Wasit Journal for Pure Science

Journal Homepage: https://wjps.uowasit.edu.iq/index.php/wjps/index
e-ISSN: 2790-5241 p-ISSN: 2790-5233

Evaluation of Decision Tree and Support Vector Machine Classifiers in Comparison for Flood Prediction

Suha Abdullah Ahmed

Electrical And Computer Engineering, Institute of Graduate Studies, Altinbaş University, TÜRKİYE

*Corresponding Author: Suha Abdullah Ahmed

DOI: https://doi.org/10.31185/wjps.749
Received 10 February 2025; Accepted 23 March 2025; Available online 30 June 2025

ABSTRACT: Protection from floods is one of the most significant activities aimed at risk reduction. Flood prediction and disaster preparedness have relevance for reducing the association between floods and people and infrastructure. The goal of this study was to find out how well the Support Vector Machine (SVM) and Decision Tree classifiers work at predicting flooding based on different environmental and infrastructure factors. Data collection had been done through gathering 140 samples with a total of 21 variables where analysis had been done in order to identify significant contributory factors: topography, urbanization, and climate change. The results show that an SVM model was capable of achieving an accuracy of 91%, while the Decision Tree classifier did much better at an accuracy of 95%. The decision tree model was also more precise in flood prediction (1.00) and recall for non-flood cases (1.00), while for both models, the recall for flood cases was the same (0.88). This indicates that both the models had some false negatives for floods. The current study focuses more on machine learning applications and disaster readiness in flood risk assessment for better and more effective mitigation.

Keywords: Flood Prediction, Machine Learning, Support Vector Machine, Decision Tree, Disaster Preparedness

1. INTRODUCTION

As stated by Nurmasari (2019), water is an essential need for every human being and is essential to human existence [1]. Water is harvested and drained from upstream locations to downstream locations through the river [2]. The excess river water is hazardous as it may lead to floods and other natural disasters if there is no storage and purification system for water [3]. Flooding takes place when the water level rises above the normal threshold, whereby river water overflows and covers the nearby terrain. Floods most frequent in 2021 were from January to April, when there were 267 deaths and 501 recorded floods [4]. According to historical statistics, the worst flood disasters occurred in 2002, 2007, and 2013. 42 of the 168 subdistricts, or around 63.4% of DKI Jakarta's total subdistricts, were flooded in 2002. In DKI Jakarta, 16,041 hectares, or 24.25% of DKI Jakarta's total area, were flooded to a depth of 5 meters. The flood that year was tragic and resulted in the deaths of 21 people among a total population of 381,266.

There are numerous causes of floods, some of which include the poor filtration of water by poor soil, garbage that accumulates on riverbanks and disrupts the passage of water into the sea as well as elevating water levels at the downstream section of the river, and the intensity and length of extended rain [5].

A disaster, according to [6], is an occurrence or a chain of events endangering and disrupting the lives of people by either natural or human-made causes, resulting in damage to property, destruction of the environment, fatalities of humans or animals, and psychological effects. Disaster management is warranted to prevent or mitigate the effects of natural disasters like landslides, flooding, and earthquakes. In the post-disaster phase, disaster management entails emergency response, preparation, mediation or mitigation, rehabilitation, and reconstruction. "Mitigation" refers to any measure to lessen a disaster's effect or impact. There are two types of mitigating measures: structural and non-structural. The objective of structural mitigation is to minimize or eliminate the possibility of natural disasters. Policy measures, generation of awareness, enhancement of existing knowledge, and legislation are all examples of non-structural mitigation. The Gartner Group has been employed to define data mining as the act of digging through the majority of the data stored in storage media using pattern recognition tools, including mathematical and statistical methods, to create

new relations with meanings, habits, and patterns. To address the issue of information extraction from massive databases, data mining combines methods from various scientific fields, including machine learning, pattern recognition, statistics, databases, and visualization [7]. Data mining involves the extraction and discovery of applicable information and associated knowledge from various large-scale datasets by using statistical techniques, mathematics, artificial intelligence, and machine learning. Data mining, according to [8], is a series of processes that are utilized in investigating the value that is possible to extract from a data set in the form of previously unknown information. The following are key points of data mining based on the definitions provided [8]:

