a very complex phenomenon that still requires a lot of research to be done.

4. Conclusion

Merging GIS and ML offers a potential mechanism to reduce the cost of analysis of spatial information by decreasing the amount of time spent on data interpretation. This integration allows the interpretive outcome from a small area to be transferred to a larger, geographically similar area, without the extra time and expense of putting geographers in the field for a time sufficient to cover geographical area.

Applications in infrastructural and urban development are still being refined considering the gap in data availability (for example, obtaining data on height of buildings, patch density of urban forms, etc.) for analytical processes as well as the conversion of data between the different software. However, reasonable progress has been made towards the modeling and prediction of urban growth and infrastructural development for governmental and urban planning projects including traffic noise pollution prediction models, examining the effects of rural land use and accessibility to pedestrians and bicycle riders to ensure their safety and even the cost effects on specific variables on major roadway/construction projects. In addition, the Coronavirus (COVID-19) pandemic raised the need for more accurate and efficient prediction models pertinent to the containment efforts as well as resource allocation within cities, states and/or countries. Research using Machine Learning and GIS methods (in conjunction with statistical methods) in the health sector is still at its very early stages and as such possesses the potential to grow into a highly valuable niche of researchers. A lot of the current literatures have expressed concerns regarding the conversion and loss of data across both systems as a result of transformations, as well as the geographical scale of the analysis. Future studies can factor in these concerns for more accurate prediction models.

The real potential of ML in flood, erosion and groundwater is also not sufficiently developed yet. How, these fields intersect in analytical discussions. At the same time, most GIS applications which are desirable for ML implementation, are driven by conventional approach and standard tools of commercial GIS packages.

Author Contributions

Conceptualization: Chikodinaka V. Ekeanyanwu, Inioluwa F. Obisakin, Precious T. Aduwenye; Methodology: Chikodinaka V. Ekeanyanwu, Inioluwa F. Obisakin, Precious T. Aduwenye; writing—original draft preparation: Chikodinaka V. Ekeanyanwu, Inioluwa F. Obisakin, Precious T. Aduwenye, Nathaniel Dede-Bamfo; writing—review and editing: Chikodinaka V. Ekeanyanwu, Inioluwa F. Obisakin, Precious T. Aduwenye, Nathaniel Dede-Bamfo; supervision: Nathaniel Dede-Bamfo. All authors have read and agreed to the published version of the manuscript.”.

Conflicts of Interest

The authors declare no conflict of interest regarding the publication of this paper.

References

Abella, E. A. C., & Van Westen, C. J. (2007). Generation of a Landslide Risk Index Map for Cuba Using Spatial Multi-Criteria Evaluation. *Landslides, 4,* 311-325. https://doi.org/10.1007/s10346-007-0087-y

Abolfazl, M., Behrooz, V., Shreejana, B., Hopkins Laura, C., Swagata, B., & Behzad, V. (2020). Predicting the Hotspots of Age-Adjusted Mortality Rates of Lower Respiratory Infection across the Continental United States: Integration of GIS, Spatial Statistics and Machine Learning Algorithms. *International Journal of Medical Informatics, 142,* Article ID: 104248. https://doi.org/10.1016/j.ijmedinf.2020.104248

Adulaimi, A. A. A., Pradhan, B., Chakraborty, S., & Alamri, A. (2021a). Traffic Noise Modelling Using Land Use Regression Model Based on Machine Learning, Statistical Regression and GIS. *Energies, 14,* Article No. 5095. https://doi.org/10.3390/en14165095

Adulaimi, A. A., Pradhan, B., Chakraborty, S., & Alamri, A. (2021b). Developing Vehicular Traffic Noise Prediction Model through Ensemble Machine Learning Algorithms with GIS. *Arabian Journal of Geosciences, 14,* Article No.1564. https://doi.org/10.1007/s12517-021-08114-y

Allen, C., Tsou, M.-H., Aslam, A., Nagel, A., & Gawron, J.-M. (2016). Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. *PLOS ONE, 11,* Article ID: e0157734. https://doi.org/10.1371/journal.pone.0157734

