Performance and Feasibility of Machine Learning for Multi-hazard Humanitarian Forecasting

A literature survey

Ewa Smura[1]
Supervisor(s): Cynthia Liem[1]**, Marijn Roelvink**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

22.06.2025

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Ewa Smura
Final project course: CSE3000 Research Project
Thesis committee: Cynthia Liem, Marijn Roelvink, Jing Sun

Abstract

Natural disasters frequently cause casualties and property losses. Predicting and mitigating the impact of such threats is crucial to the work of humanitarian organizations. The interactions between hazards are best represented through a multi-hazard approach, and machine learning models are well suited for natural hazard prediction. This study presents a systematized literature survey of machine learning in multi-hazard disaster forecasting in the years 2019-2025, focusing on the used models and performance metrics, their applications and feasibility of use, as well as potential cross-applications. There is a wide variety of models and metrics used. The most commonly used models are random forest and support vector machine and the most prevalent performance metric is the ROC-AUC score. The machine learning models generally perform well, with AUC scores above 0.8, though patterns in performance are difficult to examine. Feasibility is defined here as readiness to be used in practice, and the models are rated in the factors that define it. Most of the articles are feasible. Consideration of cross-application is rare and should be extended. This research summarizes the main trends in the field of disaster forecasting, providing a clear reference point for other academics.

1 Introduction and Background

Natural hazards occur all across the world, often with devastating consequences to local communities, including the cost of human lives, severe economic impacts and damage to various infrastructure [1]. The ability to prepare for or avoid disasters at such scale is therefore very valuable, leading to the existence of the field of disaster forecasting. Early warning systems provide the ability to respond hours earlier [2], while susceptibility maps can guide long-term development and inform infrastructure decisions [3].

Most papers in the field of hazard forecasting focus on predicting a single hazard. However, the forecasting of single hazard threats is often not as accurate, as many regions are subject to multiple hazards with complicated interrelations [4]. To define the consideration of more than one hazard at once, including the potential interaction between them, the term “multi-hazard“ has been introduced in the 1992 Agenda 21 released by the United Nations Environment Programme [5], which calls among other things for “complete multi-hazard research“ as a part of disaster-management and human settlement planning in disaster-prone areas.

With the recent surge in development of machine learning models, they have often been applied to predict hazards due to their ability to process large amounts of data and find overarching patterns. However, the usage presents challenges, as the performance, accuracy and trustworthiness of the models can be doubted, in tandem with data availability and reliability. The main research question of the paper is: **What is the performance of machine learning models for multi-hazard disaster prediction and their feasibility of application in humanitarian forecasting?**

The field of machine learning in multi-hazard forecasting is relatively recent and has been growing quite rapidly in recent years [6], and the most recent survey was executed in 2019 [7]. Given this, the purpose of the current study is to conduct a literature survey of machine learning in multi-hazard humanitarian forecasting, with a focus on examining the models used, their performance, as well as their feasibility of practical use within the field of hazard forecasting, and particularly for the usage of humanitarian organizations.

The structure of the paper is as follows: Section 2 lays out the methodology of the research. Section 3 contains the results of the survey, together with the discussion and limitations. Section 4 examines the potential ethical implications of the research, while Section 5 contains the conclusions and the recommendations for future work.

2 Methodology

In order to answer the research question, a literature survey was executed following the SALSA approach [8]. The SALSA approach specifies four stages of executing a literature survey. The stages are search, appraisal, synthesis and analysis. Subsection 2.1 describes the search stage, detailing the strategies used to find the initial set of papers. Subsection 2.2 details the inclusion criteria used to evaluate the articles, as well as reports on the final scope of the survey. Subsection 2.3 introduces the sub-questions of the research question, explains their groupings and the reasoning behind them, and finally subsection 2.4 lays out the methods used in analyzing the gathered data and drawing conclusions. It is important to note as part of the methodology that no Large Language Models were used in any way in the course of creating this work.

2.1 Search

The first stage of the SALSA method, search, included defining the exact scope of the survey, defining key terms and building the exact search query, as well as collecting the records found through the query. The scope of the survey was decided by the last relevant literature survey done in the field, which took place in 2019 [7]. Given the literature published since then has not been included in a survey(save a more linguistically focused survey on the progress of the field done in 2023 [6]), and the time constraints for this work, it was decided the survey would consider academic literature published between 2019 and 2025.

The search query was also informed by the last survey, since it contained, in addition to other things, an overview of the most common terminology used in the field. The search engine that was used to find the papers was Scopus(CITE), as other search engines provided results that were either not relevant to the subject matter or too numerous without sufficient filtering options. The final search query used was the combination of the following two:

query 1: *(multi-hazard OR compound hazard) AND (forecasting OR humanitarian forecasting OR disaster prediction) AND (machine learning OR AI OR prediction OR accuracy)*, which individually yielded 47 results with a very high relevance rate.

query 2: *((earthquake* OR lightning OR (sea AND surge*) OR landslide* OR (extreme AND rainfall*) OR (extreme AND waves) OR (extreme AND wind*) OR (river AND flood*) OR (volcanic AND eruption*) OR (extreme AND temperature*) OR hail OR tornado* OR drought* OR conflict* OR displacement* OR migration* OR (food AND security) OR typhoon*) AND (humanitarian OR forecasting OR (disaster AND prediction)) AND ((multi-hazard) OR compound OR cascade OR multi-risk OR (multi AND hazard) OR interrelationship) AND ((machine AND learning) OR (artificial AND intelligence) OR ai))*, which individually yielded 176 results, with a medium relevance rate.

Note that the quotation marks in the queries have been omitted. The two queries were joined by an “AND“, together yielding a reasonably high relevance rate. The results were then filtered within the search engine to exclude articles published before 2019, as well as articles in languages other than English. The number of papers gathered was 144. Additionally, the snowballing method [9] was used throughout the course of the survey, gathering 36 articles in total, so the final number of papers before the appraisal stage was 180.

2.2 Appraisal

The next step was initial appraisal, which included defining detailed inclusion/exclusion criteria for the gathered papers and screening them to find the relevant literature. The first of the inclusion criteria was the purpose of the research to be the prediction, forecasting or development of an early warning system for a natural hazard. The second required that the paper include the development, usage or application of a machine learning model. Machine learning is defined here as a subfield of artificial intelligence, concerning the usage and development of statistical algorithms that are able to perform tasks without being explicitly programmed to do so. This includes more traditional machine learning algorithms such as logistic regression, support vector machine or random forest, but also the machine learning sub-field of neural networks and deep learning. The paper also had to be multi-hazard in nature. This meant more than one type of hazard was included within the article and in the application of the models used. Finally, the performance of the machine learning model used or developed in the study had to be reported on, including the declaration of the performance metric used and the numerical values of the performance according to said metric.

Additionally, as mentioned above, the exclusion criteria removed from consideration articles written in a language other than English and articles published before 2019. Using this set of criteria, the initial gathered papers were scanned and narrowed down into the final set of surveyed papers. As seen in Figure 1, of the 180 papers gathered initially, 148 were eventually rejected, an additional 4 could not be accessed, and 28 were read and included in the survey. An overview of all of the surveyed literature can be found in Appendix A.1.

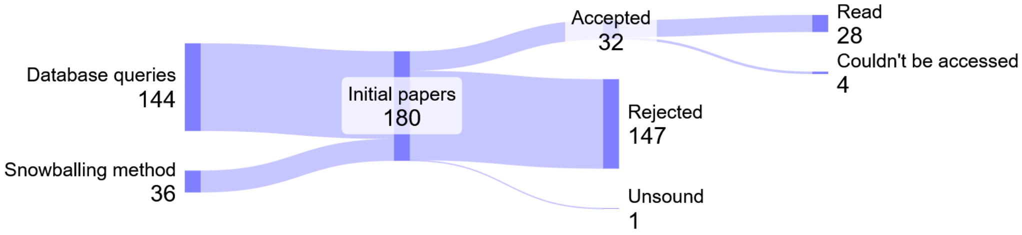

Figure 1: Sankey diagram of considered articles

2.3 Synthesis

This subsection details the synthesis stage, including the sub-questions of the main research question, their grouping and the reasoning behind those choices. The sub-questions were created to aid in answering the main research question, and are as follows:

1. What are the machine learning models used in the papers?
2. What are the metrics used to report on the performance of the models?
3. How does the choice of metric depend on model and domain?
4. What is the performance per metric, domain and model?
5. How to define and judge feasibility of practical application?

