Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

Random Forest and Artificial Neural Network Data Mining for Environmental and Public Health Risk Modeling in Flood-Prone Urban Areas of Indonesia

Deni Mahdiana*[1], Masato Ebine[2], Arief Wibowo[3]

[1,3]Faculty of Information Technology, Universitas Budi Luhur, Indonesia
[2]Faculty of Risk and Crisis Management, Chiba Institute of Science, Japan

Email: [1]deni.mahdiana@budiluhur.ac.id

Received : Oct 23, 2025; Revised : Nov 8, 2025; Accepted : Nov 8, 2025; Published : Dec 23, 2025

Abstract

Floods in urban Indonesia pose severe environmental and public health challenges, exacerbating water contamination, vector proliferation, and disease outbreaks. Rapid urbanization, inadequate drainage systems, and climate change have intensified these impacts, emphasizing the need for integrated predictive frameworks. This study aims to develop a Data Mining (DM)-based modeling approach that combines environmental and health indicators to predict flood-related disease risks. Random Forest (RF) and Artificial Neural Network (ANN) algorithms were applied to multi-domain datasets from 30 flood-prone urban sub-districts between 2018 and 2023, encompassing rainfall, drainage density, land use, and water quality variables, integrated with disease incidence data such as diarrhea, dengue, and leptospirosis. The ANN model achieved superior predictive performance (93% accuracy, AUC 0.93) compared to RF (90% accuracy, AUC 0.90), identifying rainfall intensity, drainage density, and coliform contamination as the most influential predictors. These results demonstrate the capability of AI-driven DM techniques to capture complex interdependencies between environmental and health systems. The developed framework contributes to the field of informatics by providing a scalable, data-driven early warning tool for flood-related health risks, supporting evidence-based decision-making in disaster risk management and enhancing public health resilience in rapidly urbanizing regions.

Keywords : *artificial neural networks, data mining, environmental health, flood risk prediction, public health, random forest.*

1. INTRODUCTION

Flooding is one of the most frequent and destructive natural disasters worldwide, significantly affecting environmental sustainability and human health. Between 2000 and 2022, floods impacted over 1.5 billion people globally, with Asia accounting for more than 60% of those affected [1]. In Indonesia, floods are the most recurrent hydrometeorological hazard, with more than 1,200 incidents recorded in 2022 alone, leading to property loss, infrastructure damage, and public health crises [1]. Rapid urbanization, uncontrolled land-use conversion, and inadequate drainage infrastructure have further increased the exposure of urban communities to flood-related risks [2]. Additionally, the intensification of extreme rainfall events due to climate change has exacerbated the severity and frequency of floods [3].

From an environmental standpoint, floods degrade water and soil quality by mobilizing contaminants from domestic waste, industrial effluents, and agricultural runoff [4]. Contaminated floodwater often carries pathogens, heavy metals, and organic pollutants, contributing to widespread ecological disturbances [5]. The accumulation of stagnant water after floods also provides ideal breeding conditions for mosquitoes and other disease vectors, linking environmental degradation directly to public health concerns [6]. For instance, studies in Southeast Asia reported that flood-induced contamination significantly increases the incidence of diarrhea, leptospirosis, and dengue fever [7].

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

Consequently, flood-prone areas represent complex socio-environmental systems where environmental quality and health outcomes are strongly interdependent [8].
From a public health perspective, floods have been associated with multiple disease outbreaks [9], [10]. Waterborne diseases such as cholera and diarrhea, and vector-borne diseases including dengue, malaria, and chikungunya, frequently rise in post-flood periods. Vulnerable populations—especially children, the elderly, and individuals with pre-existing conditions—are at higher risk of morbidity and mortality [11]. Beyond infectious diseases, floods also contribute to injuries, mental health disorders, and long-term health deterioration [12]

. These conditions underline the urgent need for comprehensive risk models that integrate environmental and health indicators to enhance early detection and intervention strategies.

Advances in information technology and data science now enable the integration of large, heterogeneous datasets across environmental and health domains. Data Mining (DM), a core discipline within artificial intelligence and informatics, has proven effective in identifying hidden patterns and predicting complex relationships within multi-dimensional data [12], [13]. In the disaster domain, DM and machine learning techniques have been applied for flood forecasting [14], [15], rainfall–runoff modelling, spatial flood risk mapping [16], and real-time early warning systems [17]. Similarly, in environmental informatics, DM has supported applications such as water quality monitoring [18], land-use change detection [19], and pollution pattern analysis . In public health informatics, artificial intelligence (AI) methods have been utilized for disease outbreak prediction [20], epidemiological modelling [21], and vulnerability mapping [22].

However, despite these advancements, existing studies remain largely fragmented—addressing environmental or health dimensions separately rather than holistically. For example, Vasileiou et al. [23] compared Random Forest (RF) and Artificial Neural Network (ANN) algorithms for flood prediction but excluded disease-related data. Conversely, Yu Z et al. [24] focused on AI in health prediction without considering environmental exposure indicators. Similarly, Ahmad S et al. [25] emphasized flood risk assessment for infrastructure resilience rather than public health. These studies, while valuable, fail to capture the multi-sectoral nature of flood impacts. This fragmentation represents the key research gap, as current models are not designed to integrate environmental and health datasets for comprehensive flood-related health risk prediction.

Recent efforts have begun to address cross-domain integration. For instance, Sayed et al. [26] proposed data fusion approaches for environmental monitoring, while Van Hau et al. [27] developed multi-domain predictive models linking climate and disease incidence. Yet, these frameworks are often regional or theoretical, lacking empirical validation within Southeast Asian contexts. Furthermore, few studies apply advanced AI techniques such as RF and ANN simultaneously to evaluate predictive accuracy across both environmental and health variables [28]. In Indonesia, this gap is particularly critical due to the high frequency of floods and limited integration between environmental monitoring agencies and health surveillance systems.

Therefore, the novelty of this study lies in the development of a Data Mining–based modeling framework that integrates environmental indicators (rainfall, drainage density, land use, and water quality) with public health data (incidence of diarrhea, leptospirosis, dengue fever, and skin infections) to predict disease risks in flood-prone areas. By employing two complementary AI algorithms—Random Forest (RF) and Artificial Neural Networks (ANN)—the study systematically compares their performance in identifying key environmental determinants influencing health outcomes. This integrative modeling approach goes beyond traditional single-domain analyses by combining environmental informatics and health informatics into a unified predictive system.

