Journal of Intelligent Information and Communication

Research Paper

Feature Selection Random Forest Model for Flood Prediction in Watershed Sensor Networks

Adéla Svoboda[1, *] and Jakub Dvořák[1]

[1] Department of Computer Systems, Brno University of Technology, 61669 Brno, Czech Republic
*Corresponding author: adela.s@fit.vut.cz

Abstract. Real-time flood forecasting is impacted by non-linear rainfall-runoff behavior, redundancy of hydrological parameters, and noise in a watershed's sensor network. This paper proposes a feature-selection random forest model for flood risk prediction based on multi-source watershed sensors. The variables for rainfall, upstream water level, discharge, soil moisture, terrain reaction, and transmission quality should first be arranged as time-window descriptors. Before Random Forest training, unstable and overlapping variables are removed using a relevance-redundancy-importance approach. The suggested model is evaluated experimentally against baselines for support vector machines, gradient boosting, traditional Random Forest, LSTM, and XGBoost under normal operation, missing observations, sensor noise, and delayed transmission. On the generated watershed dataset, the suggested model obtained 96.1% accuracy, 95.4% macro-F1, and 94.8% flood-event recall while reducing the feature dimension by 58.7% and maintaining an average inference latency of 7.8 ms. Therefore, a minimal number of hydrological features are picked to increase the prediction accuracy and interpretability. A model for a watershed flood warning system that needs rapid risk classification and sensor-level explanations is presented as an embeddable machine-learning framework.

Keywords: *Watershed Sensor Network, Flood Prediction, Random Forest, Hydrological Forecasting, Real-Time Warning*

Received on 22 October 2022, Accepted on 14 February 2023, Published on 23 February 2023

Introduction

Now that many are aware of the disaster risk, they want to know how to lessen the risk of floods in their lives. All aspects of a river basin, including rainfall gauges, water-level stations, flow meters, soil-moisture probes, upstream reservoir data, and low-power communication nodes, may now be monitored thanks to the growth of the hydrological sensor network. Before a flood peak occurs, the sensors continuously gather data to explain the interactions between rainfall intensity, antecedent moisture, terrain reaction, and channel storage [1]. Sensor networks may perform a fine-grained temporal resolution short-term warning based on a large number of stations and offer data streams for real-time flood risk classification [2]. In order to capture non-linear hydrological interactions and avoid explicitly calculating all of the physical runoff equations, machine learning has been used in flood forecasting [3]. Data-driven models are ideal for small- and medium-sized watersheds that urgently need flood response and have a limited lead time [4].

However, because of the data's variability, redundancy, and frequent incompleteness, flood prediction from watershed sensor networks remains challenging. Soil moisture may reflect accumulated antecedent circumstances rather than immediate flood risk, rainfall intensity can change abruptly in a matter of minutes, and the upstream water level responds slowly [5]. Wireless sensor networks also include difficulties such as packet loss, battery-induced sampling gaps, calibration drift, and anomalous spikes produced by debris, sediment or momentary communication breakdown [6]. Train the typical machine learning model using only the retrieved hydrologic parameters, and the performance may drop if many other irrelevant or strongly associated

variables are introduced. Random Forest is robust to non-linear interactions and mixed-scale features, but if there are redundant features, the model size will increase, feature interpretability will be lowered, and the relevance will be spread over variables that describe roughly the same hydrological process [7]. Research on flash-flood vulnerability and ensemble learning has also revealed that, in reality, feature quality is sometimes more essential than the selection of a classifier [8].

Random Forest Model for Feature Selection in Watershed Sensor Networks for Flood Prediction. In the suggested architecture, create time-window hydrological descriptors from several sensor data sources, evaluate each candidate feature for relevance, redundancy, and embedded tree importance, and then train a tiny Random Forest classifier to predict flood-risk categories. Maintain a low-latency inference speed for real-time warnings, decrease the number of redundant sensing variables, increase the recall rate of flood occurrences, and suppress false alarms. Connect the chosen features to hydrological significance so that the model can estimate flood risk accurately and provide explanations for warnings, such as which signals related to rainfall, water level, discharge, and soil moisture are causing the alert.

Related Work

Hydrological Sensor Network Modeling

In the case of a flood, flood monitoring devices may now continuously gather hydrological data. Rainfall stations, river-stage sensors, ultrasonic water-level gauges and soil-moisture probes can be used to monitor changes in hydrology before a substantial increase in discharge. Distributed monitoring is ideal for urban drainage areas with a short runoff concentration time and hilly watersheds, according to research on wireless sensor networks [9]. The data-collection, transmission and warning modules of an IoT flood-prediction system can be deployed to assist machine-learning models with near real-time data rather than delayed manual records [10]. Although cloud-based flood prediction systems have improved the effectiveness of centralized analysis and edge sensing, these systems are still limited by communication delays and local data stability [11].

For this reason, numerous physical phenomena and time dependencies are seen by watershed sensors. Rainfall variables are event-driven, upstream water level is delayed by channel routing, soil moisture records antecedent wetness, and discharge sensors may be impacted by hydraulic backwater or debris. Therefore, flood-stage forecasting research has concentrated on the creation of time windows and lagged feature representations [12]. The issue of local flooding brought on by increased rainfall intensity, changes in land cover, decreased drainage capacity, and uneven topography can be addressed by combining machine learning and GIS variables, according to studies on urban flood prediction [13]. Therefore, rather than employing raw measurements directly, an organized structure of sensor observations should be used to improve flood prediction ability.