6. What are the intended practical applications of the models?
7. What factors influence the feasibility of the intended use of the models?
8. What are possible cross-applications of the models in humanitarian forecasting?
9. What is the feasibility of practical application of the models?

They can be grouped into four sub-groups. Firstly, sub-questions 1 through 3 were created to determine which machine learning models are most commonly used in the field of multi-hazard forecasting, which metrics are chosen to evaluate their performance and how those choices are influenced by each other or the specific hazards being predicted. This is relevant because one of the goals of this survey is to provide a general overview of the last years of machine learning in multi-hazard forecasting, and the popularity of models and metrics is relevant to that end. Additionally, the results of these questions help prioritize which model-metric combo's are most relevant to examine in the next sub-question. Secondly, sub-question 4 corresponds to analyzing the overall effectiveness and performance of the models detailed in the surveyed papers, as machine learning models that perform badly and cannot be relied upon cannot be used in practice. This sub-question is crucial, as it corresponds directly to the main research question, and therefore forms a sub-group of its own. The results again help inform the answers to the next sub-group, since performance is one of the factors of feasibility.

Sub-questions 5, 6, 7 and 9 form another group, which is focused on defining and attempting to evaluate feasibility of usage, meaning how practical and likely it is that the models detailed in the surveyed papers can be of use to the functioning of local governments and humanitarian organizations. Since feasibility is not a clearly defined concept, multiple smaller sub-questions were created to help make the process and reasoning clearer. Feasibility forms the other main component of the overall research question, and so these sub-questions form another sub-group. Lastly, sub-question 8 deals with the potential cross-applications of the models, meaning the possible ways a model developed for one scenario could be repurposed to work in another. While not directly taken from the main research question, this subject is of interest due to its connection to humanitarian action, as it is humanitarian organizations that are most interested in developing wider scale disaster prediction models that could cover their entire area of operation. As it is independent from the other sub-questions, it forms a group of its own.

2.4 Analysis

This subsection details the last stage, analysis [8]. It lays out the methods used in data collection and analysis as well as how the research questions were answered. The process was divided into several stages. During the initial pass over the articles, if the paper was accepted into the survey, information about the examined hazards, the machine learning models and the performance metrics used in the study were also written down in a note-taking app. Once the initial pass was finished and the final set of literature decided, this was compiled into the sets of all of the mentioned hazards, models or metrics.

Since the hazards ranged from general to very specific and as they are not the focus of the study, they were grouped into nine categories, which are presented in Table 1. The total list of models or metrics was not modified beyond merging synonyms, such as recall and sensitivity. As the last part of this step, tables of all of the articles and the hazards, models and metrics were created to aid in compiling the results.

Landslides	landslides, debris fall, rock fall, debris flow
Flood	flood, flash flood
Fire	wildfire, forest fire
Earthquake	earthquake, seismic
Drought	drought
Erosion	coastal erosion, gully erosion, soil erosion, land subsidence
Snow avalanche	snow avalanche
Storm	storm, lightning, wind events
Extreme precipitation	heavy rainfall, hail

Table 1: Categorization of hazards

Following the initial steps, all of the articles were read thoroughly. The relevant parts of the text were highlighted, and general notes were taken from each one, detailing the model performance, the final result, the intended uses as well as any mention of cross-application. It informed a general overview of the surveyed literature, enabling the author to decide on the feasibility factors, which are detailed in Section 3.3. To ensure fairness and avoid mistakes from memory, a last pass was done over the articles, assigning and noting the reasoning behind the given scores in another Excel table. Lastly, once all of the data had been gathered into detailed tables, they were used to derive statistics and identify connections or trends, both by hand and through the use of aggregation functions, for example to count the number of appearances of each model type. The findings were compiled into a concise, readable format and presented in section 3.

3 Results and discussion

This section contains the explanation and analysis of the results obtained through the literature survey, as well as their discussion. The subsections correspond to the sub-question groupings detailed above. Subsection 3.1 gives an overview of the hazards, models and performance metrics appearing in the articles, presenting results for sub-questions 1-3. Subsection 3.2 reports on the performance of the models, answering sub-question 4, while subsection 3.3 reflects on their feasibility of application, corresponding to sub-questions 5, 6, 7 and 9. Finally, subsection 3.4 considers the possible opportunities for cross-application of the models and methods reported in the studies, reporting the findings of sub-question 8.

3.1 Models and metrics

This subsection will deal with the results of sub-questions 1 through 3.

3.1.1 Models

This subsection reports on the first sub-question: “What are the machine learning models used in the papers?“. The models used in the papers are varied, with 22 of the 28 papers using more than one model. In total, there are 36 distinct models present, with 84 implementations across all papers, counting the repetitions. 22 of the models appear only once, with a further 6 appearing twice. Only eight models appear three or more times across the surveyed papers. The most popular individual model is random forest(RF) [10], appearing in 16 papers. The most common reasons cited for it's inclusion is it's ability to handle large datasets without overfitting and robustness to outliers and missing values, both of which prove relevant in the field of disaster forecasting. Additionally, many papers mention its successful usage in previous literature. It is followed by support vector machine(SVM) [11], included in ten, with the main reasons being it's ability to perform well in high-dimensionality datasets, as well as precedent in literature. Boosted regression trees(BRT) [12] and logistic regression(LR) [13] have six and five appearances respectively. Finally, extreme gradient boosting(XGBoost) [14] and generalized linear models(GLM) [15] appear four times each, while maximum entropy algorithm(MaxEnt) [16] and K-nearest-neighbours(KNN) [17] three times each. Table 2 gives an overview of the distribution of the most common models in each category, with rarely used models in the last row of every category, and summarized by “Others“ if necessary. While the table does present a complete overview of the papers surveyed in the study, the complete overview of the all of the models can be found in Appendix A.2.

Categorization		
Category	Model	Articles
Supervised	SVM	[18], [19], [20], [21], [22], [23], [24], [25], [26], [27]
	LR	[28], [18], [29], [27], [30]
	GLM	[31], [32], [26], [33]
	MaxEnt	[34], [24], [35]
	KNN	[18], [29], [30]
	CART, MARS, BJSR, NB	[18], [20], [3], [23], [36], [30]
Unsupervised	DA	[26], [33]
	K-means	[37], [30]
	PCA, SPLS	[31], [37]
Deep learning/NN	MLP	[18], [19]
	FNN, CNN, DeepNDF, Others	[38], [39], [40], [29]
Ensemble	RF	[28], [18], [19], [20], [41], [21], [22], [31]
	RF cont.	[3], [42], [24], [43], [25], [27], [33], [30]
	BRT	[20], [31], [3], [24], [32], [33]
	XGBoost	[28], [18], [25]
	BART, AdaBoost, CatBoost, Others	[18], [22], [40], [44], [43], [25]

Table 2: Overview of machine learning models in the surveyed papers

This sub-question is ultimately a supporting question to determining performance, and since so many of the models appear only several times and reporting on individual data points does not carry much relevance, the overview of the frequency of model usage was used to help analyze performance. Though RF and SVM are the most popular models individually, as a group boosting models appear often, with the most popular being BRT and XGBoost. Additionally, though the individual models rarely repeat, deep learning and neural network models appear ten times. The least popular are unsupervised learning models, appearing only seven times. Curiously, the number of models used in a study does not appear correlated with the number of hazards being evaluated. Only six studies use a single model, of which two use RF, two their own custom built models, and, as a point of interest, the remaining two use MaxEnt, which is only used three times overall.