1. Data mining automatically uses existing data.
2. There is excessive data to be processed.

Identifying patterns or relationships that would offer valuable leads for things is the goal of data mining. Classification is one of the processes that might be conducted in data mining. When Carolus von Linne (otherwise known as Carolus Linnaeus) initially classified plants according to their physical attributes, classification was first used on plants that classified a particular species. As [9] points out, he is also referred to as the father of categorization. There are target categorical variables in categorization. Researchers have come up with the following models and methods to solve cases of categorization [9].

a. Decision tree, b. Classifier of Bayes/naive Bayes, c. Artificial neural network, d. Statistical analysis, e. Genetic algorithm, f. Rough sets, g. K-nearest neighbor classification, h. Rule-based method, i. Memory-based reasoning, and j. Support vector machine. Artificial intelligence has also been employed to overcome the limitation in some past research. Another related study is known as Comparison of Data Mining Methods for Rainfall Prediction. Using C4.5, Naïve Bayes, and KNN algorithms subjected a range of AI algorithms to the test of forecasting rainfall. With an 11.97% error rate and an 88.03% rate of correctness, the study results indicated that out of the three algorithms employed, the most correct rainfall forecast was performed by the C4.5 algorithm [4]. But yet another research work is "Comparison of Data Mining Techniques in Flood Prediction Using Naïve Bayes and KNN Algorithms." With 88.94% accuracy and an error rate of around 11.06%, the KNN algorithm performs better compared to the other algorithm in predicting floods, as shown by the final results of both algorithms [11]. "Comparison of Random Forest Classification Algorithm and Support Vector Machine on Prediction Status of BPS Work Unit's Disaster Mitigation and Preparedness Index (IMKB) in Indonesia for the Year 2020" is the title of the following paper with the same topic. The conclusions determine that Random Forest classification is superior to SVM in terms of accuracy, precision, and recall values. In particular, Random Forest was superior to SVM with 78.22% accuracy, 75.54% precision, and 76% recall [11]. A branch of computer science referred to as artificial intelligence (AI) seeks to create computer programs that can carry out tasks requiring human intelligence. AI claims to endow computers with human learning, planning, problem-solving, and decision-making capabilities [12]. Machine learning, which is an area of artificial intelligence, focuses on the creation of computer models and algorithms that are able to learn from data and experience without being coded [13]. To help computers make choices based on data, identify patterns, and forecast outcomes is the general purpose of machine learning [14].

A supervised learning technique called machine learning uses sample data—labeled input-output pairs in advance—to train the model. For instance, an image and a label for the objects in the image are fed into a model. Upon being provided with new inputs, the model can make new predictions or classifications after having learned to recognize patterns in data [15]. Linear regression, KNearest Neighbor, Naïve Bayes, Support Vector Machine (SVM), Random Forest, and Neural Network are just a few of the algorithms utilized within supervised learning. Machine learning models based on unsupervised learning are provided with unlabeled input data. Taha and others [16] A comprehensive set of data mining as well as machine learning algorithms has been used in order to analyze large amounts of complex medical data for assisting physicians in deciding heart diseases. The study aims at using Sequential Minimal Optimization (SMO), Naive Bayesian, Random Forest, Stochastic Gradient Descent (SGD), Multilayer Perceptron, Adaboost, and Logistic Regression as classifiers and also comparing its performance with the Heart Stalog Dataset. To improve the performance of the selected classifiers, the Harmony Search Algorithm (HSA) was utilized as effective feature selection tools. Feature selection is a well-known problem for classification and dimension reduction in high-dimensional data sets. Only the most relevant attributes from the data sets will be chosen in the process of feature selection. The outcomes show the accuracy was enhanced in Random Forest from 87.037% and AdaBoost from 84.0741% to 92.123% and 94.234%, respectively, after executing HSA on a minimum number of attributes. The goal is to find hidden patterns or structures in the data. Models have the ability to simplify data by lowering its dimensionality or consolidating similar data according to equivalent features [17]. These are some of the challenges that include wastewater treatment, overcoming several barriers to enhancing energy efficiency, operating in a more stringent water quality standards environment, and maximizing opportunities for resource recovery. Computational models have been increasingly set up as today's powerful tools that respond to these diverse issues towards promoting the operational and financial efficacy of the different WWTPs. In this study, the application of several artificial intelligence (AI) algorithms to wastewater treatment plants (WWTPs), energy optimization, inflows into WWTPs, anomaly detection, and predictions of effluent properties is involved. In the area of wastewater treatment AI algorithms, essential gaps and prospects for the future include the data-driven models' capacity for transfer learning [18]. The aim is to identify which model works better in allowing physicians to diagnose disease at the earliest possible time, minimizing the risk of disastrous outcomes.