Remote Sensing and flood-risk maps also offer supporting data for the impact of topography and space. Based on remote sensing and topographical data, flood-risk mapping has been carried out using Random Forest and ensemble learning, which have demonstrated good results in finding nonlinear flood-prone areas [14]. Analysis of meteorological parameters demonstrates that different Areas will have varied amounts of precipitation, runoff concentration, slope, etc., and varying drainage densities [15]. Since susceptibility maps are often geographical and event-independent, classification based on streaming observations is necessary for real-time dynamic risk prediction for watershed sensors. Therefore, the machine-learning model for sensor networks should blend event-level temporal descriptors with stable basin properties.

Data Mining Methods for Flood Prediction

Numerous machine learning applications have been used in various regions for runoff modeling, water level prediction, flood forecasting, and susceptibility evaluation. Compared to Support Vector Machines or Neural Networks, Random Forest is less susceptible to feature scaling and can describe non-linear hydrological linkages [16]. Ensemble flood-susceptibility experiments have demonstrated that tree-based techniques often perform better than single learners by lowering variation and considering diverse conditioning effects [17]. While XGBoost and boosting-based models can further improve the discrimination performance of some flood-prone mapping tasks, they may be more susceptible to noisy labels and require careful hyperparameter tweaking [18]. Recurrent Neural Networks and neural forecasting models are good at capturing long-term dependencies in

deep learning, but their training data and computational needs are very large for small watershed applications [19].

Studies on water-level forecasting have demonstrated that statistical machine learning may produce high-accuracy short-term forecasts when rainfall and upstream stage are appropriately connected [20]. Deep time-series analysis has also improved water-level forecasting for big rivers, although the model complexity is not always acceptable for low-cost sensor stations [21]. It is typically unclear whether a particular warning signal was triggered by a flood-prediction model, even though artificial neural networks are capable of learning its intricate form [22]. Given that the goal of the operational flood warning is to detect a flood scenario reliably, the recall rate for such events should be reasonably high; otherwise, a lack of warning might lead to catastrophic loss of life and damage. At the same time, there will be too many false alarms and a loss of trust in the warning system. Therefore, rather than using a single accuracy score, assessment indicators that take event-level reliability and class imbalance into consideration must be used.

Studies on rainfall-runoff models have also investigated hybrid models that integrate machine learning with conceptual hydrology [23]. The above models can increase the accuracy of modeling for delayed runoff response, but they are highly susceptible to input selections. The antecedent rainfall window, upstream-stage lag, soil-moisture memory, and discharge trend all have an impact on flood forecast outcomes. If all the possible lagged variables are included, the dimensionality will be rather big, and the classifier may overfit to local noise. In order to choose only hydrologically significant and non-redundant variables before to model training, an explicit feature selection step must be included to data mining for watershed flood prediction.

Random Forest and Feature Importance Analysis

Because Random Forest can manage non-linear interactions, heterogeneous numerical variables, missing-tolerant preprocessing, and a small training set, it is still appropriate for hydrological prediction. Random Forest can develop adaptable correlations between meteorological forcing and runoff response, making it appropriate for the scale of flood discharge modeling [24]. Interpretability of Random Forest is also affected by the stability of the feature set. The significance of each may be diminished by repetition when correlated rainfall windows, nearby water-level stations, and other soil-moisture statistics are included. As a result, we are unable to identify if a warning is brought on by excessive precipitation, an increase in the water level upstream, saturated soil, or an unusual acceleration of flow.

As a result, feature selection requires both of the criteria. A consistent feature selection strategy can decrease noise and weak predictors in the data before generating an ensemble of trees [25]. Using complimentary rather than redundant feature groups can improve classification resilience, as demonstrated by ensemble feature selection and transformed-subset learning [26]. According to Shapley-value studies of feature selection, a feature must provide incremental gains over other variables in order to be deemed useful [27]. The above approaches are applicable for watershed sensor networks since many of the variables share the same hydrological source but differ in lag, scale and dependability.

Explainable Artificial Intelligence (XAI) offers an all-weather strategy for understanding tree models. According to XAI research, the explanation of a model should contain domain knowledge and not only be a list of numerical rankings [28]. Predictor-effect visualization can highlight non-linear thresholds and interactions in black-box supervised learning [29]. Tree-specific explanation approaches further integrate local decisions with global feature contributions and are suited for Random Forest models [30]. Because the model now takes into account growing upstream water levels, greater rainfall accumulation, or saturated soil conditions, this explanation can assist assess whether the recall of a flood event has improved. In order to create a single flood-warning system, the current study integrates feature selection, Random Forest classification, and feature-level interpretation.

Feature-Oriented Random Forest Prediction Framework

Temporal Hydrological Feature Encoding

A series of multiple-source sensors in a watershed monitoring network serve as the input for the suggested model. At each time step, the system maintains rainfall intensity, cumulative rainfall, upstream water level, downstream water level, anticipated discharge, soil moisture, reservoir release signal, slope-weighted runoff

index and communication quality. Due to the delay and accumulation of flood response, each sensor stream is turned into a sliding time-window representation rather than being considered a separate observation. Twelve sampling intervals make up the window's length, which translates to three hours at 15-minute sampling intervals.

$$X_t = [r_t, h_t, q_t, m_t, u_t, z_t] \quad \text{Eq.(1)}$$

Here, r_t denotes rainfall descriptors, h_t denotes water-level descriptors, q_t denotes discharge descriptors, m_t denotes soil-moisture descriptors, u_t denotes upstream regulation descriptors, and z_t denotes sensor reliability descriptors. This compact notation groups physically related variables and avoids treating all raw channels as independent.