3.1.2 Metrics

This subsection presents the results for the second sub-question: “What are the metrics used to report on the performance of the models?“. The metrics used across the surveyed papers are numerous, with 20 different metrics repeated a total of 84 times. Table 3 gives an overview of which metrics are applied by which articles. Area under the receiver operating curve(ROC-AUC) is by far the most common metric, appearing in 23 out of the 28 surveyed articles. The receiver operating curve(ROC) is a graph that can be used to show how well a classification model performs by plotting the true positive rate to the false positive rate at different thresholds [45]. The area under the curve(AUC) is a numerical value that represents well the meaning in the graph. It takes values between 0 and 1, with a value below 0.5 meaning the model is guessing, and values between 0.7 to 1 considered acceptable to excellent.

To help with readability, the metrics have been grouped by the model type they generally evaluate and ones used less than three times grouped together. Metrics used in classification, originating from the confusion matrix [46] appear most often - 10 distinct metrics appearing a total of 43 times. The most popular are recall at ten appearances and precision at eight [47], followed by F1 at seven and accuracy at six [48]. These metrics are often used together, particularly precision with recall, which occur together eight times. Specificity and negative predictive value(NPV) occurring four and three times respectively [48]. The other metrics used less than three times can be found in Table 3. Regression is measured by metrics such as root mean squared error(RMSE), mean absolute error(MAE), R squaredR^2 and total sum of squares(TSS) [49]. They are the second most common group among the surveyed literature, appearing 13 total times in 7 articles - RMSE four times and the others three times each. Lastly, several metrics used to evaluate clustering or other specifics appear several times and form the last two groups in Table 3. The complete list of metric details can be found in Appendix A.3.

Categorization		
Category	Metric	Articles
Classification	ROC-AUC	[28],[18], [38], [34], [20], [39], [41], [21], [22], [31], [3], [44]
	ROC-AUC cont.	[23], [42], [24], [43], [35], [32], [36], [29], [26], [27], [33]
	recall	[18],[19], [38], [20], [41], [40], [31], [42], [32], [30]
	precision	[19], [38], [20], [41], [40], [31], [32], [30]
	F1	[19], [38], [41], [40], [35], [25], [30]
	accuracy	[18],[19], [40], [29], [27] , [30]
	specificity	[20], [31], [42], [32]
	NPV	[20], [31], [32]
	kappa, FAR, HSS, CSI	[18], [39], [27]
Regression	RMSE	[22], [36], [29], [27]
	MAE	[22], [29], [27]
	R^2	[22], [36], [29]
	TSS	[23], [42], [25]
Clustering	R-index, CCI, Silhouette	[21], [42], [37]
Others	IoU, Gini co.	[40], [42], [35]

Table 3: Overview of performance metrics in the surveyed papers

As a point of interest, only one article uses neither ROC-AUC nor at least one of the confusion matrix metrics - it is Rocchi et al. [37], which makes use only of the silhouette method. It is important to note that the naming is not uniform across the surveyed literature, and the synonyms have been grouped together, with the most common term as the name. For example precision is also called positive predictive value(PPV), and recall can be alternatively named sensitivity, true positive rate or possibility of detection [47].

3.1.3 Connections

This subsection deals with the third sub-question: “How does the choice of metric depend on model and domain?“. There does not appear to be a very clear connection between the metrics and the models chosen in a study, mostly due to the prevalence of ROC-AUC as a metric, as well as the majority of studies using more than one model. There are five studies that did not use ROC-AUC, three of which use recall, precision, F1 and accuracy. Rocchi et al. is the only user of the silhouette method, which is well-matched with their combination of PCA and K-Means [37]. Lastly, Ye et al. chose to only use F1 and TSS [25]. Eight of the papers chose to exclusively use ROC-AUC as a metric. Of those, five use RF, three BRT, two GLM, two MaxEnt, and two FDA. The seven articles making use of the regression metrics all use RF, SVM or LR, with the exception of Tang et al. using Bayesian spacial joint regressions [36]. However, since RF, LR and SVM are the most commonly used models among the whole study, used in 19 articles out of 28, this is rather to be expected. The connection between the predicted hazards and the choice of metric also seems nonexistent. The variety of model and metric types, combined with the fact the majority of surveyed articles use several models and several metrics, makes it very difficult to identify a pattern, and indeed one has not been found by the author. It is also connected to the fact the most common occurring hazards, landslides and floods, both appear in 21 papers, with fire appearing in 9, erosion in 6 and earthquakes in 5. The average study, therefore, applies RF and SVM to predict landslides and floods, and measured their performance using ROC-AUC.

ROC-AUC seems close to a universal metric, appearing in 23 out of the 28 studies. While it’s popularity is undeniable, and most of the surveyed articles are confident in it’s choice as a metric, some argue that it is not sufficient as an individual metric. Pourghasemi et al. mention “The success rate is not the best tool to evaluate the prediction capacity of the models, however, as it uses the same training data that generated the predictions. It does help to assess the correlation of the multi-hazard maps to the inventory of hazards“ [23, p. 11]. Many articles also use additional metrics, often mentioning the need for additional detail or verification as the reason.

3.2 Performance

This subsection deals with sub-question number 4: “What is the performance per metric, domain and model?“. It analyzes the most often occurring models and their performance in predicting various hazards, as given by the AUC metric. Since the ROC-AUC metric is by far the most common among the surveyed literature, it is the scores in this metric that are compared and counted, with results from articles that did not use ROC-AUC not included in the results of this section. An AUC score of 0.5 to 0.6 is poor, 0.6 to 0.7 average, 0.7 to 0.8 good, 0.8 to 0.9 very good and 0.9 to 1 excellent [45]. The models occurring most commonly are RF, SVM, BRT, LR and XGBoost.

The RF model [10] is used thirteen times for landslide prediction, ten times for floods, four times for fires and twice for erosion. It performs very well overall, with an average AUC score of 0.93 in landslides, 0.91 floods, 0.89 in fire prediction and 0.9 in erosion. It is also used once each for droughts and earthquakes, scoring 0.79 and 0.78. SVM [11] is used eight times each for landslide and flood prediction, with an average AUC score of 0.84 and 0.89 respectively. It is also used four times for fires(AUC of 0.76), twice for erosion(AUC of 0.94) and once for snow avalanche prediction(0.89). While both models generally gain very good or excellent scores, there are six studies where both models were applied to landslides and floods, and RF does outperform SVM, by 15% in landslide prediction, and 4% in flood prediction. The BRT model [12] is utilized seven times for landslides, with an average score of 0.89, as well as twice each for flood(0.89), fire(0.78) and erosion(0.88), and once for droughts(0.79). While generally it performed slightly worse than both RF and SVM, it does outperform SVM in landslide prediction. The LR model [13] appears five times for landslides, with an average AUC score of 0.86, four times for floods with a score of 0.88, as well as once each for fire(0.94) and earthquake(0.69) prediction. Lastly, the XGBoost model is used twice for landslides and floods, scoring excellent for landslides with an average AUC score of 0.95, and less well for floods, with AUC of 0.83. It is also used once for fire(0.99) and once for earthquake prediction(0.78). RF does outperform all of the other models listen in landslide prediction, with the exception of XGBoost which gains a slightly better AUC score of 0.95, but it's two data points compared to RF's thirteen lend it less credibility.

Unfortunately, the possibilities for effective analysis in this regard are extremely limited. The wide variety of hazards, models and performance metrics used in the surveyed articles leads to a very sparse matrix, without many chances for comparison. The vast majority of present combinations occurs only once or twice, and detailing singular data points is neither informative not insightful. Given the very noticeable differences in performance of predicting different hazards for the same model, it would not be informative to present the average model performance in all of the hazards it was applied to. The models, too, are significantly different, and treating, for example, all of the deep learning models as one group would be misleading and not lead to meaningful results. The same issue applies to the various performance metrics - the F1 score gained by one model cannot effectively be compared to the ROC-AUC score gained by a different one in predicting the same hazard. Due to these severe limitations, further analysis of model performance is outside of the scope of the survey.

3.3 Feasibility and intended use of the models

This subsection will deal with the answers to sub-questions 5-9 with the exception of sub-question 8.