2. THE PROPOSED METHODOLOGY

This approach outlines the process of comparing Support Vector Machine (SVM) and Decision Tree (DT) classifiers for flood prediction. Their performance is evaluated with respect to several different metrics against past hydrological and meteorological records. Collect data related to flood forecasting, such as historical rainfall, river flow rates, soil moisture, land use, topography, and meteorological indices. Data sources can include government agencies, weather stations, and remote sensing systems.

2.1 Data Pr-processing

Modeling relies heavily on the input data preparation. Modeling includes the assessment of input data quality and eventually enhancing the nature of inputs, selected stages, and time intervals. Modeling affects forecast outputs, like reliability and accuracy, directly. Pre-processing includes data cleaning. There must be a process of data cleaning in order to eliminate all the low-quality data, including the missing data.

Following data set gathering, preprocessing entails the assessment of data quality and, by extension, enhancing the input types and activities employed. It has a direct influence on the projection model's performance and accuracy.

Besides, raw data in the real world is not clean. There is noise. The big dataset register contains a lot of anomalous values that influence the conclusions of the study. Because great models are many times reliant on great data for data, the process of extraction, datasets are not always clean and can contain noise, too little data, redundant or incorrect data, and missing values. Lost data is a common error in many factual data types. Noise removal or elimination is employed to rectify lost values. The study applied the method of filling in missing character traits to meet the requirement of completeness. Errors in information like wrong feature values resulted in missing values for features. In this step, the result depends on data state; thus, a proper dataset is an essential factor in the outcome of categorized results, as this process decreases data complexity and offers better conditions for additional analysis. If the total number of missing data points is large enough, it is not possible to eliminate all the variables with missing values from the sample. In order to determine any null values, each row was compared to each column. If there are null values, then the mean value will replace all other values in the same time period in the data set of the column. Feature selection was the following task during preprocessing.

2.2 Data Cleaning

Missing Values: Use imputation methods or delete incomplete records.
Outlier Detection: Identify and resolve anomalies that may skew the results.
Normalization/Scaling: Especially applicable to SVMs because they are scale-sensitive features.

2.3 Feature Selection

Choose the most prominent features influencing flooding. Techniques such as correlation analysis. Feature selection plays a significant role in the performance of flood forecasting models. While Decision Trees provide high interpretability and training efficiency, SVM optimally generalizes by utilizing the best subsets of features. The choice of feature selection methods, e.g., correlation-based selection methods such as Pearson and Spearman correlation, can enhance model precision and performance. Pearson Correlation computes the linear correlation among features and removes highly correlated features. If the correlation between two features is high (above a specified threshold, e.g., 0.8), then one of them is removed to prevent redundancy and multicollinearity.

2.4 Data Splitting

A- Training and Testing Sets:

Divide the dataset into training and test subsets such that the test set is representative of the entire data distribution.

B- Cross-Validation:

Use techniques such as k-fold cross-validation to assess the stability of the model and reduce the influence of variance in training data.

2.5 Classification

A- Support Vector Classifier

One of the most popular supervised learning techniques that is highly regarded for its high effectiveness in classification problems is the Support Vector Classifier (SVC).

SVC has high performance and high flexibility and is therefore widely applied in many varied applications like medical diagnosis, text classification, and image classification. In this essay, the Support Vector Classifier and its basic principles, applications, and daily applications will be discussed in detail. A support vector classifier is a type of supervised learning method used for binary and multiclass classification. SVC aims to find the optimal hyperplane that maximally separates the classes in the feature space, which is different from the usual classification approach. To get

the maximum margin—that is, the distance between the hyperplane and the closest data points of both classes—the hyperplane is positioned relative to the support vectors [18].

Key Components of Support Vector Classifier:[19]

Hyperplane: A hyperplane is an option boundary in SVC that separates the classes in feature space. The binary classification hyperplane is a plane in two dimensions or a plane in three dimensions. In multiclass classification, different class combinations are separated through multiple hyperplanes.

Support Vectors: Support vectors are the closest points' positions to the hyperplane and also have a significant role in deciding its position. These decide the process of optimization that finds the best hyperplane and establishes the margin.

Kernel Trick: SVC can solve nonlinear classification problems thanks to the kernel trick, which transforms the input's original space to a higher dimension where there are linearly separable classes. Common kernel functions are the sigmoid polynomial with linear and radial basis functions (RBF).