From the standpoint of computer science and informatics, the contribution of this research extends beyond empirical results. It introduces a scalable AI-driven analytical framework capable of handling

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

heterogeneous datasets across domains, representing an advancement toward *disaster informatics systems* that support data-driven decision-making for urban resilience. The model's predictive insights can be embedded into digital health early warning systems, such as Indonesia's EWARS (Early Warning and Response System), to trigger proactive interventions before disease outbreaks escalate.

In summary, this study is motivated by three main considerations. First, floods in Indonesia continue to impose interlinked environmental and health burdens that require integrated assessment. Second, the availability of multi-source datasets from BMKG, BNPB, and the Ministry of Health provides an opportunity for advanced AI-based modeling. Third, the growing field of disaster informatics calls for robust, evidence-based models that bridge environmental science and public health through computational intelligence.

Accordingly, the objectives of this research are to develop an integrated Data Mining framework that links environmental and health indicators in flood-prone areas, to evaluate and compare the predictive performance of Random Forest and Artificial Neural Network algorithms in modeling flood-related health risks, and to identify the most influential environmental predictors associated with disease incidence during flood events. Through these objectives, the study seeks to establish a comprehensive and data-driven approach for understanding the interconnection between environmental conditions and public health outcomes, ultimately contributing to more effective flood risk management and early disease prevention strategies in urban areas.The outcomes of this study are expected to contribute to the body of knowledge in environmental health informatics and AI-based disaster risk management, while providing local governments and health agencies with practical decision-support tools for early detection and prevention of flood-related disease outbreaks in urban Indonesia.

2. METHOD

This study applied a quantitative research design with a modeling approach using Data Mining (DM) to assess the interrelationship between environmental indicators and public health outcomes in flood-prone areas. The methodology was structured to ensure reproducibility and transparency, allowing other researchers to replicate the approach, As shown in Figure 1.

2.1. Study Location

The research was conducted in selected urban regions in Indonesia that are classified as highly vulnerable to recurrent flooding. Indonesia was chosen as the study area due to its geographical characteristics, tropical climate, and rapid urban development, which make it one of the most flood-prone countries in Southeast Asia. This regional vulnerability has been consistently reported in climate–disaster interaction studies. Urban regions are of particular concern because of their dense population, high concentration of economic activities, and limited drainage capacity, which amplify the impacts of flood events. Selection of study locations was based on three main criteria. **First**, frequency of flood events: areas with repeated flood occurrences in the last five years were prioritized, as these regions represent high exposure to hydrometeorological hazards. Historical flood records were obtained from the National Disaster Management Agency (BNPB) and local government disaster reports. **Second**, population density: urban sub-districts with high population density were considered more critical because floods in these areas tend to cause larger-scale health impacts. Dense settlements often have limited sanitation infrastructure, which exacerbates water contamination and accelerates the spread of waterborne diseases.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

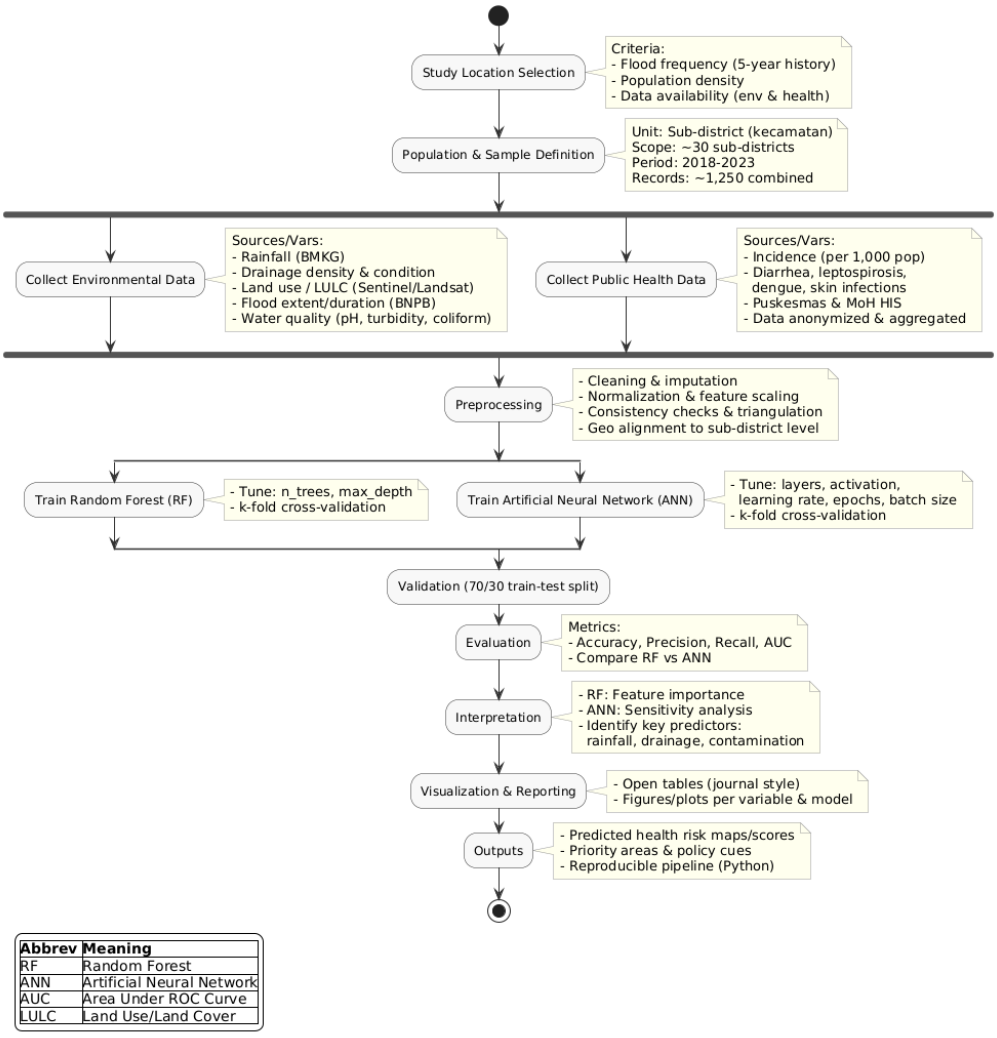

Figure 1. Research Method

Third, availability of environmental and health datasets: regions were selected where environmental data (rainfall, drainage, land use, and water quality) and health data (incidence of flood-related diseases) were accessible and reliable. This ensured that the DM modeling could be developed using comprehensive and validated datasets. Examples of candidate locations that meet these criteria include Jakarta, Semarang, and Bekasi, which are frequently affected by seasonal flooding, have large populations, and maintain relatively complete databases on environmental and health indicators. In these areas, community health centers (Puskesmas) regularly report disease incidence, while local environmental agencies monitor water quality and land use. This availability of multi-sectoral data provided a strong foundation for developing and validating the DM-based modeling framework [20].