Only when there is a tiny gap and when the nearby observations are trustworthy are missing values filled in. To prevent creating a false hydrological trend through overly forceful replacement, a longer gap is maintained in the validity mask.

$$x_i'(t) = a_i(t)x_i(t) + (1 - a_i(t))x_i(t-1) \quad \text{Eq.(2)}$$

The reliability coefficient $a_i(t)$ is determined by sensor status, packet delay, and local variance. When a sensor becomes unstable, the model uses the previous reliable observation but retains the low reliability state as an additional descriptor.

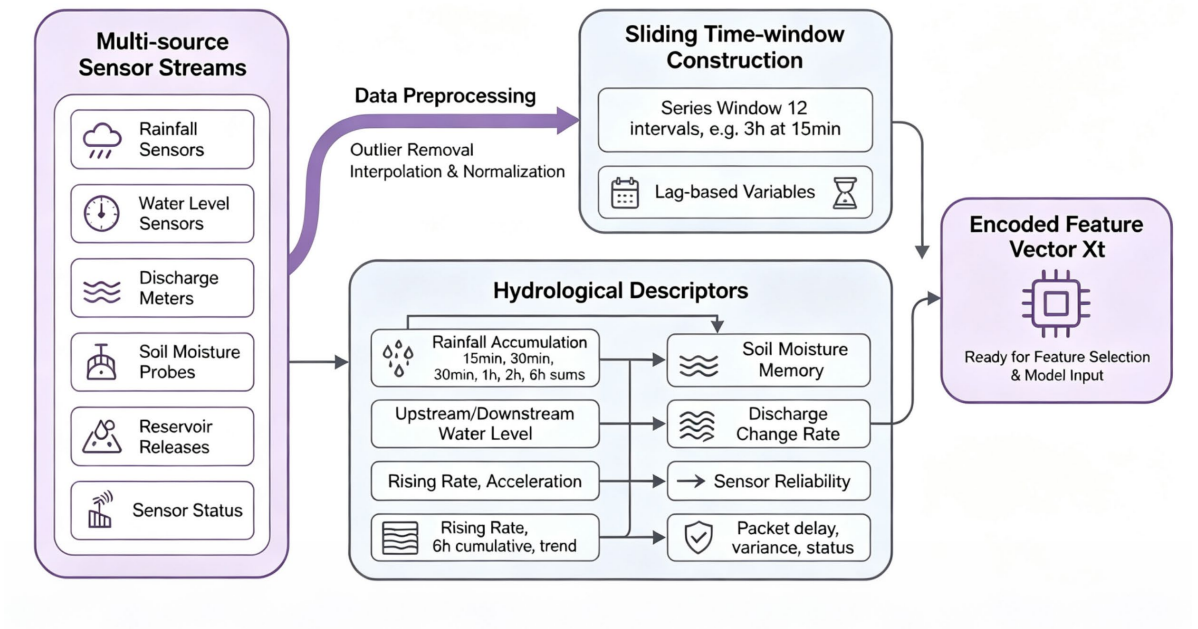

Figure 1. Temporal hydrological feature encoding for watershed sensor data

Rainfall accumulation is represented at multiple temporal horizons to describe both short burst rainfall and antecedent wetness. In the constructed dataset, 15 min, 30 min, 1 h, 2 h, and 6 h cumulative rainfall is used as candidate variables.

$$R_k(t) = \sum_{j=0}^{k-1} r(t-j) \quad \text{Eq.(3)}$$

The classifier can differentiate between slow basin-wide saturation and rapid flash-flood triggers with the aid of the chosen horizons. In a 6-hour timeframe, the average flood-event sample had gathered a total rainfall of 42.6 mm, whereas the non-flood sample had only 11.8 mm.

Normalized level, first-order rising rate, and acceleration are used to illustrate water-level dynamics. A reasonably high-rise speed can be used to offer an early warning of a flood wave before it reaches a critical height.

$$g_h(t) = \frac{h(t) - h(t-1)}{\Delta t} \quad \text{Eq.(4)}$$

A positive and persistent $g_h(t)$ indicates that the river stage is responding to upstream rainfall or reservoir release. In the training data, flood events show a mean peak rising rate of $0.37\ \mathrm{m/h}$, compared with $0.08\ \mathrm{m/h}$ under normal conditions. Figure 1 shows the temporal hydrological feature encoding process. It links raw sensor streams, reliability correction, sliding-window transformation, and candidate descriptor construction into a single input pipeline.

Joint Relevance-Redundancy Feature Selection

There are 164 hydrological descriptors in the first set of candidate pools. Included are the five levels of rainfall accumulation, basin static characteristics, soil-moisture memory, upstream-downstream lag, water-level slope and acceleration, discharge growth rate, and sensor reliability statistics. Increase the tree's complexity by feeding Random Forest all of the descriptions directly. Thus, prior to classifier training, the suggested selection procedure will ascertain feature relevance, redundancy, and event-level contribution.

$$\mathrm{Rel}(f_i) = \frac{I(f_i; y)}{H(y) + \varepsilon} \quad \text{Eq.(5)}$$

The relevance score $\mathrm{Rel}(f_i)$ is calculated by normalized mutual information between feature f_i and flood-risk label y. This score is insensitive to monotonic nonlinear dependence and can identify variables whose distribution differs clearly between normal, warning, and flood states.

$$\mathrm{Red}(f_i, F_s) = \max_{f_j \in F_s} \left|\rho(f_i, f_j)\right| \quad \text{Eq.(6)}$$

The redundancy score is the maximum absolute correlation between a candidate feature and the already selected subset F_s. The maximum form is deliberately strict, because rainfall variables from adjacent windows and water-level variables from neighboring stations often carry almost identical information.