5. How to define and judge feasibility of practical application?
6. What are the intended practical applications of the models?
7. What factors influence the feasibility of the intended use of the models?
9. What is the feasibility of practical application of the models?

3.3.1 Intended applications of the models

This subsection answers sub-question 6. “What are the intended practical applications of the models?“. The given result of all but four of the surveyed articles is a multi-hazard map. Interestingly, the naming of the maps is inconsistent and seemingly interchangeable. The maps created are called (multi-hazard) susceptibility maps, risk index maps, exposure maps, probability maps or risk maps. While “multi-hazard susceptibility map“ seems to be the most common term, many of these are used interchangeably or no difference can be found between the style of map created. While most of the maps include the specific hazard combinations as part of the multi-hazard map, some only mark the level of risk, without mentioning which hazards the given spot may be vulnerable to. This is not in any way correlated with the specific naming of the type of map, and seems just as likely with any of the particular types.

In line with the relative similarity of the generated results, the intended applications of the models and their results are echoed in the various surveyed papers. The created maps are meant to give insights for disaster risk reduction and resilience planning, aid in prioritizing high-risk zones for hazard management and mitigation strategies, serve as a tool in sustainable and safe land-use planning and infrastructure development, as well as be helpful for the conservation of natural resources and the environment. The intended users are named to be planners, policy makers, decision makers, stakeholders, local governments as well as emergency managers.

Two of the four articles that do not create a multi-hazard map echo those same intentions. Padmaja et al. created a model that provides disaster segmentation and say it offers “actionable insights for disaster preparedness and mitigation“ [40, p. 159] while Rocchi et al. assign a general risk score to each municipality and say their tool can “can assist governments and stakeholders in decision making and prioritization of interventions“ [37, p. 1]. The last two articles focus more on applications in early warning systems. Dal Barco et al. whose model assigns a daily risk index per municipality in the studied region, and whose model's intended applications are to aid in the mitigation planning process, report the model “can also be used as an early warning system when applied to short term weather forecasts“ [19, p. 17]. Leinonen et al. developed a nowcasting model for thunderstorms and intended for it to serve as an early warning system that can issue storm warnings a short time in advance [39].

This sub-question grants some insight into the intentions behind the surveyed literature. Both the final products and their intended applications are very closely aligned across the articles. It is interesting to note that humanitarian organizations are not directly mentioned among the intended users, focusing more on local authorities. This is perhaps connected to how few of the models attempt to predict the time a disaster may occur. Most anticipatory actions carried out by humanitarian organizations require early warning of a specific event, several hours to several months in advance, depending on the disaster. This does not seem to be a focus or intention currently common in the field, though some articles do mention it as a possible extension point. Another point of interest is that none of the articles report on cooperation with the local authorities or the other detailed stakeholders, nor do they mention contacting anybody to share the created maps or models with them. This could be simply because they did not deem it relevant to the academic research, but it may also suggest a widespread approach to the potential applications that is purely academic, with little care for the results being utilized.

3.3.2 Definition and factors of feasibility

This subsection answers sub-question 5: “How to define and judge feasibility of practical application?“, as well as sub-question 7: “What factors influence the feasibility of the intended use of the models?“. As the questions are strongly related to each other, the results for them are presented together. In the context of what this work is examining, the feasibility of practical application is meant as the usefulness and readiness of the tools developed in the surveyed literature in preventing or foreseeing natural disasters that have the potential to cause harm to humans. Simply put, it is the ease with which a stakeholder could use the developed tool, and how effective that tool would be. To judge this concept, the surveyed articles and their final result have been evaluated on four factors of their development, each on a scale from 1 to 5, with 1 being unsuitable for usage in disaster prediction and humanitarian forecasting, and 5 being excellent, summing up to maximum 20. The result is then scaled to a percentage value for easy understanding.

The development factor represents the simple fact that for a model to be useful, it needs to be fully developed and functional. The score reflects how finalized the model development is. A model that is still a work-in-progress would receive a low score, while one that is fully fleshed out and complete would receive a high score. It is worth noting that this is partially based on what the articles report as potential future work and the importance of that work, so a model may receive a high development score if the authors did not elaborate on potential future development. The reliability factor corresponds to the integrity and reproducibility of the models. If the development process, like reliable data sources, feature selection or choice of model hyper-parameters, is reported on and thoroughly done, an article would receive a high score. If the model creation is not reproducible or the results could not be trusted without additional verification, a low score would be assigned.

The performance of a model is also crucial, as a model whose results cannot be trusted is not suitable to be used, particularly when an incorrect prediction about a disaster may well cost human lives. A model would receive a score of five if it performs excellently, a three if the performance is satisfactory, and lower if it performs badly. As a note, if an article does not report the performance clearly, a score of three is given as an average score. Lastly, the amount of detail given by a model is represented by its own factor. To be effective, a model needs to give sufficient information on disaster types, locations and potentially time frames. The way the information is presented should also be readable. For example, a map giving a general multi-hazard risk index to each district in a province would receive a lower score than an article that created detailed landscape maps that report on the specific hazard combinations present.

Note that these factors were created by the author as a way to represent what is meant by feasibility in the context of this research. It is not a strict definition, nor a commonly used one, and is meant more as an understandable structure to present the findings than as completely reliable and objective model.

3.3.3 Feasibility of the models

This subsection answers sub-question 9: “What is the feasibility of practical application of the models?“. As detailed in subsection 3.3.2, the feasibility of application is judged through four factors: development, reliability, performance and detail. Table 4 gives the feasibility scores for each article. To keep the table small, the factor names were shortened to “Dev“, “Rel“, “Perf“ and “Det“, respectively. The overall average sum was 15.75, corresponding to a 78.75% feasibility score. Seven of the articles scored 90% or above, demonstrating good feasibility, while nine scored below 75% and twelve received a total score in between, which the authors considers below average and average, respectively.

The performance scores were based only on the performance of the model used to create

Reference	Dev.	Rel.	Perf.	Det.	Total	Score
[28]	5	5	4	4	18	90%
[18]	5	5	4	3	17	85%
[19]	2	3	3	2	10	50%
[38]	4	4	4	3	15	75%
[34]	4	3	5	3	15	75%
[20]	4	2	5	3	14	70%
[41]	3	3	4	3	13	65%
[39]	4	5	5	3	17	85%
[21]	3	3	4	4	14	70%
[22]	5	4	4	4	17	85%
[40]	4	3	5	2	14	70%
[31]	5	4	4	4	17	85%
[3]	3	3	4	5	15	75%
[44]	4	3	3	4	14	70%
[23]	4	5	4	4	17	85%
[42]	3	2	4	4	13	65%
[24]	4	2	4	4	14	70%
[43]	4	5	5	4	18	90%
[37]	3	4	3	2	12	60%
[35]	3	4	4	5	16	80%
[32]	4	5	5	4	18	90%
[36]	5	5	4	4	18	90%
[29]	4	4	5	5	18	90%
[25]	5	5	4	5	19	95%
[26]	4	2	5	4	15	75%
[27]	4	5	5	5	19	95%
[33]	5	4	4	4	17	85%
[30]	4	4	4	5	17	85%

Table 4: Feasibility scores of the articles

the final results. Most likely as a consequence, the best scores overall were received in the performance factor, with an average score of 4.2 compared to the average scores of development, reliability and detail, which are 3.96, 3.79 and 3.79 respectively. This is possibly caused by the majority of articles testing several models and choosing the ones with the best performance to generate the results. A strong limitation of this method is that it can only rely on what the authors of the surveyed articles have chosen to report on. The development factor is particularly vulnerable to this, since it is influenced almost exclusively by the self-assessed completeness. Therefore, a model that is very well-developed but reports several additional potential improvements may receive the same or even lower score than one that reports no improvements at all, even if it is less extensive. Additionally due to various metrics there may be slight differences in performance scores given to models that use ROC-AUC as a metric compared to ones only using alternative metrics. It is also important to note the scores were assigned by the author by hand, and cannot be guaranteed to be completely objective.