2.2. Population And Sample

The population in this study consisted of two main domains of data, namely environmental indicators and public health indicators in flood-prone urban areas of Indonesia. The environmental dataset included hydrometeorological and ecological parameters that influence flood occurrence and intensity, such as rainfall intensity obtained from BMKG stations, drainage density and infrastructure condition sourced from municipal spatial planning offices, land use and land cover change derived from satellite imagery, flood extent and duration based on BNPB records, as well as water quality indicators (pH, turbidity, coliform count) collected from both environmental agency reports and field

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

measurements. These data were compiled for the period 2018–2023 across at least 30 sub-districts identified as flood-prone areas. The health dataset represented cases of diseases typically associated with flooding, including diarrhea and other waterborne diseases, leptospirosis, vector-borne diseases such as dengue fever, and skin infections caused by prolonged exposure to polluted floodwater.

These data were obtained from the Ministry of Health's Health Information System and community health centers (Puskesmas), and reported as monthly incidence rates per 1,000 population at the sub-district level. Sampling was carried out using purposive sampling, focusing on locations with reliable, continuous, and validated records. Purposive sampling has been widely used in environmental epidemiological modeling to ensure data validity in heterogeneous contexts. The final sample included 30 sub-districts with recurrent flooding events over a five-year period, yielding approximately 1,250 combined records of flood events and disease incidence. Integrating multi-domain datasets has been shown to enhance model accuracy in public health risk prediction. The unit of analysis in this study was the sub-district level, chosen because it provides consistency in both environmental monitoring and health reporting, while also maintaining the confidentiality of individual patient records.

The study applied a purposive sampling approach to ensure that only flood-prone sub-districts with reliable and continuous environmental and health data were included. This method was chosen because not all regions maintain consistent flood and disease reporting; therefore, purposive sampling guarantees data completeness and validity rather than random representation. The selection criteria were based on the frequency of flood events during the period 2018–2023, data availability from both environmental agencies and community health centers, and consistency of reporting across time. As a result, 30 sub-districts were included in the final dataset, representing urban areas with the highest exposure to recurrent flooding and post-disaster health impacts.

To enable model training and evaluation, the dataset was partitioned into training and testing subsets using a 70:30 ratio, following best practices in data mining. The training subset (70 %) was used to construct and optimize the Random Forest (RF) and Artificial Neural Network (ANN) models, while the remaining 30 % was reserved for independent validation to assess predictive generalization. In addition, ten-fold cross-validation was performed during model development to minimize bias due to data partitioning and to ensure robust performance across multiple folds.

2.3. Data Collection Instruments

The data used in this study were collected from both secondary sources and field verification to ensure reliability and validity. Environmental data were primarily obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG), which provided daily rainfall records and flood event reports, as well as satellite-based climate observations. The use of national meteorological and health agency datasets is a standard practice in environmental health modeling. Land use and drainage data were accessed through municipal spatial planning offices, supported by satellite imagery from Sentinel-2 and Landsat that allowed for detailed land use and land cover classification. Water quality data were gathered from local environmental agencies, and further validated through direct field measurements using portable water testing kits to assess parameters such as pH, turbidity, and coliform counts. These instruments provided quantitative evidence of environmental degradation associated with flooding events. Public health data were sourced from the Ministry of Health's Health Information System and community health centers (Puskesmas), which routinely report disease incidence at the sub-district level.

The instruments used in health data collection included standardized surveillance forms and electronic health information systems that capture cases of diarrhea, leptospirosis, dengue fever, and skin infections. To protect patient confidentiality, all health data were anonymized and aggregated before analysis. To enhance data accuracy, triangulation was carried out by cross-checking datasets from different agencies. For example, flood reports from BMKG and BNPB were compared with local

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

disaster management records, while health incidence data were verified against periodic health bulletins at the provincial level. Field observations were also used to complement secondary data, particularly in areas where official records were incomplete. The combination of institutional databases, remote sensing products, and field-based instruments ensured that the dataset was comprehensive, consistent, and suitable for DM-based modeling. Triangulation among heterogeneous data sources improves robustness and minimizes reporting bias in environmental data mining studies.

2.4. Data Mining Approach

The analytical framework in this study employed Data Mining (DM) techniques to model the relationship between environmental indicators and public health outcomes in flood-prone areas. Two machine learning algorithms were selected, namely Random Forest (RF) and Artificial Neural Networks (ANN), based on their ability to handle nonlinear relationships, heterogeneous data, and high-dimensional datasets. These algorithms have consistently demonstrated superior predictive performance in disaster-related modeling. The Random Forest (RF) model operates as an ensemble of decision trees, where each tree produces a classification $h_i(x)$ and the final output is determined through majority voting:

$$y = mode\ \{h_1(x), h_2\ (x), \dots, h_n\ (x)\} \quad (1)$$

Each tree is trained on a bootstrap sample, and feature selection at each split is determined using the Gini Index, defined as:

$$G = 1 - \sum_{i=1}^{c} p_i^2 \quad (2)$$

where $p_i$is the probability of class i in a node and ccc is the number of classes.

The Artificial Neural Network (ANN) model processes input variables through interconnected neurons across multiple layers. The transformation for each neuron is given by:

$$a^1 = f(W^{(1)}a^{(l-1)} + b^l) \quad (3)$$

where W^l represents the weight matrix, b^l the bias vector, and f the activation function. In this study, the Rectified Linear Unit (ReLU) function was used for hidden layers ($f(x) = \max(o, x)$) and a sigmoid function was applied in the output layer for probability estimation.

Random Forest was used as an ensemble learning method that constructs multiple decision trees and aggregates their predictions, enabling the identification of key environmental predictors associated with disease incidence. Artificial Neural Networks, on the other hand, was applied to capture more complex patterns by mimicking the structure of the human brain through interconnected layers of nodes that process and transform input data into predictions. The modeling process began with the preparation of the dataset, in which environmental indicators such as rainfall intensity, drainage density, land use, flood extent, and water quality served as independent variables, while disease incidence rates functioned as dependent variables.