The classification problem for the linearly separable training vectors xi is as defined below**:** *[20]*

$f(\mathbf{x}) = \text{sgn}(\boldsymbol{\omega}T\,\mathbf{x} + b)$

where ω is normal to the hyperplane and b is a bias term, which should satisfy the following conditions:

$$yi\ (\boldsymbol{\omega}T\ \mathbf{x}i + b) \geq 1,\ i = 1, 2, \ldots, l \qquad (1)$$

By finding the optimum separation hyperplane, SVc aims to set the margin to the maximum amount between positive and negative examples. Under the conditions in equation below, being a convex quadratic programming problem, the optimum separator hyperplane is that which achieves the minimum value of 1 ωT ω, with 2/ω being the margin. For the linear non-separable case, constraints of equation (2) are relaxed by introducing a new set of non-negative slack variables ξi i 1, 2, l as a constraint violation measure, as noted below)

$yi\ (\boldsymbol{\omega}T\ \mathbf{x}i + b) \geq 1 — \xi i\ ,\ i = 1, 2, \ldots, l$.

The following formula is minimized by the ideal hyperplane:

$$\frac{1}{2}\omega^t\omega + \lambda\sum_{i=1}^{l}\xi i \qquad (2)$$

where –λ is a parameter used to penalize variables ξi, subject to constraints in equation above. figure (1) shows the flowchart of the SVM

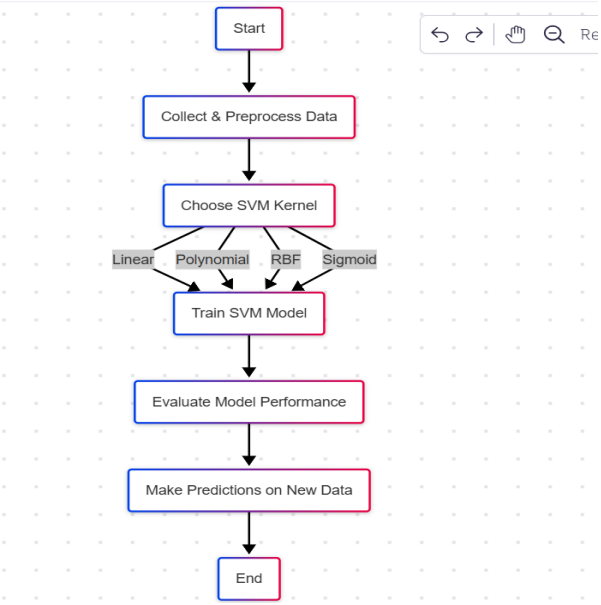

FIGURE 1. – SVM Flowchart

B- Decision Tree Classifier

One of the most popular techniques used for data classification is the decision tree classifier (DTC). The most important characteristic of DTC is the capability to decompose complex decision-making processes into the identification of a solution that is comprehensible and simpler to interpret [21].

Using decision trees, classification models with a tree structure can be established. A decision tree divides a dataset into smaller and smaller subgroups. A decision tree is constructed step by step in this manner. A decision A tree with decision nodes and leaf nodes is the ultimate output. A decision node includes at least two branches. A leaf node contains the classification decision. Typically, a node at the top of a tree for the best predictor is referred to as the root node. Decision trees can be used for both numerical DTs and categorical data [22] as predictive models that make use of a tree-like analysis of the data. They are particularly intended for supervised data mining. DT induction [23] includes two phases mainly. Once the entire tree is built, an algorithm is used to scan duplicate and repetitive subtrees. DT is

pruned if there is such a subtree. Pruned DTs are developed faster and more stable. Sub-regions are created within the data set via decision tree modeling [24].

C4.5 uses an algorithm based on information that is theoretically quantified in the form of "gain" and "gain ratio" to construct a decision tree from training data. All samples have the same format when given a training set (TS). By and large, the food product training set (TS) divides into two classes: acceptable level (AL) and unacceptable level (UL). Subsequently, the information (I) required to determine the class of a TS element is provided by

$$I(TS) = (|AL|)/(|TS|)\ \log2\ |(|AL|)/(|TS|) - (|UL|)/(|TS|)\log2\ (|UL|)/(|TS|) \quad (3)$$

$$\text{Entropy} = -\sum(p(i) * \log2(p(i))) \quad (4)$$

where p(i) is the proportion of data points in the set that belong to class i.