Al-Mayahi, M., Alaa, H., Al-Abadi, M., & Fryar, A. E. (2021). Probability Mapping of Groundwater Contamination by Hydrocarbon from the Deep Oil Reservoirs Using GIS-Based Machine-Learning Algorithms: A Case Study of the Dammam Aquifer (Middle of Iraq). *Environmental Science and Pollution Research, 28,* 13736-13751. https://doi.org/10.1007/s11356-020-11158-4

Al-Ruzouq, R., Abdallah, S., Abdullah Gokhan, Y., AlaEldin, I., Sunanda, M., Mohamad Ali, K., & Gibril Mohamed, B. A. (2019) Dam Site Suitability Mapping and Analysis Using an Integrated GIS and Machine Learning Approach. *Water, 11,* Article No. 1880. https://doi.org/10.3390/w11091880

Anh Nguyen, K., & Chen, W. (2021). DEM- and GIS-Based Analysis of Soil Erosion Depth Using Machine Learning. *International Journal of Geo-Information, 10,* Article No. 452. https://doi.org/10.3390/ijgi10070452

Anselin, L. (2000). Part 2 The Link between GIS and Spatial Analysis: GIS, Spatial Econometrics and Social Science Research. *Journal of Geographical Systems, 2,* 11-15. https://doi.org/10.1007/s101090050023

Arabameri, A., Chandra Pal, S., Rezaie, F., Asadi Nalivan, O., Chowdhuri, I., Saha, A., Lee S., & Moayedi, H. (2021). Modeling Groundwater Potential Using Novel GIS-Based Machine-Learning Ensemble Techniques. *Journal of Hydrology: Regional Studies, 36,* Article ID: 100848. https://doi.org/10.1016/j.ejrh.2021.100848

Avand, M., Moradi, H., & Ramazanzadehlasboyee, M. (2021). Using Machine Learning Models, Remote Sensing, and GIS to Investigate the Effects of Changing Climates and Land Uses on Flood Probability. *Journal of Hydrology, 595,* Article ID: 125663. https://doi.org/10.1016/j.jhydrol.2020.125663

Bai, Z., Liu, Q., & Liu, Y. (2021). Landslide Susceptibility Mapping Using GIS-Based Machine Learning Algorithms for the Northeast Chongqing Area, China. *Arabian Journal of Geosciences, 14,* Article No. 2831. https://doi.org/10.1007/s12517-021-08871-w

Bui, Q.-T., Nguyen, Q.-H., Manh, P. V., Hai, P. M., & Tuan, T. A. (2019). Understanding Spatial Variations of Malaria in Vietnam Using Remotely Sensed Data Integrated into GIS and Machine Learning Classifiers. *Geocarto International, 34,* 1300-1314. https://doi.org/10.1080/10106049.2018.1478890

Burns, E., Laskowski, N., & Tucci, L. (2022). *What Is Artificial Intelligence (AI)?* SearchEnterpriseAI (Blog). https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence

Carrara, A., Guzzetti, F., Cardinali, M., & Reichenbach, P. (1999). Use of GIS Technology in the Prediction and Monitoring of Landslide Hazard. *Natural Hazard, 20,* 117-135. https://doi.org/10.1023/A:1008097111310

Chang, Z. L., Zhen, D., Zhang, F., Huang, F. M., Chen, J. W., Li, W. B., & Guo, Z. Z. (2020). Landslide Susceptibility Prediction Based on Remote Sensing Images and GIS: Comparisons of Supervised and Unsupervised Machine Learning Models. *Remote Sensing, 12,* Article No. 502. https://doi.org/10.3390/rs12030502

Chen, W., Peng, J. B., Hong, H. Y., Shahabi, H., Pradhan, P., Liu, J. Z., Zhu, A.-X., Pei, X. J., & Duan, Z. (2018). Landslide Susceptibility Modelling Using GIS-Based Machine

Learning Techniques for Chongren County, Jiangxi Province, China. *Science of the Total Environment, 626,* 1121-1135. https://doi.org/10.1016/j.scitotenv.2018.01.124

Costache, R., Pham, Q. B., Corodescu-Roşca, E., Cîmpianu, C., Hong, H. Y., Linh, N. T. T., Chow, M. F. et al. (2020). "Using GIS, Remote Sensing, and Machine Learning to Highlight the Correlation between the Land-Use/Land-Cover Changes and Flash-Flood Potential. *Remote Sensing, 12,* Article ID: 1422. https://doi.org/10.3390/rs12091422