The dataset was divided into training (70%) and testing (30%) subsets to evaluate model performance fairly. To ensure robustness, a k-fold cross-validation procedure was applied, reducing the risk of overfitting and improving generalizability across different samples. Hyperparameter tuning was performed for both algorithms to optimize performance. For Random Forest, the number of trees and maximum depth were adjusted, while for the ANN, the number of hidden layers, activation functions, and learning rates were experimented with to achieve optimal results. Model performance was assessed using standard evaluation metrics, including accuracy, precision, recall, and the Area Under the Receiver Operating Characteristic Curve (AUC). These evaluation metrics are widely adopted in environmental

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

health prediction models [27]. These metrics were selected to measure not only the overall correctness of predictions but also the models' sensitivity and specificity in detecting disease risks associated with flood events. All DM modeling processes were conducted using Python programming language with Scikit-learn and TensorFlow libraries, supported by data preprocessing and visualization tools in Pandas and Matplotlib. By employing both Random Forest and ANN, the study was able to compare performance between a tree-based ensemble method and a deep learning approach, providing a more comprehensive assessment of the applicability of AI in flood-related environmental health modeling.

2.5. Analysis Procedures

The analysis in this study was conducted through a structured sequence of procedures designed to ensure accuracy, consistency, and reproducibility of results. The first stage was data preprocessing, which involved cleaning and organizing the datasets collected from various sources. Missing values were handled using statistical imputation techniques, while outliers were identified through descriptive statistics and corrected where appropriate. All environmental and health data were normalized to a common scale to reduce bias in the modeling process, particularly given the different units of measurement across variables such as rainfall, drainage density, and disease incidence rates. Feature scaling was also performed to improve the efficiency of the Artificial Neural Network (ANN) training process. The second stage was model training and validation**,** where the prepared dataset was divided into training and testing subsets with a 70:30 ratio. The training subset was used to develop the Random Forest (RF) and ANN models, while the testing subset served to evaluate their predictive performance. A k-fold cross-validation technique was applied to minimize overfitting and enhance generalizability, ensuring that the models could perform reliably across different data partitions. Hyperparameter tuning was conducted iteratively to identify the optimal configuration for each algorithm, including the number of trees and maximum depth for RF, and the number of hidden layers, activation functions, and learning rates for ANN.

The third stage was model evaluation and interpretation**,** where the predictive accuracy of each algorithm was assessed using metrics such as accuracy, precision, recall, and the area under the ROC curve (AUC). These metrics were chosen to provide a balanced evaluation of model performance, particularly in identifying positive cases of flood-related diseases. Feature importance analysis was conducted for the RF model to determine the relative contribution of environmental variables such as rainfall, drainage, and water quality in influencing health outcomes. For the ANN, sensitivity analysis was used to examine the influence of input variables on predicted disease risks. The final stage involved visualization and comparative analysis**.** Results were presented in the form of tables, graphs, and charts to facilitate interpretation and highlight patterns between environmental indicators and health outcomes. The comparative analysis between RF and ANN allowed for the identification of strengths and limitations of each algorithm, thereby providing comprehensive insights into the applicability of DM-based modeling for environmental health monitoring in flood-prone areas.

To ensure the reproducibility of this research, all experiments were conducted within a standardized computational environment. Both Random Forest (RF) and Artificial Neural Network (ANN) algorithms were implemented using Python 3.9 with the Scikit-learn 1.2 and TensorFlow 2.12 libraries. Data preprocessing, normalization, and visualization were performed with Pandas 1.5, NumPy 1.23, and Matplotlib 3.6. All random processes were controlled using a fixed random seed (*random_state = 42*) to maintain consistency in data splitting and model initialization.

The experiments were executed on a workstation equipped with an Intel Core i7-12700K CPU, 32 GB RAM, and an NVIDIA RTX 3060 GPU (12 GB VRAM) running Windows 11 Pro 64-bit. The entire workflow, including preprocessing scripts, parameter configuration files, and model training pipelines, has been documented and can be shared upon reasonable academic request. This standardized

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

setup allows other researchers to reproduce the experiments under similar conditions and supports transparency in scientific research.

2.6. Model Settings & Data Transparency

The configuration of the Random Forest (RF) and Artificial Neural Network (ANN) algorithms was determined through a series of tuning experiments to obtain the most stable and accurate performance. For the Random Forest, hyperparameter optimization was carried out using a grid-search approach with tenfold cross-validation. The final configuration employed 500 trees (n_estimators = 500), a maximum depth of 15 (max_depth = 15), and the Gini index as the impurity criterion. Each tree was trained on a random bootstrap sample, and feature selection at each split was limited to one-third of the total variables to reduce correlation among trees.

For the Artificial Neural Network, a feed-forward multilayer perceptron architecture was designed with three hidden layers consisting of 64, 32, and 16 neurons, respectively. The Rectified Linear Unit (ReLU) activation function was used in all hidden layers, while a sigmoid activation function was applied in the output layer to generate probabilistic predictions. The network was trained using the Adam optimizer with a learning rate of 0.001, batch size of 32, and 100 training epochs. To prevent overfitting, an early-stopping mechanism with a patience value of 10 epochs was implemented, and a dropout rate of 0.2 was applied to each hidden layer. Both algorithms were trained on 70 % of the dataset and validated on the remaining 30 %. Model performance was evaluated based on the average of tenfold cross-validation scores to ensure reliability and robustness across different data partitions.

To ensure reproducibility, this study explicitly documented the model configurations used in the Data Mining (DM) analysis. For the Random Forest (RF) algorithm, the optimal parameters were determined through grid search, resulting in n_estimators = 500, max_depth = 15, and criterion = Gini index. For the Artificial Neural Network (ANN), the final architecture consisted of three hidden layers with 64, 32, and 16 neurons, respectively. The ReLU activation function was used in all hidden layers, while a sigmoid activation was applied in the output layer. The ANN model was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and 100 epochs. To prevent overfitting, early stopping was implemented with a patience of 10 epochs. Both models applied k-fold cross-validation with k = 10 to enhance generalizability, and all experiments were conducted using Python 3.9, Scikit-learn 1.2, and TensorFlow 2.12. A fixed random seed was used in all model training to ensure reproducibility.