The information acquired on one feature is a distinction between the information which must be employed to decide on a TS element and the information required to discriminate a TS element when meaning for the function was previously established. Therefore, the knowledge attained on xk is

$$\text{Gain}(xk, TS) = I(TS) - I(xk,. TS)$$

To amend this flaw, an error-based method is used to prune the decision tree simply replace an entire subtree with a node with leaves [25]. figure (2) shows the flowchart

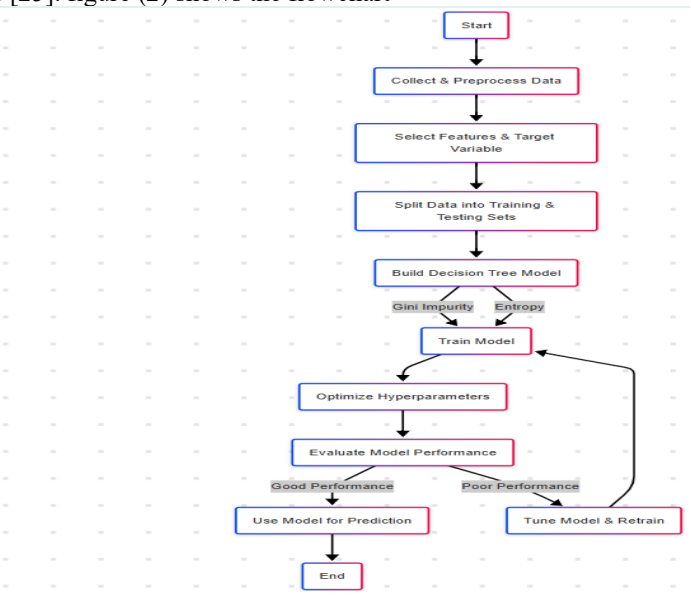

FIGURE 2. – SVM Flowchart

3. RESULTS

In this section, we will discuss the proposed steps to process flood disister classification using SVM and decision tree

3.1. Data Collection

140 rows and 21 columns make up the data set of the numerical flood calculation in this study. Every row consists of a sequence of factors that describe the flood. Agricultural practices, drainage systems, climate, topography, and drainage were among the factors that were gathered, as shown in Table 1.

Four artificial intelligence technology pre-processing procedures were used, and since the data was gathered from the site and compiled immaculately, there was no missing data. These procedures are as shown: Using Python, the initial portion of the

This made it difficult to collect the data, which was collected over a period of one year due to the challenges and obstacles that the researchers faced in order to collect the data in a way that allows it to be analyzed using artificial intelligence techniques. The data was collected so that all the parameters were present at the same time, and it was also present (whether there was a flood or not).

This is because artificial intelligence requires three steps for analysis (training, testing, and investigation).

Table 1. - Sample of the Data Set

Monsoon Intensity	**1**	**0**	**0**	**3**	**1**	**1**	**8**	**0**	**1**	**4**	**3**
Topography Drainage	8	7	1	4	3	4	3	0	2	2	7
River Management	1	4	0	0	2	1	1	1	8	3	2
Deforestation	8	4	7	1	0	4	2	7	1	1	0
Urbanization	0	8	3	4	4	2	3	1	4	8	0
Climate Change	4	8	7	8	4	4	7	1	1	0	3
Dams Quality	4	3	1	4	3	0	3	3	2	1	2
Siltation	3	1	1	7	3	0	4	1	4	1	3
Agricultural Practices	3	4	4	0	3	7	0	1	1	7	3
Encroachments	4	0	1	8	3	1	7	1	1	0	2
IneffectiveDisasterPreparedness	2	9	0	1	1	1	1	3	2	4	0
Drainage Systems	1	7	7	2	2	3	2	1	9	0	9
Coastal Vulnerability	3	2	3	4	2	1	1	3	2	3	1
Landslides	3	0	7	7	0	1	0	1	7	3	2
Watersheds	1	3	1	4	0	4	4	1	3	4	1
Deteriorating Infrastructure	4	1	0	4	4	4	1	8	4	4	4
Population Score	7	3	8	0	1	0	0	0	0	3	1
Wetland Loss	1	3	2	1	2	8	3	8	4	3	8
Inadequate Planning	7	4	3	7	3	3	4	1	1	1	8
Political Factors	3	3	3	1	1	2	0	0	1	0	1
Flood Probability	0	1	0	0	0	0	0	0	0	0	0

The dataset provides insights into the factors influencing flood probability, revealing both expected and unexpected patterns.