According to the adaptive threshold, a feature is retained if it is highly relevant and not excessively redundant. As the size of the selected subset rises, the threshold is raised; consequently, numerous duplicate descriptors cannot be added late. After about 35 candidates were kept, the initial redundancy tolerance of 0.92 decreased to 0.68. When the subset is sufficiently informative, this schedule can avoid duplicate expansion by incorporating the strongly linked variables early on.

$$\mathrm{Imp}(f_i) = T^{-1} \sum_{t} \sum_{v \in V_t(i)} \Delta G_v \quad \text{Eq.(7)}$$

The term ΔG_v denotes the Gini reduction produced by feature f_i at node v. Features that split flood and non-flood samples near upper tree levels usually obtain higher importance.

The final utility score combines relevance, embedded importance, and redundancy penalty. It also includes a reliability term so that unstable sensors are not selected simply because they correlate with a specific historical event.

$$U(f_i) = 0.40\mathrm{Rel}(f_i) + 0.35\mathrm{Imp}(f_i) - 0.20\mathrm{Red}(f_i, F_s) + 0.05Q_i \quad \text{Eq.(8)}$$

Q_i is the average reliability of the sensor group associated with the feature. After selection, the feature dimension is reduced from 164 to 68, producing a 58.7% compression ratio while retaining the dominant hydrological information.

The feature subset is finalized by a marginal event-recall criterion. A feature is accepted only if it improves validation macro-F1 or increases recall of flood samples without raising false alarms sharply.

$$\Delta E_k = F1_m(F_s \cup f_k) - F1_m(F_s) \quad \text{Eq.(9)}$$

When ΔE_k remains below 0.0015 for six consecutive ranked candidates, the selection stops. This strategy avoids adding weak variables that only increase feature dimension. Class imbalance is handled through event-sensitive weighting because flood samples account for only 18.4% of the dataset. Warning and flood categories are therefore assigned higher influence during validation scoring.

Ensemble Flood-Risk Decision Mechanism

The selected feature vector is used to train a Random Forest classifier. Each tree receives a bootstrap sample and a random subset of candidate features at each split. This reduces variance and prevents a single dominant rainfall or water-level variable from controlling all decision paths. The forest contains 260 trees, the maximum depth is limited to 20, and the minimum leaf size is set to 5.

$$G(v) = 1 - \sum_{c} p_{v,c}^{2} \qquad \text{Eq.(10)}$$

$G(v)$ is the Gini impurity at node v. A split is preferred when it reduces impurity while keeping enough samples in child nodes for stable flood-risk estimation.

$$s_v = \arg\max_{f_i,\tau} [G(v) - G(v_L) - G(v_R)] \qquad \text{Eq.(11)}$$

The split rule s_v searches the selected feature f_i and threshold τ that maximize impurity reduction. In the final model, upstream water-level rising rate, 2 h cumulative rainfall, and soil-moisture memory appear most frequently among upper-level splits.

$$P(c \mid x) = T^{-1} \sum_{t} P_t(c \mid x) \qquad \text{Eq.(12)}$$

The predicted class is the maximum probability category among normal, alert, and flood states. In the test set, floodstate probability above 0.72 corresponds to 94.8% event recall and 6.3% false-alarm rate. Figure 2 presents the Random Forest decision mechanism with selected hydrological variables.