Costache, R., Pham, Q. B., Sharifi, E., Linh, N. T. T., Abba, S. I., Vojtek, M., Vojteková, J., Nhi, P. T. T., & Khoi, D. N. (2019). Flash-Flood Susceptibility Assessment Using Multi-Criteria Decision Making and Machine Learning Supported by Remote Sensing and GIS Techniques. *Remote Sensing, 12,* Article No. 106. https://doi.org/10.3390/rs12010106

Devarakonda, P., Ravi, S., Nobrega Rodrigo, A. A., & Wu, J. H. (2021). Application of Spatial Multicriteria Decision Analysis in Healthcare: Identifying Drivers and Triggers of Infectious Disease Outbreaks Using Ensemble Learning. *Journal of Multi-Criteria Decision Analysis, 29,* 23-26. https://doi.org/10.1002/mcda.1732

Effati, M., & Mahyar, V. S. (2022). Examining the Influence of Rural Land Uses and Accessibility-Related Factors to Estimate Pedestrian Safety: The Use of GIS and Machine Learning Techniques. *International Journal of Transportation Science and Technology, 11,* 144-157. https://doi.org/10.1016/j.ijtst.2021.03.005

Ekeanyanwu, C. V., Bose, P., Beavers, M., Yuan, Y. H., & Feranmi Obisakin, I. (2022). Modeling and Mapping Flood Hazard with a Flood Risk Assessment Tool: A Case Study of Austin, Texas. *Journal of Geographic Information System, 14,* 332-346. https://doi.org/10.4236/jgis.2022.144018

Feng, Y. j., Liu, Y., & Michael, B. (2016). Modeling Urban Growth with GIS Based Cellular Automata and Least Squares SVM Rules: A Case Study in Qingpu-Songjiang Area of Shanghai, China. *Stochastic Environmental Research and Risk Assessment, 30,* 1387-1400. https://doi.org/10.1007/s00477-015-1128-z

Garosi, Y., Sheklabadi, M., Conoscenti, C., Reza Pourghasemi, H., & Van Oost, K. (2019). Assessing the Performance of GIS-Based Machine Learning Models with Different Accuracy Measures for Determining Susceptibility to Gully Erosion. *Science of the Total Environment, 664,* 1117-1132. https://doi.org/10.1016/j.scitotenv.2019.02.093

Guzzetti, F., CesareMondini, A., Cardinali, M., Fiorucci, F., Santangelo, M., & Chang, K.-T. (2012) Landslide Inventory Maps: New Tools for an Old Problem. *Earth-Science Reviews, 112,* 42-66. https://doi.org/10.1016/j.earscirev.2012.02.001

Iyanda, A., Boakye, K., & Lu, Y. M. (2021). COVID-19: Evidenced Health Disparity. *Encyclopedia, 1,* 744-763. https://doi.org/10.3390/encyclopedia1030057

Juliev, M., Mergili, M., Mondal, I., Nurtaev, B., Pulatov, A., & Hübl, J. (2019). Comparative Analysis of Statistical Methods for Landslide Susceptibility Mapping in the Bostanlik District, Uzbekistan. *Science of the Total Environment, 653,* 801-814. https://doi.org/10.1016/j.scitotenv.2018.10.431

Kamel, B., Maged, N., Peng, G. C., & VoPham, T. (2019). An Overview of GeoAI Applications in Health and Healthcare. *International Journal of Health Geographics, 18,* Article No. 7. https://doi.org/10.1186/s12942-019-0171-2

Khan, F. M., Kumar, A., Puppala, H., Kumar, G., & Gupta, R. (2021). Projecting the Criticality of COVID-19 Transmission in India Using GIS and Machine Learning Methods. *Journal of Safety Science and Resilience, 2,* 50-62. https://doi.org/10.1016/j.jnlssr.2021.05.001

Khan, Q., Liaqat, M. U., & Mohamed, M. M. (2022). A Comparative Assessment of Modeling Groundwater Vulnerability Using DRASTIC Method from GIS and a Novel Classification Method Using Machine Learning Classifiers. *Geocarto International, 37,* 5832-5850. https://doi.org/10.1080/10106049.2021.1923833