3.2 Applying SVM to the flood Classification

The data were split at random into training and test sets, 70%) allocated to training and 30% reserved for testing. In order to achieve a proper estimate and avoid the effects of partitioning data, classification techniques were all tested in terms of 10-fold cross-validation, over an average of 10 different partitions. This allowed classification performance measures to be compared in full.

3.3 Performance Evaluation of Model Applied

After partitioning the data into training and testing sets, the model development commenced using the training dataset, consisting of 98 cases. Figure (3) show the confusion matrix

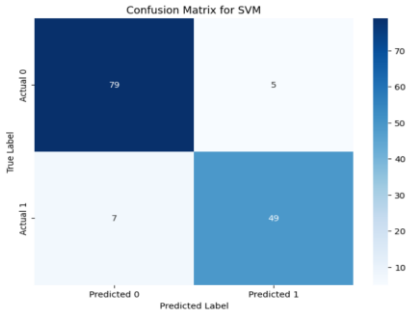

FIGURE 3. - Confusion matrix for SVM

Table 2. – SVM Accuracy

	precision	recall	f1-score	support
0	0.92	0.94	0.93	84
1	0.91	0.88	0.89	56
accuracy	0.91			
macro avg	0.91	0.91	0.91	140
weighted avg	0.91	0.91	0.91	140

Accuracy: 91% → The model correctly predicted 91% of all instances.

Macro Average (0.91) → Equal weight for both the classes, denoting balanced performance.

Weighted Average (0.91) → Accounts for the number of samples in each class, with generally good performance overall. The model is better but slightly prejudiced towards Class 0 (No Flood) against Class 1 (Flood) because the recall value for Class 1 is lower (0.88). If flood prediction is crucial, then recall for Class 1 (i.e., detection of true flood events) would be higher. Modifying the threshold of the decision, employing alternative algorithms, or addressing class imbalance (if it exists) might also improve predictions.

3.4 Applying Decision Tree to the flood Classification

The dataset was randomly split into training and test datasets, with 70%) allocated for training and the remaining 30% for testing. To ensure a fair assessment and mitigate data partitioning effects, all classification methods underwent evaluation using 10-fold cross-validation, averaged across 10 separate partitions. This approach allowed for a comprehensive comparison of classification performance metrics.

3.5 Performance Evaluation of Model Applied

After partitioning the data into training and testing sets, the model development commenced using the training dataset, consisting of 98 cases. Figure (4) show the confusion matrix

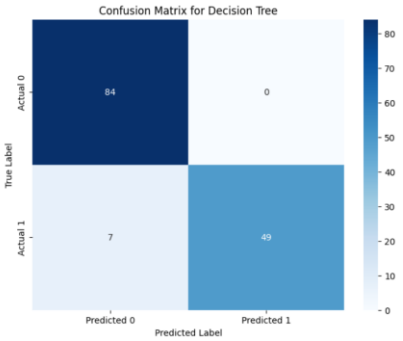

FIGURE 4. - Confusion matrix for Decision Tree

Table 3. – Decision Tree Accuracy

	precision	recall	f1-score	support
0	0.92	1	0.96	84
1	1	0.88	0.93	56
accuracy			0.95	140
macro avg	0.96	0.94	0.95	140
weighted avg	0.95	0.95	0.95	140

Accuracy: 95% → The model correctly predicted 95% of all instances.

Macro Average (0.96 Precision, 0.94 Recall, 0.95 F1-score) → Suggests an overall high-performing model.

Weighted Average (0.95 for all metrics) → Confirms balanced outstanding performance.

The model reliably identifies No Flood instances (100% recall for Class 0) but does not identify all actual flood instances (88% recall for Class 1).

Accuracy for floods (1.00) is very good, i.e., when the model predicts a flood, it's almost always correct.

But lower recall for floods (0.88) means that not all flood events are detected. This could be an issue in real-world cases where missing a flood event is costly.

3.6. Comparison between models

Table 4 displays a comparison solely for the testing dataset. This comparison identifies the superior classifier based on this dataset. The findings indicate that the ANN model outperformed LDA, demonstrating higher accuracy and efficiency in classification.