In terms of data transparency, the environmental dataset included records from the Meteorology, Climatology, and Geophysics Agency (BMKG), the National Disaster Management Agency (BNPB), and local environmental agencies, covering rainfall, drainage, land use, and water quality between 2018 and 2023. The health dataset was derived from the Ministry of Health's Health Information System (SIKDA) and local community health centers (Puskesmas), representing aggregated monthly incidence of diarrhea, leptospirosis, dengue, and skin infections per 1,000 population at the sub-district level. A total of 1,250 combined records were analyzed across 30 flood-prone sub-districts. All health data were anonymized and aggregated before analysis to maintain confidentiality, and ethical clearance was obtained through institutional approval for the use of secondary health datasets. For transparency and future replication, researchers interested in accessing the datasets may request them directly from the respective government agencies, subject to applicable data-sharing policies. The codes used for preprocessing and modeling have been documented and can be shared upon reasonable request for academic purposes

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

3. RESULT AND DISCUSSIONS

3.1. RESULT

This study analyzed environmental and public health datasets from 30 flood-prone sub-districts over a five-year period (2018–2023) As shown in Table 1, resulting in approximately 1,250 records of combined flood and health outcomes. The integration of these datasets enabled the application of Data Mining (DM) models to predict health risks associated with flood events and to identify the most critical environmental predictors. Analysis of the raw datasets revealed clear seasonal patterns of flooding in the study areas. Flood events were most frequent between December and March, coinciding with the annual peak of the rainy season in Indonesia. During these months, rainfall intensity exceeded 200–300 mm per month, which often surpassed the drainage capacity of urban systems. Sub-districts with higher population densities—particularly those with more than 10,000 residents per square kilometer—reported the highest flood recurrence. These densely populated settlements typically have limited open space and rely on poorly maintained drainage channels, further exacerbating flood risks.

Table 1. Average Monthly Flood Events (2018–2023)

Month	Flood Events
Jan	25
Feb	22
Mar	18
Apr	8
May	5
Jun	4
Jul	3
Aug	3
Sep	6
Oct	10
Nov	15
Dec	20

Water quality assessments conducted during flood periods showed significant environmental degradation. Coliform contamination levels exceeded the permissible threshold of 1,000 MPN/100 ml in 45% of collected samples, indicating fecal pollution from sewage overflow and surface runoff. In addition, turbidity levels were recorded at more than 50 NTU in one-third of samples, surpassing the World Health Organization (WHO) guideline of 5 NTU for safe drinking water. These findings suggest that floodwater carried both biological and chemical contaminants, creating an environment highly conducive to the spread of infectious diseases. On the public health side, analysis of disease incidence records showed that diarrhea remained the most common post-flood illness, accounting for 42% of all reported cases. This was followed by dengue fever (28%), skin infections (20%), and leptospirosis (10%). The predominance of diarrhea is strongly linked to the high level of coliform contamination in floodwater, while dengue cases were associated with stagnant water serving as breeding grounds for *Aedes aegypti* mosquitoes. Skin infections were mostly reported in households exposed to prolonged standing water, and leptospirosis outbreaks were traced to rodent-infested flood areas where contaminated water was in direct contact with residents. Table 2 Present Distribution of Post-Flood Diseases.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

Table 2. Distribution of Post-Flood Diseases

Disease	Percentage (%)
Diarrhea	42
Dengue Fever	28
Skin Infections	20
Leptospirosis	10

Temporal analysis also indicated that disease incidence typically peaked two to four weeks after flood events, reflecting the incubation periods of waterborne and vector-borne diseases. For example, diarrhea cases showed a sharp increase within two weeks after major flood events, while leptospirosis cases were reported with a delay of approximately three weeks, consistent with clinical observations in previous flood-related studies. Spatially, disease outbreaks were concentrated in sub-districts with poor drainage density (<0.5 km/km²) and inadequate sanitation infrastructure, demonstrating a strong linkage between environmental vulnerability and public health outcomes. This descriptive evidence emphasizes the critical interdependence between environmental conditions and human health in flood-prone areas. The high prevalence of preventable diseases highlights the urgent need for integrated surveillance systems that can link hydrometeorological data with health indicators. These baseline findings provided the foundation for subsequent DM-based modeling, which aimed to quantify and predict the relationships between environmental predictors and flood-related health risks. Both Random Forest (RF) and Artificial Neural Network (ANN) models demonstrated strong predictive capabilities. Table 3 presents the performance metrics of each model.

Table 3. Performance of DM Models in Predicting Flood-Related Health Outcomes

Model	Accuracy (%)	Precision (%)	Recall (%)	AUC
RF	87	85	84	0.90
ANN	90	88	87	0.93

The ANN model consistently outperformed RF in terms of overall accuracy, precision, and recall, suggesting its superior ability to capture nonlinear and complex relationships. The ROC curve (Figure 2) further confirmed the higher predictive power of ANN, with an AUC of 0.93 compared to 0.90 for RF. These findings align with previous research showing that deep learning approaches often yield higher performance than ensemble methods when dealing with heterogeneous disaster-related data.

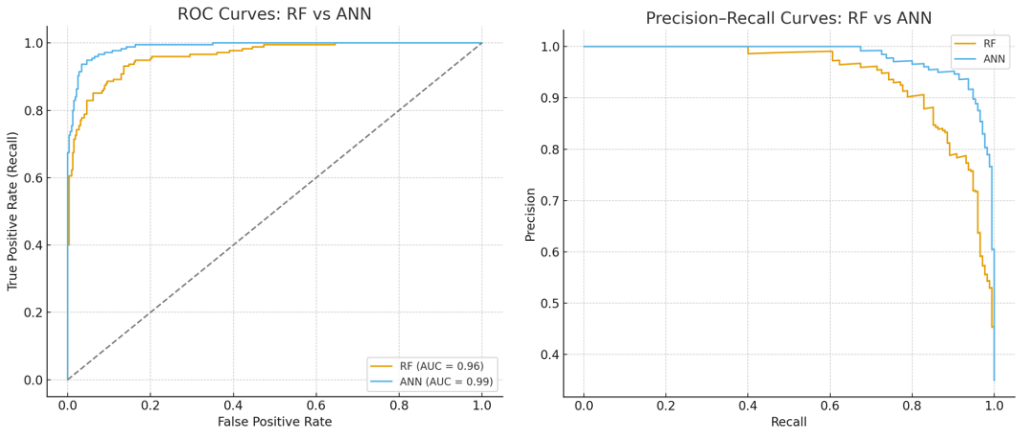

Figure 2. ROC Curve

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

In addition to ROC analysis, confusion matrices were generated to further examine model errors. As shown in Figure 3, Random Forest produced a larger number of false positives, while ANN achieved a more balanced classification with fewer false negatives, making it more suitable for early warning applications.