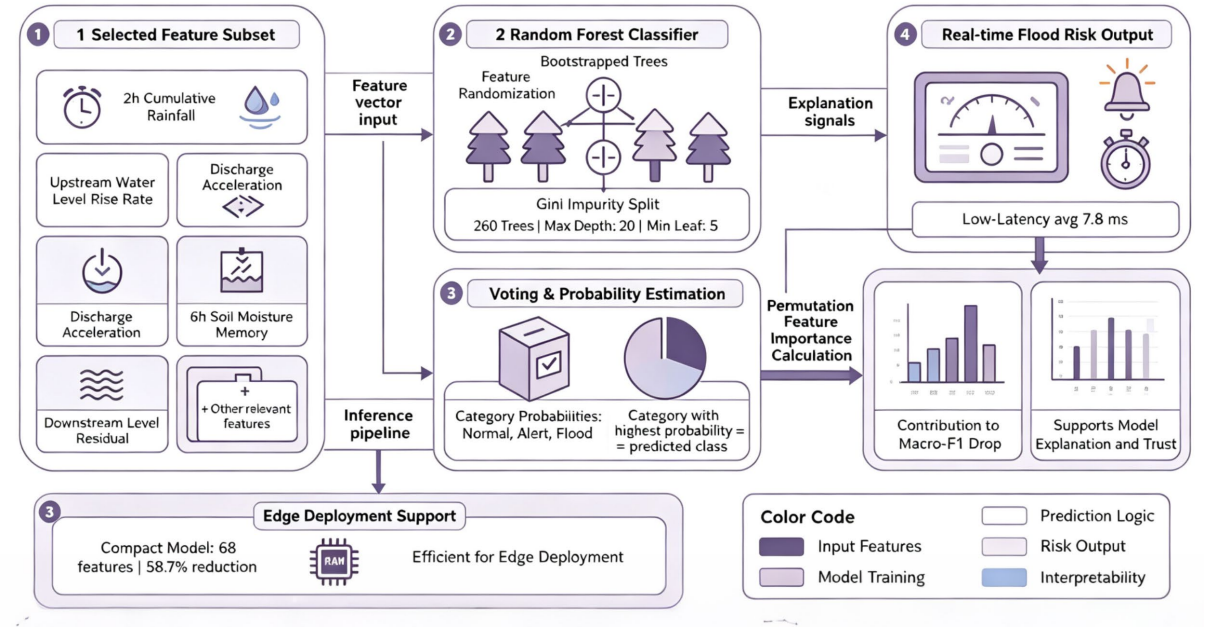

Figure 2. Random Forest flood-risk decision mechanism with selected features

To support interpretation, the model calculates a permutation contribution score for each selected feature. This score measures how much macro-F1 decreases when a feature is randomly permuted within the validation set.

$$\Phi_i = F1_m(x) - F1_m(\pi_i(x)) \qquad \text{Eq.(13)}$$

A larger Φ_i indicates that the feature is essential for flood-risk discrimination. The top five features are 2 h cumulative rainfall, upstream water-level rising rate, discharge acceleration, 6 h soil-moisture memory, and downstream level residual. A compactness-aware model score is used for hyperparameter selection, and the selected model achieves a composite score of 0.914, compared with 0.872 for conventional Random Forest without feature selection.

A compactness-aware model score is used for hyperparameter selection. It jointly considers macro-F1, flood recall, selected feature ratio, and inference latency.

$$J_c = 0.50F_m + 0.30R_f - 0.10D_f - 0.10L_i \quad \text{Eq.(14)}$$

Here， F_m denotes macro－F1， R_f denotes flood event recall， D_f denotes the normalized selected feature radioland L_i denotes normalized inference latency with this criterion the selected model achieves a composite score of 0.914， compared with 0.872 for conventional Random Forest without feature selection.

Experimental Design and Metric Definition

Sensor Dataset and Preprocessing Configuration

The experiment uses a watershed sensor dataset from 34 monitoring sites in a medium-sized river basin. The records are as follows: rainfall, water level, discharge, soil moisture, reservoir release and sensor-status logs from 2019 to 2022. All of the samples are sliding windows, and the sampling duration is fifteen minutes. Samples that overlap but are chronologically ordered are produced because the window length is specified at 12 intervals and the step is fixed at 2 intervals. A total of 46,280 labelled windows have been collected, of which 28,940 are normal samples, 8,825 are alert samples and 8,515 are flood samples. Train, validation and test sets are temporally split at a ratio of 60:20:20 to prevent leaking of adjacent event periods.

Outlier elimination, range verification, reliability-weighted interpolation, and station-wise normalization are examples of sensor preprocessing. If a spike's local derivative is greater than the 99.5th percentile and there is no continuous change in the nearby stations, the spike is eliminated. Only the training-set statistics are used to normalize all variables, and the same parameters are used to transform validation and test samples.

Evaluation Metrics for Flood Prediction

The accuracy of classification and the dependability of operational alerts are the two. Overall accuracy is the ratio of correctly categorized samples, although it is not ideal because normal samples are more common in the dataset. Therefore, the following are simultaneously reported: precision, recall, macro-F1, false-alarm rate, feature reduction ratio, model size and average inference latency. To guarantee that fewer flood alerts are overlooked during classification errors, flood-event recall is utilized. To ascertain whether only the flood class has improved or if all categories have performed well, recall is computed for that class and compared with macro-F1. Additionally, the False-Alarm Rate is comparatively high; otherwise, there would be an excessive number of alarms, which would cause consumers to lose trust in the system. Determine the reduction ratio of features by comparing the number of features in the original set with the selected set.

The revised model has a FRR of 58.7% and, on an Intel i7 Edge workstation, has lowered the average inference time per sample to 7.8ms. The basin-level use of several windows that require ongoing evaluation necessitates the higher efficiency.

Baseline Methods and Training Protocol

Support Vector Machine, conventional Random Forest, gradient boosting decision trees, XGBoost, and LSTM are compared to the suggested model. Gradient boosting and XGBoost employ tunable tree depth and learning rate, while the SVM baseline is a radial basis kernel. Temporal dependencies are modeled using the two recurrent layers and a hidden size of 64 in the LSTM baseline. Conventional Random Forest uses the same number of trees as the proposed model but does not select any of the 164 candidate features. All baselines employ the same chronological split and an identical preprocessing pipeline.

To choose hyperparameters, use the flood-event recall and macro-F1 on the validation set. To evaluate the robustness, three disturbance situations were added: transmission delays of one to four sampling intervals, Gaussian noise intensities of 2% to 10%, and missing sensor ratios ranging from 5% to 30%. Using various random seeds, five disturbance settings have been repeated. As a result, the evaluation design will additionally evaluate the clean-sample's operational performance under faulty watershed sensor networks.

Comparative Results and Hydrological Interpretation

Prediction Accuracy and Error Distribution

The predictions of the models—SVM, gradient boosting, Random Forest, XGBoost, and LSTM—are displayed in Figure 3 along with a comparison with the suggested model. The macro-F1 and overall accuracy of the suggested approach are 95.4% and 96.1%, respectively, as shown in Figure 3(a). Tree ensembles are more advantageous than just adding more features, as Random Forest has attained an accuracy of 93.2% and a macro-F1 score of 91.8%. Flood-event recall is shown in Figure 3(b). The proposed model has a recall of 94.8%, and the others are 86.1% for SVM, 90.4% for gradient boosting, and 91.6% for LSTM. First, compared to the downstream region, a greater rate of water level rise and antecedent rainfall conditions are retained earlier [31]. The distribution of errors in the normal, alert, and flood states is displayed in Figure 3(c). Most of the residual mistakes occur in the area between the alert and flood categories, and the hydrological change is not sudden. The false-alarm rate is 6.3%, which is lower than the 9.8% of classic Random Forest since duplicate rainfall windows do not enhance the weight of recurrent warning evidence.