As an example of the reasoning behind the score assignments, Ye et al. received 5 for development, since their only note is a need for a little extra detail. They report very thoroughly on their data sources, the variables and model creation, so they also receive a 5 for reliability. The best model scores are between 0.8 and 0.9 in F1-score, which led them to receive a 4 in performance. Lastly, they create many maps, including individual ones per hazard and per model, as well as a very detailed multi-hazard map, also gaining a 5 in detail.

3.4 Potential cross-application

This subsection answers sub-question number 8: “What are possible cross-applications of the models in humanitarian forecasting?“. The capacity for cross-application is meant here as the potential to apply the model or methodology outside of the case study area, or extend it to new hazards. Fifteen of the articles do not mention cross-application in any way [20, 41, 22, 31, 3, 44, 23, 42, 24, 37, 35, 32, 36, 25, 33]. A further seven report that their methodology could be replicated in other areas [28, 34, 21, 43, 26, 27, 30]. Choubin et al. use the word “framework“ without additional details given [18], while Padmaja et al. writes “The proposed model provides a robust framework for integrating real-time remote sensing data into early warning systems“, but also does not provide any additional details [40, p. 159]. These models have been counted in the methodology replication group, bringing the number up to nine articles.

Only four of the articles truly call their models cross-applicable. Dal Barco et al. report that “The same algorithm can be applied to other geographic areas, with the same or different indicators and assessment endpoints, provided that the algorithm is retrained or fine-tuned on local data.“ [19, p. 16]. Hasnanoui et al. describe their model as a scalable, adaptive and reproducible framework. They mention it can be used in other regions, though warn that it's application has not been tested for any region other than the case study [38]. Leinonen et al. note that their model architecture “can use a combination of many data sources to predict various hazards from thunderstorms“ and that they chose “different approaches for different hazards in order to demonstrate the flexibility of the model, which can be adapted to the users' needs with minor changes to the training procedure“ [39, p. 8]. Lastly, Ullah et al. say that “the proposed method is also applicable to other environmental hazards“, as well as claim that “‘it can be applied to similar geo-environments, especially in mountainous areas with sparse data“ [29, p. 14].

It is important to note that the results presented in this section are based only on the information given directly in the articles - no additional interpretation is done by the author to judge the potential for cross-application. The results indicate that cross-application is not a common consideration in the field of machine learning multi-hazard disaster prediction. This is likely due to the unique relationship between the data available for a given region and the decisions made in model development - there is no model without data, and as the potential data sources strongly vary between regions, a lack of universal data sources stands as the biggest challenge to cross-application.

4 Responsible Research

This section reflects on the ethics of this study, including the methodology and the recommendations made by the author. Humanitarian forecasting and more generally disaster prediction is by itself a highly ethical pursuit, as it can save human lives and prevent much damage. Nevertheless, there are some ethical risks present in the execution of this study.

The validity and trustworthiness of the surveyed articles is an implicit assumption made by the author, since the survey only considers literature published in academic journals. However, the author has strong reasons to believe one of the articles found during the initial pass was created mostly through the use of an LLM, as it lacked cohesiveness, repeated fragments of text and cited non-existent research. This calls into doubt the assumed reliability, which poses something of an ethical concern, but since the other articles showed none of the same flaws and verifying their validity in detail is outside of the scope of this survey, the author chose to upkeep the general assumption of trustworthiness.

The research does not deal with sensitive data relating to human subjects, so that is not a concern. The validity of the data sources reported by the surveyed literature, however, impacts the validity of the literature itself. In line with the above assumption and due to time constraints, the data sources have not been validated and are assumed to be trustworthy. Additionally, as detailed in section 2, the survey is systematized and not systemic. As a result, articles relevant to the research may have been omitted due to language differences or being published before 2019. This reduces the completeness of the survey and implies possible bias in the results towards English language research, but was found to be unavoidable given the time constraints and scope of the research.

Lastly, the conclusions and recommendations made in this work are logically sound and reasoned to the best of the author's ability. As reliable disaster predictions can make the difference between life and death, an unsound model recommendation could have severe consequences. In addition, no individual is free of bias, and it is a relevant ethical risk that the conclusions may have been unknowingly influenced by the author's personal beliefs. Given the small scope of the research, it is highly unlikely that the recommendations made in this work would be followed blindly by relevant parties without additional verification.

5 Conclusions and Future Work

This section contains the conclusions of this work, as well as the author's recommendations for future research. The contribution of this work is an up-to-date systematized literature survey of machine learning multi-hazard disaster prediction, including the types of models and performance metrics used, the performance of the models, their feasibility of use as well as the potential for cross-applications. The survey contains 28 articles published between 2019 and 2025.

The most widely predicted hazards are landslides and floods, and there is a wide variety of both models and metrics used. The most common models are RF and SVM, followed by BRT and LR. Boosting models are also becoming more popular and show a lot of potential. The most used metric is ROC-AUC, with precision and recall also appearing often. No apparent connection has been found between the model, hazard and metric choice, which is likely due to the amount of diversity and potential combinations in all three. The overall performance of machine learning models is very good and reaffirms the suitability of machine learning methods for disaster prediction and humanitarian forecasting. The best performing model by AUC-ROC score overall is RF, which performs excellently for landslides and floods. It narrowly outperforms SVM and XGBoost. Due to the sparseness of the model metric hazard combinations, it is difficult to identify patterns in performance, and further work is needed on this subject.

The most common result of the studies is a multi-hazard susceptibility map. Their intended applications are broadly to give insights for disaster risk reduction, aid in prioritizing and planning of mitigation strategies and help in sustainable and safe land-use planning. Feasibility of practical application was defined as the usefulness and readiness of the created models to be applied in disaster prediction in practice. It was decided it would be judged with four factors: development, representing the development stage of the model; reliability, responsible for integrity and reproducibility; performance, correspondingly simply to the reported performance metric scores; and detail, representing the amount of information given by the model. The four factors sum up to a general percentage score. Seven of the models have excellent feasibility, with scores above 90%, nine score below average with scores below 75% and twelve have acceptable feasibility. Cross-application is not considered by most of the surveyed articles, but nine report their methodology can be replicated in other areas, and four of the models can be retrained on data from another area.

Recommended further work for this survey is to broaden the scope to include work published before 2019, as well as extend it into a fully systemic survey. Additionally, further work is needed to better analyze the model performance, the data sources of the surveyed articles should be verified, and a more in-depth analysis of factors that make a model suitable for cross-application carried out. Recommendations for further development in the field would be to focus on developing models that predict specific hazard occurrences in time, prioritize creation of cross-applicable models, as well as generally more international cooperation.

A Appendix

The appendix contains supporting materials that were not necessary to include in the main body of the work, but are still relevant to include for more detailed readers and for the sake of further clarity or reproducibility.

A.1 Surveyed papers

This subsection contains details of all of the surveyed articles, presented in Table 5 and Table 6.

A.2 Model details

This appendix contains the details on the models present during the work. The full model names, the shortcuts used for them, as well as the citations can be found in Table 7. Note the table does not contain custom ensemble models unique to one study.

A.3 Metric details

This subsection contains the naming details and citations of the performance metrics present in the work, presented in table 8. Note that the "Full name(s) of metric" column contains all of the synonyms found in the surveyed articles.

References

[1] P. Lemke, J. Ren, R. Alley, I. Allison, J. Carrasco, G. Flato, Y. Fujii, G. Kaser, P. Mote, R. Thomas, and T. Zhang, "IPCC, 2007. Climate change 2007. Synthesis report. Contribution of working groups I, II & III to the fourth assessment report of the intergovernmental panel on climate change. Geneva," 2007.

[2] M. Reichstein, V. Benson, J. Blunk, G. Camps-Valls, F. Creutzig, C. Fearnley, B. Han, K. Kornhuber, N. Rahaman, B. Schölkopf, J. Tárraga, R. Vinuesa, K. Dall, J. Denzler, D. Frank, G. Martini, N. Nganga, D. Maddix, and K. Weldemariam, "Early Warning of Complex Climate Risk with Integrated Artificial Intelligence," *Nature Communications*, vol. 16, no. 1, 2025.