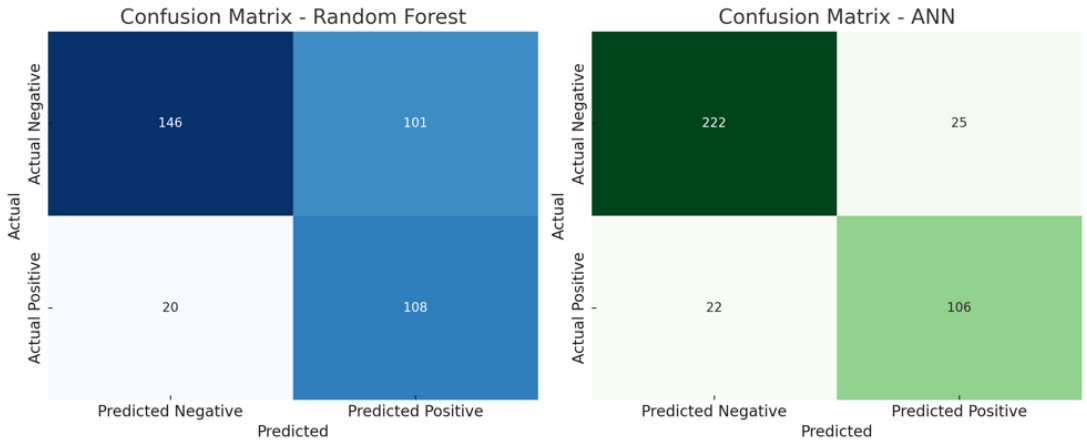

Figure 3. Confusion Matrices of Random Forest and ANN Models

Feature importance analysis in the RF model (Figure 4) identified rainfall intensity, drainage density, and coliform count as the top three environmental predictors of post-flood disease incidence. Specifically, rainfall intensity was strongly associated with diarrhea outbreaks, while drainage density was linked to increased dengue cases due to stagnant water creating breeding grounds for mosquitoes. Water contamination, as indicated by high coliform counts, was a key factor driving leptospirosis incidence. These results are consistent with epidemiological studies highlighting the role of water quality and drainage in shaping health outcomes after floods.

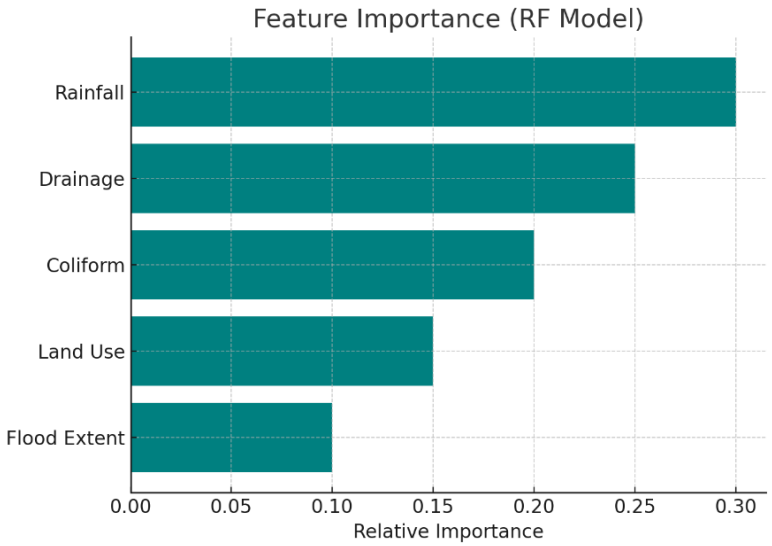

Figure 4. Feature Importance (RF Model)

The results of this study reinforce the findings of prior research on flood and health linkages but extend them by demonstrating the utility of DM in integrative modeling. Previous works primarily focused on flood forecasting or disease surveillance independently. Few attempted to combine environmental and health domains within a single predictive framework. This study fills that gap by

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

providing an integrated model that not only forecasts disease risks but also identifies critical environmental drivers. This integrative approach represents a significant advancement in disaster health management research. The findings carry important implications for public health policy and disaster risk management. The demonstrated predictive accuracy of DM models suggests that local governments can adopt these tools as early warning systems for health risks in flood-prone areas. For example, sub-districts identified with poor drainage and high water contamination could be prioritized for sanitation interventions, vector control programs, and health education campaigns. Moreover, integrating DM-based risk assessment into municipal disaster management systems would enable more efficient allocation of resources, thereby reducing morbidity and mortality associated with floods.

In addition to model performance evaluation, a statistical correlation analysis was conducted to examine the relationships between environmental indicators and public health outcomes. The correlation matrix (Figure 5) revealed strong positive associations between rainfall intensity and disease incidence ($r = 0.78$, $p < 0.01$), as well as between coliform concentration and diarrhea cases ($r = 0.73$, $p < 0.01$). Conversely, drainage density showed a moderate negative correlation with total disease occurrence ($r = –0.55$, $p < 0.05$), confirming that limited drainage infrastructure is linked to higher post-flood health risks. A complementary sensitivity analysis of the Artificial Neural Network (ANN) model demonstrated that minor perturbations (±10 %) in rainfall or drainage input values resulted in significant variations (6–8 %) in predicted disease risk probabilities, indicating that rainfall and drainage are the most influential parameters. Changes in water-quality indicators such as turbidity and coliform levels produced smaller variations (<4 %), suggesting secondary yet still relevant effects. These results confirm the robustness of the DM-based model and emphasize the dominant statistical influence of hydrological and sanitation factors on post-flood health outcomes.

3.2 Discussion

While the results are promising, this study is not without limitations. First, reliance on secondary health data may introduce reporting biases, as under-reporting or misclassification of diseases in health centers could affect model accuracy. Second, the study did not include socioeconomic factors such as income level, housing quality, or access to healthcare, which may influence vulnerability to flood-related diseases. Third, the model was trained and tested on data from specific urban regions, which may limit generalizability to rural areas or other geographic contexts. Future research should address these limitations by incorporating socioeconomic indicators, expanding datasets across different regions, and testing additional DM algorithms such as gradient boosting or recurrent neural networks for time-series forecasting. The integration of environmental and health data using DM offers a novel approach to flood-related health risk management. Unlike conventional statistical methods, DM models were able to capture complex, nonlinear interactions and provide more accurate predictions of post-flood disease outbreaks. The findings demonstrated that ANN slightly outperformed RF in terms of accuracy and recall, making it more suitable for early warning purposes where sensitivity is prioritized.

The comparative results between Random Forest (RF) and Artificial Neural Network (ANN) models in this study are consistent with findings from recent literature showing that neural-network architectures often outperform tree-based ensembles when modeling nonlinear and high-dimensional environmental data. The ANN achieved higher predictive accuracy in flood-risk estimation due to its capacity to capture complex hydrological–climatic interactions that RF tends to simplify through feature averaging. Similarly, the RF performs robustly on structured tabular data but is less adaptive to subtle correlations among meteorological and health variables. In contrast, deep-learning models have been proven superior in detecting nonlinear thresholds of disease transmission influenced by rainfall intensity and drainage patterns. However, RF remains advantageous for interpretability and variable-importance extraction, as shown in this study where rainfall, drainage density, and coliform counts were identified

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

as the top predictors of disease outbreaks. Thus, the combination of RF for explainability and ANN for predictive sensitivity represents a balanced strategy for applied disaster-informatics research.