Table 4. - Comparison between models

Metric	SVM	Decision Tree	Difference
Accuracy	91%	95%	+4% (Improved overall performance)
Precision (No Flood - Class 0)	0.92	0.92	Same (Model maintains precision for non-flood cases)
Recall (No Flood - Class 0)	0.94	1	+6% (Perfectly classifies non-flood cases)
F1-Score (No Flood - Class 0)	0.93	0.96	3%
Precision (Flood - Class 1)	0.91	1	+9% (No false positives for floods)
Recall (Flood - Class 1)	0.88	0.88	Same (Still misses 12% of actual flood cases)
F1-Score (Flood - Class 1)	0.89	0.93	4%
Macro Average (Precision, Recall, F1-Score)	0.91, 0.91, 0.91	0.96, 0.94, 0.95	Improved across all
Weighted Average (Precision, Recall, F1-Score)	0.91, 0.91, 0.91	0.95, 0.95, 0.95	Improved across all

Better Accuracy (+4%), The decision tree classifier is more accurate in predicting flood and non-flood cases.

Perfect Recall for No Flood Cases

The new classifier never makes a mistake by misclassifying any non-flood cases (100% recall for Class 0).

This means better generalization for areas unlikely to flood.

Better Precision for Flood Cases

The new classifier predicts flood only when it is certain (100% precision).

No false alarms (false positives) for floods.

Same Recall for Floods (88%) The downside is that it still fails to detect 12% of actual floods (false negatives). This could be dangerous in real-life use where the failure to pick up a flood event has catastrophic consequences. (1)

4. CONCLUSION

To forecast flood occurrences based on different environmental and urbanization parameters, the present study compared the performance of Support Vector Machine (SVM) and Decision Tree classifiers. Both models are excellent, but the result shows that the Decision Tree classifier performs better than SVM with 95% accuracy compared to 91%. The decision tree model reduced false alarms by achieving 100% recall for non-flood cases and demonstrating greater precision in predicting floods. The occurrence of some flood events was excluded, however, since the models both registered an 88% recall in flood cases. The findings indicate that urban planning, efficient drainage systems, and emergency preparedness contribute significantly to flood mitigation. The study also indicates the need to select the best classification model to ensure memory and accuracy balance when forecasting floods. Future studies need to utilize deep learning techniques, hybrid models, and feature selection to improve flood recall rates. Moreover, blending real-time observation with climate forecast models can enhance flood threat assessment and response planning. Decision Tree method ruled out false positives as it showed optimum recall (1.00) for non-flood cases and higher accuracy (1.00) for the prediction of flood. Both algorithms struggled with recalling floods (0.88), i.e., 12% of actual floods were missed. This could potentially have significant operational implications for disaster management of flood. Among the principal conclusions of this study are the following:

1. Disaster preparedness is paramount; areas where disaster management systems fail are far more likely to have floods.
2. Flood hazards are influenced by urbanization and drainage facilities: Urban sprawl and inadequate drainage greatly contribute to the flood risk.
3. Climate change influences flood patterns: Future models need to take into account increasing rainfall intensity and climate-fueled extreme weather conditions.
4. Flood prediction requires model choice with utmost care. Decision trees are superior to SVM in terms of precision and accuracy, but there is a need for additional recall improvement for all cases of floods to be identified.
5. Real-time data integration is crucial. Real-time rainfall, river level observation, and land use changes could also add to precise flood forecasting.