[3] Y. Piao, D. Lee, S. Park, H. Kim, and Y. Jin, "Multi-Hazard Mapping of Droughts and Forest Fires Using a Multi-Layer Hazards Approach with Machine Learning Algorithms," *Geomatics, Natural Hazards and Risk*, vol. 13, no. 1, pp. 2649–2673, 2022.

[4] J. C. Gill and B. D. Malamud, "Reviewing and visualizing the interactions of natural hazards," *Rev. Geophys.*, vol. 52, pp. 680–722, 2014.

[5] U. N. E. Programme, "Agenda 21," 1992.

Title	**Author**	**Year**	**Journal**	**Cite**
Multi-hazard probability assessment and mapping in Iran	H.R. Pourghasemi et al.	2019	Science of The Total Enviroment	[44]
Multi-hazard exposure mapping using machine learning for the state of Salzburg, Austria	T.G. Nachappa et al.	2020	Remote Sensing	[21]
Is multi-hazard mapping effective in assessing natural hazards and integrated watershed management?	H.R. Pourghasemi et al.	2020	Geoscience Frontiers	[23]
Assessing and mapping multi-hazard risk susceptibility using a machine learning technique	H.R. Pourghasemi et al.	2020	Scientific Reports	[42]
A machine learning framework for multi-hazards modeling and mapping in a mountainous area	S. Yousefi et al.	2020	Scientific Reports	[26]
Evaluation of multi-hazard map produced using MaxEnt machine learning technique	N. Javidan et al.	2021	Scientific Reports	[34]
Modelling multi-hazard threats to cultural heritage sites and environmental sustainability: The present and future scenarios	A. Saha et al.	2021	Journal of Cleaner Production	[32]
Evaluation of debris flow and landslide hazards using ensemble framework of Bayesian- and tree-based models	S.C. Pal et al.	2022	Bulletin of Engineering Geology and the Environment	[31]
Multi-hazard mapping of droughts and forest fires using a multi-layer hazards approach with machine learning algorithms	Y. Piao et al.	2022	Geomatics, Natural Hazards and Risk	[3]
Multi-hazard susceptibility and exposure assessment of the Hindu Kush Himalaya	J. Rusk et al.	2022	Science of the Total Environment	[35]
Multi-hazard susceptibility mapping based on Convolutional Neural Networks	K. Ullah et al.	2022	Geoscience Frontiers	[29]
Landslides and flood multi-hazard assessment using machine learning techniques	A.M. Youssef et al.	2022	Bulletin of Engineering Geology and the Environment	[27]
A Machine Learning Framework for Multi-Hazard Risk Assessment at the Regional Scale in Earthquake and Flood-Prone Areas	A. Rocchi et al.	2022	Applied Sciences	[37]
A Hybrid Multi-Hazard Susceptibility Assessment Model for a Basin in Elazig Province, TÃŒrkiye	G. Karakas et al.	2023	International Journal of Disaster Risk Science	[41]
Thunderstorm Nowcasting With Deep Learning: A Multi-Hazard Data Fusion Model	J. Leinonen et al.	2023	Geophysical Research Letters	[39]

Table 5: Surveyed papers: part 1

Title	**Author**	**Year**	**Journal**	**Cite**
Multi-resource potentiality and multi-hazard susceptibility assessments of the central west coast of India applying machine learning and geospatial techniques	P. Prasad et al.	2023	Environmental Earth Sciences	[43]
Multi-hazards (landslides, floods, and gully erosion) modeling and mapping using machine learning algorithms	A.M. Youssef et al.	2023	Journal of African Earth Sciences	[33]
Flood, landslides, forest fire, and earthquake susceptibility maps using machine learning techniques and their combination	H. R. Pourghasemi et al.	2023	Natural Hazards	[24]
Multi-hazard assessment using machine learning and remote sensing in the North Central region of Vietnam	H.D. Nguyen et al.	2023	Transactions in GIS	[22]
Machine learning-enabled regional multi-hazards risk assessment considering social vulnerability	T. Zhang et al.	2023	Scientific Reports	[30]
A machine learning approach to evaluate coastal risks related to extreme weather events in the Veneto region (Italy)	M.K. Dal Barco et al.	2024	International Journal of Disaster Risk Reduction	[19]
Development of risk maps for flood, landslide, and soil erosion using machine learning model	N. Javidan et al.	2024	Natural Hazards	[20]
Risk assessment of landslide and rockfall hazards in hilly region of southwestern China: a case study of Qijiang, Wuxi and Chishui	P. Ye et al.	2024	Environmental Earth Sciences	[25]
Machine learning and GIS-based multi-hazard risk modeling for Uttarakhand: Integrating seismic, landslide, and flood susceptibility with socioeconomic vulnerability	V. Chauhan et al.	2025	Environmental and Sustainability Indicators	[28]
A spatially explicit multi-hazard framework for assessing flood, landslide, wildfire, and drought susceptibilities	B. Choubin et al.	2025	Advances in Space Research	[18]
Transfer learning-based deep learning models for flood and erosion detection in coastal area of Algeria	Y. Hasnaoui et al.	2025	Earth Science Informatics	[38]
Deep Learning in Remote Sensing for Climate-Induced Disaster Resilience: A Comprehensive Interdisciplinary Approach	S.M. Padmaja et al.	2025	Remote Sensing in Earth Systems Sciences	[40]
Spatial joint hazard assessment of landslide susceptibility and intensity within a single framework: Environmental insights from the Wenchuan earthquake	Z. Tang et al.	2025	Science of the Total Environment	[36]

Table 6: Surveyed papers: part 2

Short name	Full Model Name	Citation
LR	logistic regression	[13]
RF	random forest	[10]
XGBoost	extreme gradient boosting	[14]
SVM	support vector machine	[11]
CART	classification and regression tree	[50]
MLP	multi-layer perceptron	[51]
NB	NaÃ¯ve Bayes	[52]
KNN	k-nearest neighbours	[17]
DeepNDF	deep neural decision forests	[53]
FNN	feedforward neural network	[54]
autoencoders	autoencoders	[55]
Bi-RNN	bidirectional recurrent neural networks	[56]
MaxEnt	maximum entropy	[16]
BRT	boosted regression trees	[12]
MARS	multivariate adaptive regression spline	[57]
AdaBoost	adaptive boosting	[58]
BGLM	Bayesian generalized linear models	[59]
SPLS	sparse partial least squares	[60]
PCA	principal component analysis	[61]
KM	k-means clustering	[62]
BART	Bayesian additive regression trees	[63]
BSJR	Bayesian spacial joint regression	[64]
CNN	convolutional neural network	[65]
LightGBM	light gradient-boosting machine	[66]
CatBoost	CatBoost	[67]
GLM	generalized linear model	[15]
FDA	functional discriminant analysis	[68]
MDA	multivariate discriminant analysis	[69]

Table 7: Models present in the study

Shortcut	Full name(s) of metric	Citation
ROC-AUC	receiver operating characteristics and area under the curve	[45]
accuracy	accuracy	[48]
precision	precision, positive predictive value(PPV)	[70]
recall	recall, sensitivity, true positive rate(TPR), probability of detection(POD)	[48]
F1	balanced f-score	[48]
FAR	false alarm ratio, false discovery rate	[71]
HSS	Heidke Skill Score	[72]
kappa	Cohen's kappa	[48]
specificity	specificity, true negative rate(TNR)	[70]
NPV	Negative predictive value	[70]
CSI	critical success index	[73]
R-Index	Rand index	[74]
RMSE	Root Mean Squared Error	[48]
MAE	mean absolute error	[48]
R^2	R squared	[48]
IoU	Intersection over union	[75]
TSS	total sum of squares	[49]
CCI	cosine clustering index	[76]
Gini co.	Gini coefficient	[77]
Silhouette	silhouette method	[78]

Table 8: Complete list of performance metrics used in the surveyed literature

[6] T. A. Owolabi and M. Sajjad, "A Global Outlook on Multi-Hazard Risk Analysis: A Systematic and Scientometric Review," *International Journal of Disaster Risk Reduction*, vol. 92, p. 103727, June 2023.