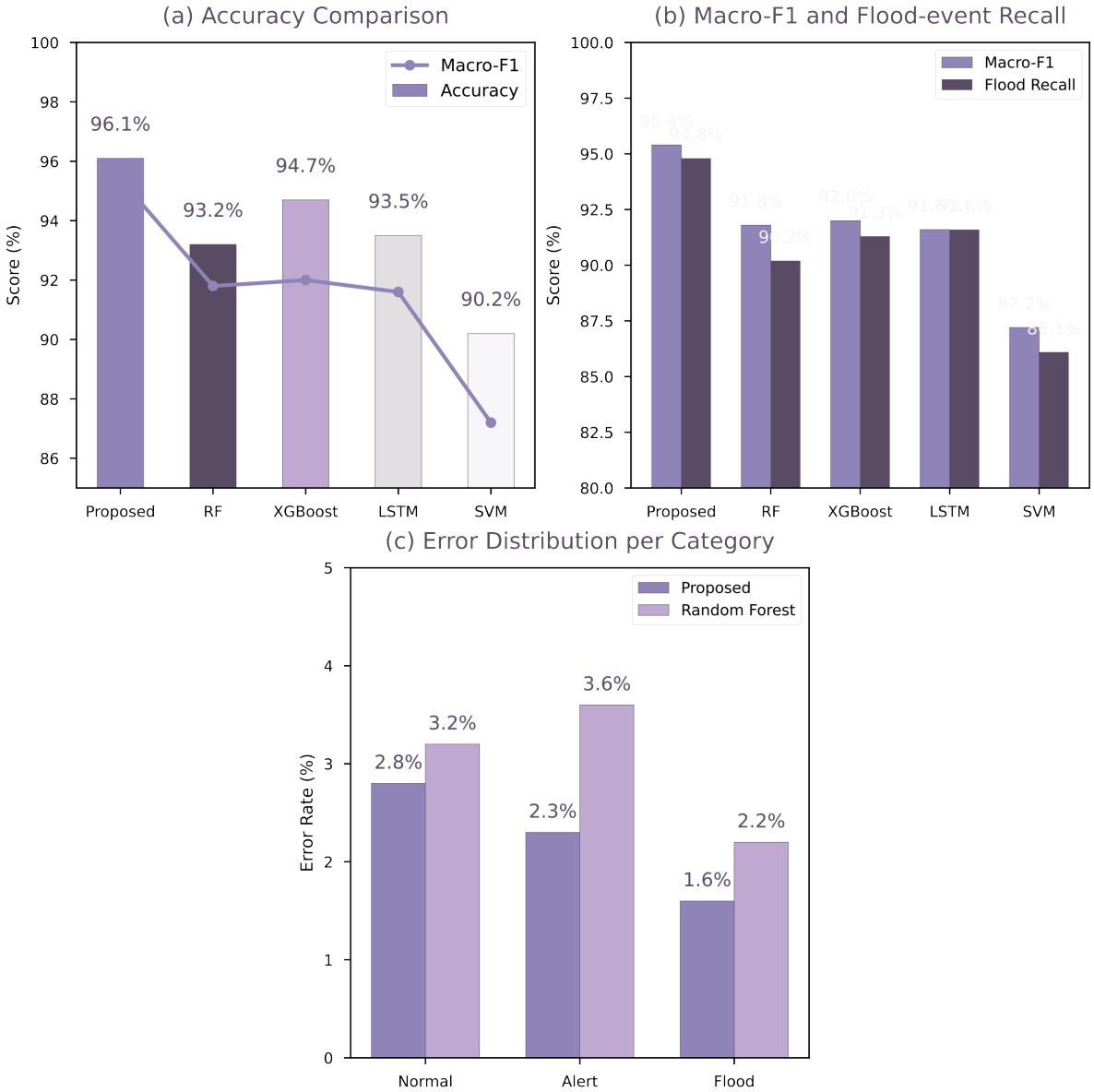

Figure 3. Prediction accuracy and error behavior. (a) Accuracy comparison. (b) Macro-F1 and flood-event recall comparison. (c) Error distribution among risk categories.

An improved model capacity is not always the cause of the higher accuracy. XGBoost obtains a high accuracy of 94.7% but has a comparatively high false-alarm rate of 8.5% because it fits weak warning signals in high-dimensional features too aggressively. LSTM can be utilized for time-series data, although it is more sensitive to missing water-level readings; consequently, with a missing rate of 20%, its macro-F1 reduces by 3.9 percentage points. The proposed model saves the most event-sensitive variables during training, and hence the tree ensemble receives cleaner hydrological evidence. Therefore, its increase in flood recall is considerably bigger than that in normal-state accuracy. Because the model has decreased the category with a high cost of misclassification, it is also advantageous operationally.