Kopeć, A., Trybała, P., Głąbicki, D., Buczyńska, A., Owczarz, K., Bugajska, N., Kozińska, P., Chojwa, M., & Gattner, A. (2020). Application of Remote Sensing, Gis and Machine Learning with Geographically Weighted Regression in Assessing the Impact of Hard Coal Mining on the Natural Environment. *Sustainability, 12,* 9338. https://doi.org/10.3390/su12229338

Kulawiak, M., & Lubniewski, Z. (2014). SafeCity—A GIS-Based Tool Profiled for Supporting Decision Making in Urban Development and Infrastructure Protection. *Technological Forecasting and Social Change, 89,* 174-187. https://doi.org/10.1016/j.techfore.2013.08.031

Kurwadkar, S., Kanel, S. R., & Nakarmi, A. (2020). Groundwater Pollution: Occurrence, Detection, and Remediation of Organic and Inorganic Pollutants. *Water Environment Research, 92,* 1659-1668. https://doi.org/10.1002/wer.1415

Lazar, A., & Shellito, B. (2005). Comparing Machine Learning Classification Schemes—a GIS Approach. In *Fourth International Conference on Machine Learning and Applications (ICMLA'05)* (p. 7). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICMLA.2005.16

Lee, S. M., Hyun, Y. J., Lee, S., & Lee, M.-J. (2020). Groundwater Potential Mapping Using Remote Sensing and GIS-Based Machine Learning Techniques. *Remote Sensing, 12,* Article No. 1200. https://doi.org/10.3390/rs12071200

Lee, S., Ryu, J.-H., Won, J.-S., & Park, H.-J. (2004). Determination and Application of the Weights for Landslide Susceptibility Mapping Using an Artificial Neural Network. *Engineering Geology, 71,* 289-302. https://doi.org/10.1016/S0013-7952(03)00142-X

Lei, X. X., Chen, W., Avand, M., Janizadeh, S., Kariminejad, N., Shahabi, H., Costache, R., Shahabi, H., Shirzadi, A., & Mosavi, A. (2020). GIS-Based Machine Learning Algorithms for Gully Erosion Susceptibility Mapping in a Semi-Arid Region of Iran. *Remote Sensing, 12,* Article No. 2478. https://doi.org/10.3390/rs12152478

Lloyd, C. T., Sturrock Hugh, J. W., Leasure Douglas, R., Jochem Warren, C., Lázár Attila, N., & Tatem Andrew, J. (2020). Using GIS and Machine Learning to Classify Residential Status of Urban Buildings in Low and Middle Income Settings. *Remote Sensing, 12,* Article No. 3847. https://doi.org/10.3390/rs12233847

McLafferty, S. L. (2003). GIS and Health Care. *Annual Review of Public Health, 24,* 25-42. https://doi.org/10.1146/annurev.publhealth.24.012902.141012

Meyer, V., Scheuer, S., & Haase, D. (2009). A Multicriteria Approach for Flood Risk Mapping Exemplified at the Mulde River, Germany. *Natural Hazards, 48,* 17-39. https://doi.org/10.1007/s11069-008-9244-4

Mojaddadi, H., Pradhan, B., Nampak, H., Ahmad, N., & Abdul Halim bin, G. (2017). Ensemble Machine-Learning-Based Geospatial Approach for Flood Risk Assessment Using Multi-Sensor Remote-Sensing Data and GIS. *Geomatics, Natural Hazards and Risk, 8,* 1080-1102. https://doi.org/10.1080/19475705.2017.1294113

Motta, M., Miguelde, C. N., & Pedro, S. (2021). A Mixed Approach for Urban Flood Prediction Using Machine Learning and GIS. *International Journal of Disaster Risk Reduction, 56, Article ID:* 102154. https://doi.org/10.1016/j.ijdrr.2021.102154

Nsengiyumva, J. B., & Roberto, V. (2020). Predicting Landslide Susceptibility and Risks Using GIS-Based Machine Learning Simulations, Case of Upper Nyabarongo Catchment. *Geomatics, Natural Hazards and Risk, 11,* 1250-1277. https://doi.org/10.1080/19475705.2020.1785555

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. https://doi.org/10.1016/j.envsoft.2016.07.005

Rabby, Y. W.,, Md Belal, H., & Joynal, A. (2020). Landslide Susceptibility Mapping in Three Upazilas of Rangamati Hill District Bangladesh: Application and Comparison of GIS-Based Machine Learning Methods. *Geocarto International, 37,* 3371-3396.