[7] A. Tilloy, B. D. Malamud, H. Winter, and A. Joly-Laugel, "A review of quantification methodologies for multi-hazard interrelationships," *Earth-Science Reviews*, vol. 196, p. 102881, 2019.

[8] M. J. Grant and A. Booth, "A typology of reviews: an analysis of 14 review types and associated methodologies," *Health Information & Libraries Journal*, vol. 26, no. 2, pp. 91–108, 2009.

[9] L. A. Goodman, "Snowball Sampling," *The Annals of Mathematical Statistics*, vol. 32, no. 1, pp. 148–170, 1961.

[10] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282, IEEE, 1995.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[13] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, 2016.

[15] P. McCullagh, *Generalized Linear Models.* New York: Routledge, 2 ed., Jan. 1989.

[16] S. J. Phillips, R. P. Anderson, and R. E. Schapire, "Maximum Entropy Modeling of Species Geographic Distributions," *Ecological Modelling*, vol. 190, no. 3, pp. 231–259, 2006.

[17] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[18] B. Choubin, A. Jaafari, and D. Mafi-Gholami, "A Spatially Explicit Multi-Hazard Framework for Assessing Flood, Landslide, Wildfire, and Drought Susceptibilities," *Advances in Space Research*, vol. 75, no. 3, pp. 2569–2583, 2025.

[19] M. Dal Barco, M. Maraschini, D. Ferrario, N. Nguyen, S. Torresan, S. Vascon, and A. Critto, "A Machine Learning Approach to Evaluate Coastal Risks Related to Extreme Weather Events in the Veneto Region (Italy)," *International Journal of Disaster Risk Reduction*, vol. 108, 2024.

[20] N. Javidan, A. Kavian, C. Conoscenti, Z. Jafarian, M. Kalehhouei, and R. Javidan, "Development of Risk Maps for Flood, Landslide, and Soil Erosion Using Machine Learning Model," *Natural Hazards*, vol. 120, no. 13, pp. 11987–12010, 2024.

[21] T. Nachappa, O. Ghorbanzadeh, K. Gholamnia, and T. Blaschke, "Multi-Hazard Exposure Mapping Using Machine Learning for the State of Salzburg, Austria," *Remote Sensing*, vol. 12, no. 17, 2020.

[22] H. D. Nguyen, D.-K. Dang, Q.-T. Bui, and A.-I. Petrisor, "Multi-Hazard Assessment Using Machine Learning and Remote Sensing in the North Central Region of Vietnam," *Transactions in GIS*, vol. 27, pp. 1614–1640, Aug. 2023.

[23] H. Pourghasemi, A. Gayen, M. Edalat, M. Zarafshar, and J. Tiefenbacher, "Is Multi-Hazard Mapping Effective in Assessing Natural Hazards and Integrated Watershed Management?," *Geoscience Frontiers*, vol. 11, no. 4, pp. 1203–1217, 2020.

[24] H. R. Pourghasemi, S. Pouyan, M. Bordbar, F. Golkar, and J. J. Clague, "Flood, Landslides, Forest Fire, and Earthquake Susceptibility Maps Using Machine Learning Techniques and Their Combination," *Natural Hazards*, vol. 116, pp. 3797–3816, Apr. 2023.

[25] P. Ye, B. Yu, W. Chen, Y. Feng, H. Zhou, X. Luo, and Y. Li, "Risk Assessment of Landslide and Rockfall Hazards in Hilly Region of Southwestern China: A Case Study of Qijiang, Wuxi and Chishui," *Environmental Earth Sciences*, vol. 83, no. 13, 2024.

[26] S. Yousefi, H. R. Pourghasemi, S. N. Emami, S. Pouyan, S. Eskandari, and J. P. Tiefenbacher, "A Machine Learning Framework for Multi-Hazards Modeling and Mapping in a Mountainous Area," *Scientific Reports*, vol. 10, p. 12144, July 2020.

[27] A. Youssef, A. Mahdi, and H. Pourghasemi, "Landslides and Flood Multi-Hazard Assessment Using Machine Learning Techniques," *Bulletin of Engineering Geology and the Environment*, vol. 81, no. 9, 2022.

[28] V. Chauhan, L. Gupta, and J. Dixit, "Machine Learning and GIS-based Multi-Hazard Risk Modeling for Uttarakhand: Integrating Seismic, Landslide, and Flood Susceptibility with Socioeconomic Vulnerability," *Environmental and Sustainability Indicators*, vol. 26, 2025.

[29] K. Ullah, Y. Wang, Z. Fang, L. Wang, and M. Rahman, "Multi-Hazard Susceptibility Mapping Based on Convolutional Neural Networks," *Geoscience Frontiers*, vol. 13, no. 5, 2022.

[30] T. Zhang, D. Wang, and Y. Lu, "Machine Learning-Enabled Regional Multi-Hazards Risk Assessment Considering Social Vulnerability," *Scientific Reports*, vol. 13, p. 13405, Aug. 2023.

[31] S. Pal, R. Chakrabortty, A. Saha, S. Bozchaloei, Q. Pham, N. Linh, D. Anh, S. Janizadeh, and K. Ahmadi, "Evaluation of Debris Flow and Landslide Hazards Using Ensemble Framework of Bayesian- and Tree-Based Models," *Bulletin of Engineering Geology and the Environment*, vol. 81, no. 1, 2022.

[32] A. Saha, S. Pal, M. Santosh, S. Janizadeh, I. Chowdhuri, A. Norouzi, P. Roy, and R. Chakrabortty, "Modelling Multi-Hazard Threats to Cultural Heritage Sites and Environmental Sustainability: The Present and Future Scenarios," *Journal of Cleaner Production*, vol. 320, 2021.

[33] A. Youssef, A. Mahdi, M. Al-Katheri, S. Pouyan, and H. Pourghasemi, "Multi-Hazards (Landslides, Floods, and Gully Erosion) Modeling and Mapping Using Machine Learning Algorithms," *Journal of African Earth Sciences*, vol. 197, 2023.

[34] N. Javidan, A. Kavian, H. Pourghasemi, C. Conoscenti, Z. Jafarian, and J. Rodrigo-Comino, "Evaluation of Multi-Hazard Map Produced Using MaxEnt Machine Learning Technique," *Scientific Reports*, vol. 11, no. 1, 2021.

[35] J. Rusk, A. Maharjan, P. Tiwari, T.-H. Chen, S. Shneiderman, M. Turin, and K. Seto, "Multi-Hazard Susceptibility and Exposure Assessment of the Hindu Kush Himalaya," *Science of the Total Environment*, vol. 804, 2022.

[36] Z. Tang, X. Zheng, J. Pan, X. Huang, L. Zhu, N. Wang, M. Xie, G. Yan, C. Wang, Z. Wang, C. Xu, and C. Song, "Spatial Joint Hazard Assessment of Landslide Susceptibility and Intensity within a Single Framework: Environmental Insights from the Wenchuan Earthquake," *Science of the Total Environment*, vol. 963, 2025.

[37] A. Rocchi, A. Chiozzi, M. Nale, Z. Nikolic, F. Riguzzi, L. Mantovan, A. Gilli, and E. Benvenuti, "A Machine Learning Framework for Multi-Hazard Risk Assessment at the Regional Scale in Earthquake and Flood-Prone Areas," *Applied Sciences*, vol. 12, p. 583, Jan. 2022.

[38] Y. Hasnaoui, S. Tachi, H. Bouguerra, and Z. Yaseen, "Transfer Learning-Based Deep Learning Models for Flood and Erosion Detection in Coastal Area of Algeria," *Earth Science Informatics*, vol. 18, no. 2, 2025.

[39] J. Leinonen, U. Hamann, I. V. Sideris, and U. Germann, "Thunderstorm Nowcasting With Deep Learning: A Multi-Hazard Data Fusion Model," *Geophysical Research Letters*, vol. 50, no. 8, p. e2022GL101626, 2023.