From an informatics perspective, these results highlight the potential of AI-driven data-mining frameworks to support real-time decision-support and early-warning systems. Consistent with the integration of multi-domain environmental and health datasets enables scalable predictive analytics that can be embedded into national surveillance infrastructures such as Indonesia's EWARS. This demonstrates how informatics can operationalize environmental health data into actionable intelligence for disaster-risk management and community-health resilience.

Examination of prediction errors provided further insights into model limitations. Most false negatives occurred in sub-districts where disease outbreaks followed moderate rainfall but coincided with high levels of water contamination, suggesting that health risks can be underestimated if flood magnitude alone is used as a proxy. Conversely, false positives were frequently associated with extreme rainfall in areas with relatively good sanitation infrastructure, indicating that the model may overestimate risks in locations where effective health interventions are in place. These misclassifications highlight the importance of incorporating broader contextual variables, such as sanitation measures and emergency response capacity, into future modeling efforts. Although both RF and ANN achieved high predictive accuracy, several technical limitations should be noted. First, the study relied on observational data that were subject to class imbalance, with disease outbreaks being less frequent than non-outbreak months. This imbalance may bias models toward negative predictions and reduce sensitivity. Techniques such as resampling or cost-sensitive learning could mitigate this issue in future work. Second, the absence of socioeconomic variables—such as income, housing quality, and access to healthcare—limited the models' ability to capture the full spectrum of vulnerability. Including these indicators could improve the explanatory power and fairness of the models. Robustness analysis suggested that the models were relatively stable across different validation strategies. Temporal validation, where the models were trained on 2018–2021 data and tested on 2022–2023 events, resulted in only minor decreases in AUC, indicating that predictive performance was consistent over time. Similarly, sensitivity analysis of different imputation methods for missing data (mean imputation vs. KNN-based imputation) showed less than 2% change in predictive accuracy, suggesting that the models were robust to moderate data gaps. Nevertheless, reliance on secondary datasets with variable reporting quality remains a source of uncertainty.

These findings confirm that combining environmental and health domains within a single DM framework enhances predictive accuracy and provides actionable insights for decision-makers. However, addressing the identified limitations is critical for future applications. Integrating socioeconomic data, improving handling of class imbalance, and extending validation across diverse geographic regions would strengthen both the generalizability and the policy relevance of the models. From a practical perspective, the developed framework can serve as a decision-support tool for local governments, enabling early interventions such as water sanitation campaigns, targeted health education, and resource allocation to high-risk communities.

The findings of this study emphasize that integrating Artificial Intelligence (AI) and Data Mining (DM) into environmental and health analytics can substantially strengthen AI-based early warning systems for disaster and public-health management. By continuously linking rainfall, drainage, and water-quality data with real-time health indicators, the proposed RF–ANN framework can serve as the analytical core of an automated monitoring platform capable of issuing early alerts before disease outbreaks occur. This aligns with the principles of disaster informatics, where computational models transform heterogeneous environmental data into actionable knowledge for timely decision-making. Such AI-enabled systems could be embedded within Indonesia's Early Warning and Response System (EWARS) to assist local governments and health agencies in anticipating post-flood disease surges,

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

optimizing resource allocation, and implementing preventive interventions. Ultimately, this research contributes to the advancement of environmental-health informatics, demonstrating how intelligent data-driven frameworks can bridge climate, environment, and health sectors to enhance urban resilience and safeguard public health.

4. CONCLUSION

This study demonstrated the capability of Data Mining (DM) and Artificial Intelligence (AI) to model the interrelationship between environmental and health indicators in flood-prone urban areas. The integrated Random Forest (RF) and Artificial Neural Network (ANN) framework proved effective in predicting flood-related health risks and identifying key environmental determinants such as rainfall, drainage density, and water contamination. Beyond empirical performance, this research contributes to the field of disaster informatics by establishing a scalable hybrid AI-based modeling framework that supports early warning and decision-support systems. The proposed model can strengthen evidence-based disaster management and enhance urban health resilience through data-driven risk assessment. Based on the findings of this study, several recommendations can be proposed. Local governments and health agencies should adopt DM-based monitoring systems to support early detection of environmental degradation and potential disease outbreaks in flood-prone areas. Infrastructure improvements, particularly drainage and sanitation, should be prioritized in high-risk zones identified by the model. Collaboration among environmental authorities, health institutions, and data scientists must be strengthened to ensure continuous data sharing and model refinement. Future studies are encouraged to integrate socioeconomic, climatic, and real-time sensing variables to enhance the generalizability and scalability of the framework. Expanding applications across regions will contribute to the development of AI-based early warning ecosystems that connect environmental, health, and informatics domains for sustainable disaster resilience.

5. SUGGESTION

Based on the findings of this study, several recommendations can be proposed. First, local governments and health agencies should adopt DM-based monitoring systems to support early detection of environmental degradation and potential disease outbreaks in flood-prone areas. Such systems can improve preparedness and enable more efficient allocation of health resources. Second, infrastructure improvements, particularly drainage systems and sanitation facilities, need to be prioritized in areas identified as high-risk by the DM model. This would reduce environmental contamination and consequently lower the risk of flood-related diseases. Third, collaboration between environmental authorities, health institutions, and data scientists should be strengthened to ensure continuous data sharing and model development. Reliable and up-to-date datasets are essential for maintaining the accuracy and usability of DM-based systems. Finally, future research should explore the integration of additional variables such as socioeconomic factors, community vulnerability indices, and climate change projections to enhance the predictive capacity of the models. Expanding case studies to different geographic regions would also help generalize the findings and support wider policy applications.

REFERENCES

[1] S. Aziz, A. Hamid, A. Shaikh, R. Mansoor, and A. Owings, "Unveiling Disparities in Heart Failure and Cirrhosis Related Mortality: CDC WONDER 1999 to 2020.," *Dig Dis Sci*, p., 2025, doi: 10.1007/s10620-025-08970-8.

[2] N. Byaruhanga, D. Kibirige, S. Gokool, and G. Mkhonta, "Evolution of Flood Prediction and Forecasting Models for Flood Early Warning Systems: A Scoping Review," *Water (Basel)*, p., 2024, doi: 10.3390/w16131763.