REFERENCES

[1] ROSYIDA, Ainun, et al. Analisis Perbandingan Dampak Kejadian Bencana Hidrometeorologi dan Geologi di Indonesia Dilihat Dari Jumlah Korban (Studi: Data Kejadian Bencana Indonesia 2018): Studi: Data Kejadian Bencana Indonesia 2018. *Jurnal Dialog Penanggulangan Bencana*, 2019, 10.1: 12-21.
[2] KARNO, R.; MUBARRAK, J. Analsis Spasial (Ekologi) Pemanfaatan Daerah Aliran Sungai di Sungai Batang Lubuh Kecamatan Rambah Kabupaten Rokan Hulu. *Jurnal Ilmiah Edu Research*, 2018, 1.7: 59-62.
[3] AL FAUZI, Rahmat. Analisis tingkat kerawanan banjir Kota Bogor menggunakan metode overlay dan scoring berbasis sistem informasi geografis. *Geo Media: Majalah Ilmiah dan Informasi Kegeografian*, 2022, 20.2: 96-107.
[4] AKBAR, Jiwa; YUDONO, Muchtar Ali Setyo. Water Level Classification for Detect Flood Disaster Status Using KNN and SVM. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, 2024, 13.3: 298-302.
[5] ANGGRAINI, Asfilia Nova. *Prediksi status banjir sungai Ciliwung untuk deteksi dini bencana banjir menggunakan Artificial Neural Network Backpropagation*. 2022. PhD Thesis. Universitas Islam Negeri Maulana Malik Ibrahim.
[6] PRIBADI, Krishna S.; YULIAWATI, Ayu Krishna. Pendidikan Siaga Bencana Gempa Bumi Sebagai Upaya Meningkatkan Keselamatan Siswa (Studi Kasus Pada SDN Cirateun dan SDN Padasuka 2 Kabupaten Bandung). *KRISHNA_S_PRIBADI_-_ITB. Pdf*, 2008.
[7] LAROSE, Daniel T. An introduction to data mining. *Traduction et adaptation de Thierry Vallaud*, 2005, 23.
[8] BRAMER, Max. *Principles of data mining*. Springer, 2007.
[9] MARDI, Yuli. Analisa Data Rekam Medis untuk Menentukan Penyakit Terbanyak Berdasarkan International Classification Of Disease (ICD) Menggunakan Decision Tree C4. 5 (Studi Kasus: RSU. CBMC Padang). *UPI YPTK Padang*, 2014, 93.
[10] ZAMRI, David, et al. Perbandingan Metode Data Mining untuk Prediksi Banjir Dengan Algoritma Naïve Bayes dan KNN: Comparison of Data Mining Methods for Predition of Floods with Naïve Bayes and KNN Algorithm. In: *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*. 2022. p. 40-48.
[11] NURKHALIZA, Ayu Aina; WIJAYANTO, Arie Wahyu. Perbandingan Algoritma Klasifikasi Support Vector Machine dan Random Forest pada Prediksi Status Indeks Mitigasi dan Kesiapsiagaan Bencana (IMKB) Satuan Kerja BPS di Indonesia Tahun 2020. *Jurnal Informatika Universitas Pamulang*, 2022, 7.1: 54-59.
[12] LIRIWATI, Fahrina Yustiasari. Transformasi Kurikulum; Kecerdasan Buatan untuk Membangun Pendidikan yang Relevan di Masa Depan. *IHSAN: Jurnal Pendidikan Islam*, 2023, 1.2: 62-71.
[13] BOCK, Marius, et al. Improving deep learning for HAR with shallow LSTMs. In: *Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 2021. p. 7-12.

[14] DARMAYANTI, Irma, et al. Prediksi potensi siswa putus sekolah akibat pandemi covid-19 menggunakan algoritme k-nearest neighbor. *JST (Jurnal Sains dan Teknologi)*, 2021, 10.2: 230-238.
[15] RANI, Veenu, et al. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, 2023, 30.4: 2761-2775.
[16] TAHA, Mohammed A.; ALSAIDI, Saif Ali Abd Alradha; HUSSEIN, Reem Ali. Machine Learning Techniques for Predicting Heart Diseases. In: 2022 International Symposium on iNnovative Informatics of Biskra (ISNIB). IEEE, 2022. p. 1-6.
[17] TIBALDI, Simone, et al. Unsupervised and supervised learning of interacting topological phases from single-particle correlation functions. *SciPost Physics*, 2023, 14.1: 005.
[18] ZAKUR, Yahya, et al. Artificial intelligence techniques applications in the wastewater: A comprehensive review. In: E3S Web of Conferences. EDP Sciences, 2025. p. 03006.
[19] ALAM, Shamshe, et al. One-class support vector classifiers: A survey. *Knowledge-Based Systems*, 2020, 196: 105754.
[20] AASIM, Muhammad, et al. A comparative and practical approach using quantum machine learning (QML) and support vector classifier (SVC) for Light emitting diodes mediated in vitro micropropagation of black mulberry (Morus nigra L.). *Industrial Crops and Products*, 2024, 213: 118397.
[21] PRIYANKA; KUMAR, Dharmender. Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, 2020, 12.3: 246-269.
[22] PAYNE; MEISEL. An algorithm for constructing optimal binary decision trees. *IEEE Transactions on Computers*, 1977, 100.9: 905-916.
[23] GRĄBCZEWSKI, Krzysztof. *Meta-learning in decision tree induction*. Cham, Switzerland: Springer International Publishing, 2014.
[24] VINAYAK, Rashmi Korlakai; GILAD-BACHRACH, Ran. Dart: Dropouts meet multiple additive regression trees. In: *Artificial Intelligence and Statistics*. PMLR, 2015. p. 489-497.
[25] ROSS, Quinlan J., et al. C4. 5: programs for machine learning. *San Mateo, CA*, 1993.