[40] S. Padmaja, R. Naveenkumar, N. Kumari, E. Pimo, M. Bindhu, B. Konduri, and P. Jangir, "Deep Learning in Remote Sensing for Climate-Induced Disaster Resilience: A Comprehensive Interdisciplinary Approach," *Remote Sensing in Earth Systems Sciences*, vol. 8, no. 1, pp. 145–160, 2025.

[41] G. Karakas, S. Kocaman, and C. Gokceoglu, "A Hybrid Multi-Hazard Susceptibility Assessment Model for a Basin in Elazig Province, Türkiye," *International Journal of Disaster Risk Science*, vol. 14, no. 2, pp. 326–341, 2023.

[42] H. R. Pourghasemi, N. Kariminejad, M. Amiri, M. Edalat, M. Zarafshar, T. Blaschke, and A. Cerda, "Assessing and Mapping Multi-Hazard Risk Susceptibility Using a Machine Learning Technique," *Scientific Reports*, vol. 10, p. 3203, Feb. 2020.

[43] P. Prasad, V. Loveson, S. Mandal, P. Chandra, and L. Kulimushi, "Multi-Resource Potentiality and Multi-Hazard Susceptibility Assessments of the Central West Coast of India Applying Machine Learning and Geospatial Techniques," *Environmental Earth Sciences*, vol. 82, no. 9, 2023.

[44] H. R. Pourghasemi, A. Gayen, M. Panahi, F. Rezaie, and T. Blaschke, "Multi-Hazard Probability Assessment and Mapping in Iran," *Science of The Total Environment*, vol. 692, pp. 556–571, Nov. 2019.

[45] F. Melo, "Area under the ROC Curve," in *Encyclopedia of Systems Biology*, pp. 38–39, Springer, New York, NY, 2013.

[46] K. M. Ting, "Confusion Matrix," in *Encyclopedia of Machine Learning* (C. Sammut and G. I. Webb, eds.), pp. 209–209, Boston, MA: Springer US, 2010.

[47] K. M. Ting, "Precision and Recall," in *Encyclopedia of Machine Learning*, pp. 781–781, Springer, Boston, MA, 2011.

[48] I. H. Witten, E. Frank, and M. A. Hall, "Chapter 5 - Credibility: Evaluating What's Been Learned," in *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)* (I. H. Witten, E. Frank, and M. A. Hall, eds.), The Morgan Kaufmann Series in Data Management Systems, pp. 147–187, Boston: Morgan Kaufmann, third edition ed., 2011.

[49] P. Bhattacharya and P. Burman, "Linear Models," in *Theory and Methods of Statistics* (P. Bhattacharya and P. Burman, eds.), ch. 11, pp. 309–382, Academic Press, 2016.

[50] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* New York: Chapman and Hall/CRC, 1984.

[51] S. Haykin, *Neural networks: a comprehensive foundation.* Prentice Hall PTR, 1994.

[52] G. I. Webb, "Naïve Bayes," in *Encyclopedia of Machine Learning* (C. Sammut and G. I. Webb, eds.), pp. 713–714, Boston, MA: Springer US, 2010.

[53] P. Kontschieder, M. Fiterau, A. Criminisi, and S. R. BulÃ^2, "$DeepNeuralDecisionForests,'' in2015IEEEInternationalConferenceonComputerVision(ICCV), pp$
$-1475, 2015.$

[54] G. Yagawa and A. Oishi, "Feedforward Neural Networks," in *Computational Mechanics with Neural Networks*, pp. 11–23, Cham: Springer International Publishing, 2021.

[55] D. Bank, N. Koenigstein, and R. Giryes, “Autoencoders,” in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (L. Rokach, O. Maimon, and E. Shmueli, eds.), pp. 353–374, Cham: Springer International Publishing, 2023.

[56] M. Schuster and K. Paliwal, “Bidirectional Recurrent Neural Networks,” *Signal Processing, IEEE Transactions on*, vol. 45, pp. 2673–2681, Dec. 1997.

[57] J. H. Friedman, “Multivariate Adaptive Regression Splines,” *The Annals of Statistics*, vol. 19, pp. 1–67, Mar. 1991.

[58] Y. Freund and R. E. Schapire, “A Desicion-Theoretic Generalization of on-Line Learning and an Application to Boosting,” in *Computational Learning Theory* (P. Vitányi, ed.), (Berlin, Heidelberg), pp. 23–37, Springer, 1995.

[59] R. Ghosal and S. K. Ghosh, “Bayesian Inference for Generalized Linear Model with Linear Inequality Constraints,” *Computational Statistics & Data Analysis*, vol. 166, p. 107335, 2022.

[60] H. Chun and S. Keleş, “Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, pp. 3–25, Jan. 2010.

[61] I. Jolliffe, “Principal Component Analysis,” in *International Encyclopedia of Statistical Science* (M. Lovric, ed.), pp. 1094–1096, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[62] X. Jin and J. Han, “K-Means Clustering,” in *Encyclopedia of Machine Learning* (C. Sammut and G. I. Webb, eds.), pp. 563–564, Boston, MA: Springer US, 2010.

[63] H. A. Chipman, E. I. George, and R. E. McCulloch, “BART: Bayesian additive regression trees,” *The Annals of Applied Statistics*, vol. 4, Mar. 2010.

[64] Z. Ma, G. Hu, and M.-H. Chen, “Bayesian Hierarchical Spatial Regression Models for Spatial Data in the Presence of Missing Covariates with Applications,” *Applied Stochastic Models in Business and Industry*, vol. 37, no. 2, pp. 342–359, 2021.

[65] S. Indolia, A. K. Goswami, S. Mishra, and P. Asopa, “Conceptual Understanding of Convolutional Neural Network- a Deep Learning Approach,” *Procedia Computer Science*, vol. 132, pp. 679–688, 2018.

[66] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” *Advances in neural information processing systems*, vol. 30, pp. 3146–3154, 2017.

[67] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: Unbiased Boosting with Categorical Features,” in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[68] F. Chamroukhi, H. Glotin, and A. Samé, “Model-Based Functional Mixture Discriminant Analysis with Hidden Process Regression for Curve Classification,” *Neurocomputing*, vol. 112, pp. 153–163, July 2013.

[69] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY: Springer, 2009.

[70] R. Trevethan, “Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice,” *Frontiers in Public Health*, vol. 5, p. 307, Nov. 2017.

[71] A. Marandon, L. Lei, D. Mary, and E. Roquain, “Adaptive Novelty Detection with False Discovery Rate Guarantee,” Oct. 2023.

[72] O. Hyvärinen, “A Probabilistic Derivation of Heidke Skill Score,” *Weather and Forecasting*, vol. 29, pp. 177–181, Feb. 2014.

[73] G. K. Mbizvo and A. J. Larner, “On the Dependence of the Critical Success Index (CSI) on Prevalence,” *Diagnostics*, vol. 14, p. 545, Mar. 2024.

[74] M. J. Warrens and H. van der Hoef, “Understanding the Rand Index,” in *Advanced Studies in Classification and Data Science* (T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto, and M. Vichi, eds.), (Singapore), pp. 301–313, Springer Singapore, 2020.

[75] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression,” Apr. 2019.

[76] M. Jahanian, A. Karimi, N. Osati Eraghi, and F. Zarafshan, “Introducing the Cosine Clustering Index (CCI): A Balanced Approach to Evaluating Deep Clustering,” *SN Computer Science*, vol. 5, p. 687, July 2024.

[77] B. Baydil, V. H. de la Peña, H. Zou, and H. Yao, “Unbiased Estimation of the Gini Coefficient,” *Statistics & Probability Letters*, vol. 222, p. 110376, July 2025.

[78] A. Starczewski and A. Krzyżak, “Performance Evaluation of the Silhouette Index,” in *Artificial Intelligence and Soft Computing* (L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, eds.), (Cham), pp. 49–58, Springer International Publishing, 2015.