Ray, S. (2019). A Quick review of Machine Learning Algorithms. In *2019 International Conference on MACHINE Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 35-39). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/COMITCon.2019.8862451

Sifa, S. F., Mahmud, T., Abdullah Tarin, M., & Haque, D. M. E. (2020). Event-Based Landslide Susceptibility Mapping Using Weights of Evidence (WoE) and Modified Frequency Ratio (MFR) Model: A Case Study of Rangamati District in Bangladesh. *Geology, Ecology, and Landscapes, 4,* 222-235. https://doi.org/10.1080/24749508.2019.1619222

Stefanidis, S., & Stathis, D. (2013). Assessment of Flood Hazard Based on Natural and Anthropogenic Factors Using Analytic Hierarchy Process (AHP). *Natural Hazards, 68,* 569-585. https://doi.org/10.1007/s11069-013-0639-5

Sun, T., Chen, F., Zhong, L. X., Liu, W. M., & Wang, Y. (2019). GIS-Based Mineral Prospectivity Mapping Using Machine Learning Methods: A Case Study from Tongling Ore District, Eastern China. *Ore Geology Reviews, 109,* 26-49. https://doi.org/10.1016/j.oregeorev.2019.04.003

Tohidi, N., & Rustamov Rustam, B. (2020). A Review of the Machine Learning in Gis for Megacities Application. In R. B. Rustamov (Ed.), *Geographic Information Systems in Geospatial Intelligence* (pp. 29-53). IntechOpen. https://doi.org/10.5772/intechopen.94033

Tsou, M.-H., Yang, J.-A., Lusher, D., Han, S., Spitzberg, B., Gawron, J. M., Gupta, D., & Li, A. (2013). Mapping Social Activities and Concepts with Social Media (Twitter) and Web Search Engines (Yahoo and Bing): A Case Study in 2012 US Presidential Election. *Cartography and Geographic Information Science, 40,* 337-348. https://doi.org/10.1080/15230406.2013.799738

van Westen, C., van Asch, T., & Soeters, R. (2006). Landslide Hazard and Risk Zonation—Why Is It Still so Difficult? *Bulletin of Engineering Geology and the Environment, 65,* 167-184. https://doi.org/10.1007/s10064-005-0023-0

Wankhede, C. (2022, January 4). *What Is Machine Learning and How Does It Work?* Android Authority (Blog). https://www.androidauthority.com/machine-learning-explained-3074635/

Wright, D. J., Goodchild, M. F., & Proctor James, D. (1997). GIS: Tool or Science? Demystifying the Persistent Ambiguity of GIS as “Tool” versus “Science”. *Annals of the Association of American Geographers, 87,* 346-362. https://doi.org/10.1111/0004-5608.872057

Xiao, T., Yin, K. L., Yao, T. L., & Liu, S. H. (2019). Spatial Prediction of Landslide Susceptibility Using GIS-Based Statistical and Machine Learning Models in Wanzhou County, Three Gorges Reservoir, China. *Acta Geochimica, 38,* 654-669. https://doi.org/10.1007/s11631-019-00341-1

Yilmaz, I. (2010). Comparison of Landslide Susceptibility Mapping Methodologies for Koyulhisar, Turkey: Conditional Probability, Logistic Regression, Artificial Neural Networks, and Support Vector Machine. *Environmental Earth Sciences, 61,* 821-836. https://doi.org/10.1007/s12665-009-0394-9

Yun, J. h., Rok, R. K., & Ham, S. (2022). Spatial Analysis Leveraging Machine Learning and GIS of Socio-Geographic Factors Affecting Cost Overrun Occurrence in Roadway Projects. *Automation in Construction, 133,* Article ID: 104007. https://doi.org/10.1016/j.autcon.2021.104007

Zhu, L., Wang, G. J., Huang, F. M., Li, Y., Chen, W., & Hong, H. Y. (2022). Landslide Susceptibility Prediction Using Sparse Feature Extraction and Machine Learning Models Based on GIS and Remote Sensing. *IEEE Geoscience and Remote Sensing Letters, 19,* 1-5. https://doi.org/10.1109/LGRS.2021.3054029