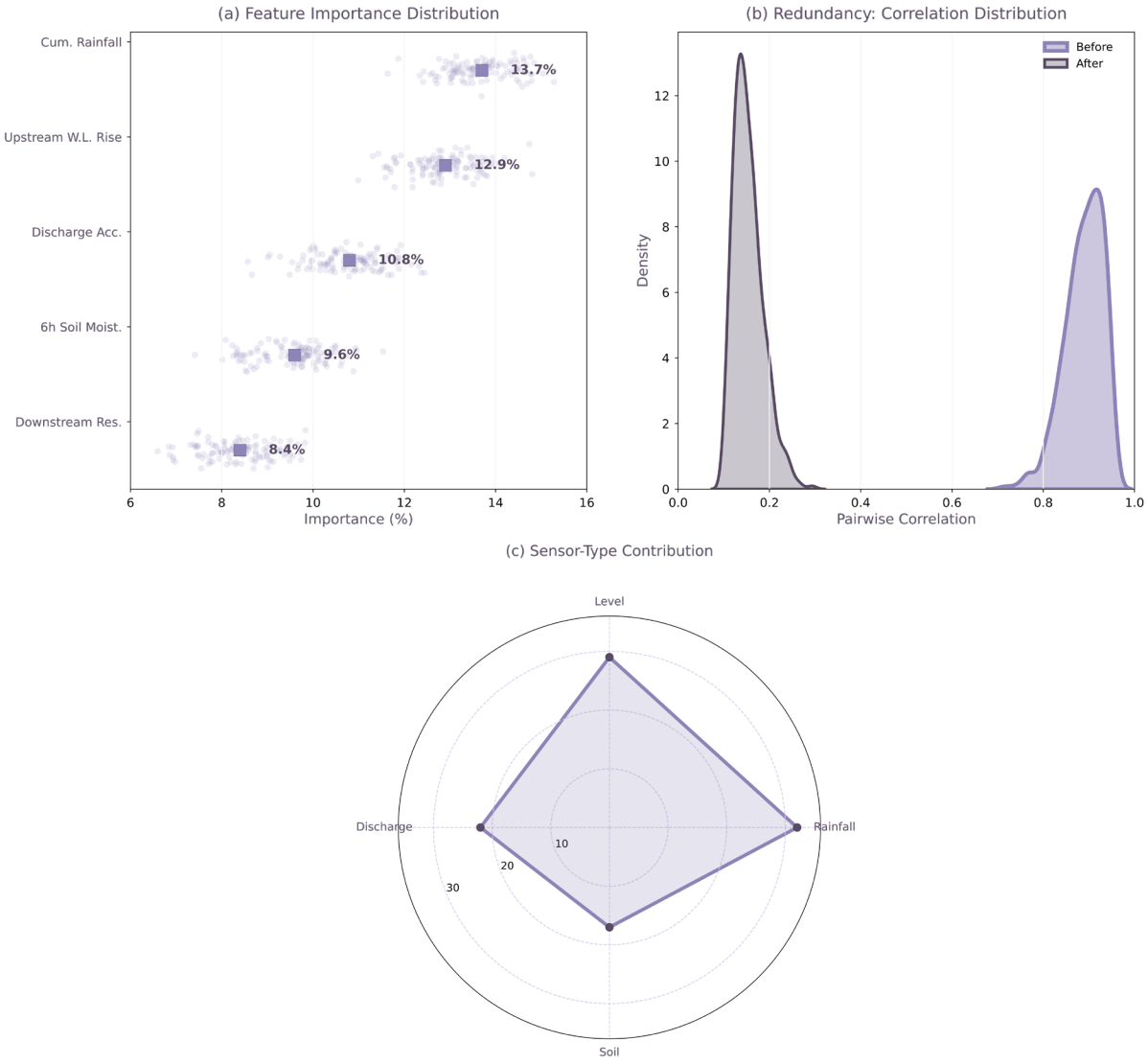

Figure 4. Hydrological feature importance visualization. (a) Selected feature ranking. (b) Redundancy reduction effect. (c) Sensor-type contribution comparison.

Feature Contribution and Event-Level Interpretation

The significance of hydrological features following selection is shown in Figure 4. The top two variables are cumulative rainfall (13.7%) and the rate of increase in upstream water level (12.9%), which are followed by discharge acceleration (10.8%), 6-hour soil-moisture memory (9.6%), and downstream level residual (8.4%), as shown in Figure 4(a). The chosen subset is hydrologically significant rather than merely statistical because the variables include rainfall forcing, channel response, antecedent saturation, and local anomaly signals [32]. The decrease in redundancy is shown in Figure 4(b). The mean of the pairwise correlations for the retained rainfall descriptors is 0.81 and 0.46, respectively; the mean correlation among level-related descriptors is 0.78 and 0.43. The sensor-type contribution is shown in Figure 4(c). Water-level and discharge factors verify a flood stage,

whereas rainfall variables serve as the first warning. In order to improve the recall of a delayed basin reaction during an extended wet period, soil-moisture memory is required.

The confusion of flood-risk statuses is depicted in Figure 5. The majority of false alerts happen when cumulative rainfall surpasses 28 mm but the upstream water-level response is modest, as seen in Figure 5(a), which compares the normal and alert data. Figure 5(b) shows the comparison of alert and flood samples, and it is discovered that missed flood cases often have moderate rainfall but a rapid rise in the upstream area; consequently, channel routing features are necessary for a short-lead-time warning [33]. Figure 5(c) depicts the event-level confusion at varied rainfall intensities. In the event of severe rainfall over 50 mm in six hours, the proposed model has a flood recall rate of 96.7%; Random Forest is only at 92.1%. Under light-to-moderate rainfall, the model still obtains a recall of 92.8% thanks to the soil-moisture memory and discharge acceleration mechanisms for rainfall uncertainty.