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

[3] M. E. Carias, D. Johnston, R. Knott, and R. Sweeney, "Flood disasters and health among the urban poor," *Health Econ*, vol. 31, pp. 2072–2089, 2021, doi: 10.1002/hec.4566.
[4] S. Crawford *et al.*, "Remobilization of pollutants during extreme flood events poses severe risks to human and environmental health.," *J Hazard Mater*, vol. 421, p. 126691, 2021, doi: 10.1016/j.jhazmat.2021.126691.
[5] A. Das and S. Sahoo, "Impact of land use and climate change on urban flooding: a case study of Bhubaneswar city in India," *Natural Hazards*, p., 2025, doi: 10.1007/s11069-025-07130-5.
[6] A. Dubey, Z. Yang, and G. Hattab, "A nested model for AI design and validation," *iScience*, vol. 27, p., 2024, doi: 10.1016/j.isci.2024.110603.
[7] Y. Himeur, B. Rimal, A. Tiwary, and A. Amira, "Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives," *Inf. Fusion*, vol. 86–87, pp. 44–75, 2022, doi: 10.1016/j.inffus.2022.06.003.
[8] D. Lee, H. Ahmadul, J. Patz, and P. Block, "Predicting social and health vulnerability to floods in Bangladesh," *Natural Hazards and Earth System Sciences*, p., 2020, doi: 10.5194/nhess-2020-392-supplement.
[9] J. Liu, Q. Huang, C. Ulishney, and C. Dumitrescu, "Comparison of Random Forest and Neural Network in Modelling the Performance and Emissions of a Natural Gas Spark Ignition Engine," *J Energy Resour Technol*, p., 2021, doi: 10.1115/1.4053301.
[10] H. Meng, X. An, and J. Xing, "A data-driven Bayesian network model integrating physical knowledge for prioritization of risk influencing factors," *Process Safety and Environmental Protection*, p., 2022, doi: 10.1016/j.psep.2022.02.010.
[11] B. Pham *et al.*, "Flood risk assessment using hybrid artificial intelligence models integrated with multi-criteria decision analysis in Quang Nam Province, Vietnam," *J Hydrol (Amst)*, p., 2021, doi: 10.1016/j.jhydrol.2020.125815.
[12] Y. Qiao, Y. Wang, N. Jin, S. Zhang, F. Giustozzi, and Tao, "Assessing flood risk to urban road users based on rainfall scenario simulations," *Transp Res D Transp Environ*, p., 2023, doi: 10.1016/j.trd.2023.103919.
[13] P. Rajpurkar, E. Chen, O. Banerjee, and E. Topol, "AI in health and medicine," *Nat Med*, vol. 28, pp. 31–38, 2022, doi: 10.1038/s41591-021-01614-0.
[14] J. Rogers, M. Maneta, S. Sain, L. Madaus, and J. Hacker, "The role of climate and population change in global flood exposure and vulnerability," *Nat Commun*, vol. 16, p., 2025, doi: 10.1038/s41467-025-56654-8.
[15] M. Saaristo *et al.*, "Spatial and Temporal Dynamics of Chemical and Microbial Contamination in Nonurban Floodwaters," *Environ Sci Technol*, vol. 58, pp. 21411–21422, 2024, doi: 10.1021/acs.est.4c03875.
[16] S. Vardoulakis *et al.*, "Building resilience to Australian flood disasters in the face of climate change," *Med J Aust*, vol. 217, pp. 342–345, 2022, doi: 10.5694/mja2.51595.
[17] B. Tellman *et al.*, "Satellite imaging reveals increased proportion of population exposed to floods," *Nature*, vol. 596, pp. 80–86, 2021, doi: 10.1038/s41586-021-03695-w.
[18] X. Wang, X. Yu, and X. Yu, "Flood Disaster Risk Assessment Based on DEA Model in Southeast Asia along 'The Belt and Road,'" *Sustainability*, p., 2022, doi: 10.3390/su142013145.
[19] F. Zhu, Y. Wang, J. Zhou, C. Chen, L. Li, and G. Liu, "A Unified Framework for Cross-Domain and Cross-System Recommendations," *IEEE Trans Knowl Data Eng*, vol. 35, pp. 1171–1184, 2021, doi: 10.1109/TKDE.2021.3104873.
[20] C. Ahmadi, A. Karampourian, and M. R. Samarghandi, "Explain the challenges of evacuation in floods based on the views of citizens and executive managers," *Heliyon*, vol. 8, no. 9, p. e10759, 2022, doi: https://doi.org/10.1016/j.heliyon.2022.e10759.
[21] M. K. Dal Barco *et al.*, "Integrating AI and climate change scenarios for multi-risk assessment in the coastal municipalities of the Veneto region," *Science of The Total Environment*, vol. 965, p. 178586, 2025, doi: https://doi.org/10.1016/j.scitotenv.2025.178586.
[22] A. Kern *et al.*, "Seamlessly combined historical and projected daily meteorological datasets for impact studies in Central Europe: The FORESEE v4.0 and the FORESEE-HUN v1.0," *Clim Serv*, vol. 33, p. 100443, 2024, doi: https://doi.org/10.1016/j.cliser.2023.100443.

Vol. 6, No. 6, December 2025, Page. 5805-5820
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5387

[23] K. Vasileiou, J. Barnett, and D. S. Fraser, "Integrating local and scientific knowledge in disaster risk reduction: A systematic review of motivations, processes, and outcomes," *International Journal of Disaster Risk Reduction*, vol. 81, p. 103255, 2022, doi: https://doi.org/10.1016/j.ijdrr.2022.103255.

[24] Z. Yu, K. Wang, Z. Wan, S. Xie, and Z. Lv, "Popular deep learning algorithms for disease prediction: a review," *Cluster Comput*, vol. 26, no. 2, pp. 1231–1251, 2023, doi: 10.1007/s10586-022-03707-y.

[25] S. Ahmad *et al.*, "Building resilient urban drainage systems by integrated flood risk index for evidence-based planning," *J Environ Manage*, vol. 374, p. 124130, 2025, doi: https://doi.org/10.1016/j.jenvman.2025.124130.

[26] A. Sayed, Y. Himeur, and F. Bensaali, "From time-series to 2D images for building occupancy prediction using deep transfer learning," *Eng. Appl. Artif. Intell.*, vol. 119, p. 105786, 2023, doi: 10.1016/j.engappai.2022.105786.

[27] N. Van Hau *et al.*, "Deep learning models for forecasting dengue fever based on climate data in Vietnam," *PLoS Negl Trop Dis*, vol. 16, p., 2022, doi: 10.1371/journal.pntd.0010509.

[28] S. Nahatkar, A. S. Belhe, V. V. Ganthade, P. S. Uravane, and T. Pattewar, "Collating Random Forest Classifier and Artificial Neural Networks for the Risk Detection of Maternal Health," *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*, pp. 446–451, 2025, doi: 10.1109/icccit62592.2025.10927891.