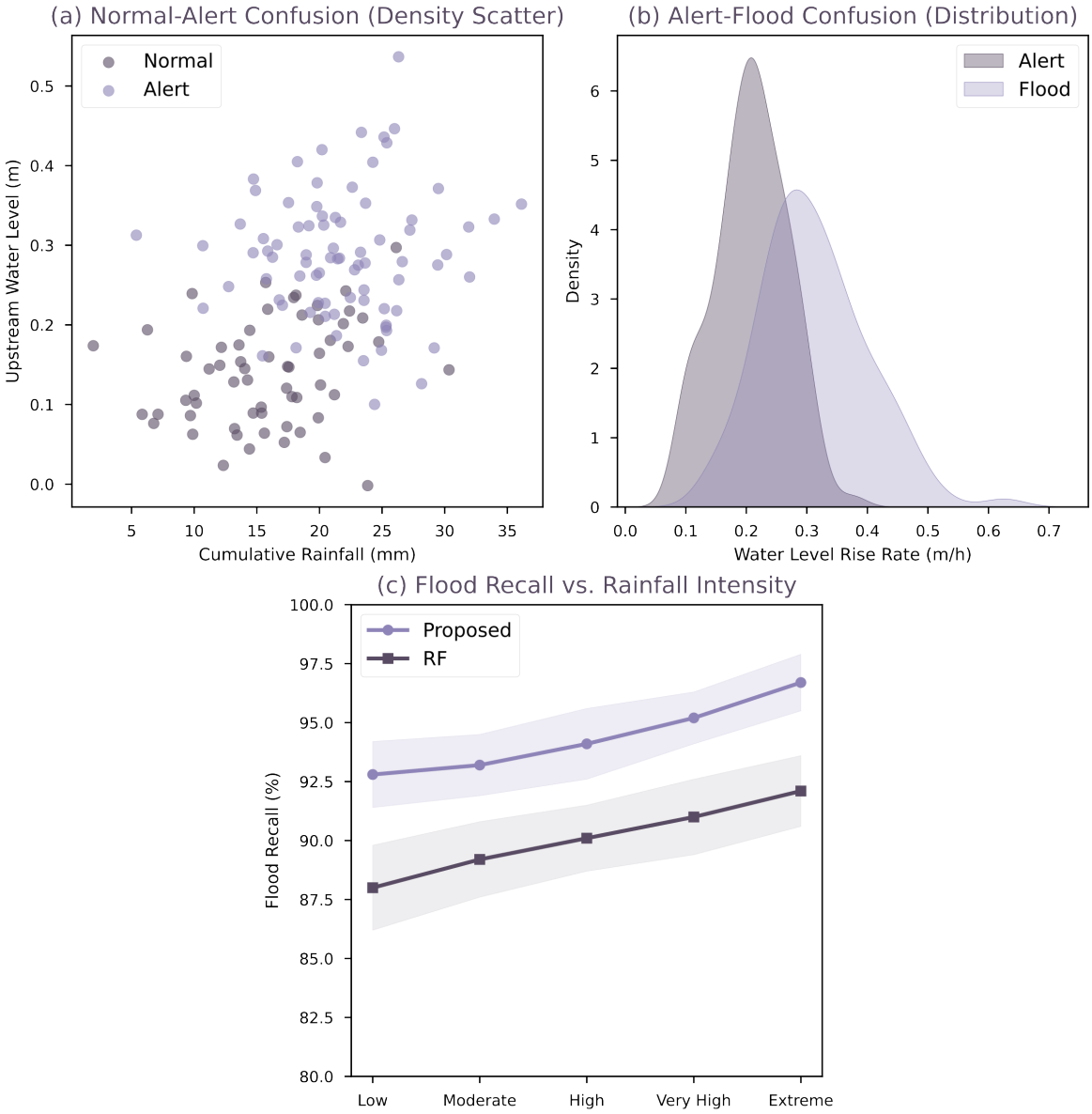

Figure 5. Flood-event classification confusion analysis. (a) Normal-alert confusion. (b) Alert-flood confusion. (c) Event-level confusion under rainfall intensity.

The selected feature set will perform better, according to the analysis above. The model will use several highly correlated rainfall windows in the absence of feature selection, increasing the likelihood that an alarm may be issued for a brief, non-hydrologically effective rainfall burst. Redundancy suppression has been carried out, and

now the model uses more complementary evidence: rainfall initiation, upstream-stage response, discharge acceleration, and soil saturation. Together, they reduce false alarms for isolated rainfall and enhance the detection rate of flood response owing to upstream accumulation rather than local rainfall [34].

Robustness and Deployment Analysis

Figure 6 illustrates the robustness under sensor disturbance. Figure 6(a) exhibits sensor dropout at 5% and 30%. At 20% missing observations, the suggested model retains a 93.6% macro-F1 score; otherwise, Random Forest is at 89.7% and LSTM is at 87.9%. The reason for this is that the chosen characteristics don't rely on a lot of weak informative sensors and have reliability indicators. The Gaussian noise is shown in Figure 6(b). The macro-F1 of the suggested model is 2.8 percentage points lower than that of XGBoost (5.1 percentage points) at a noise intensity of 10%. The transmission delay is shown in Figure 6(c). When the data are delayed by three periods, the flood recall is still 91.5% because antecedent rainfall and an upstream-stage trend have supplied predictive evidence for the downstream peak arrival [35].

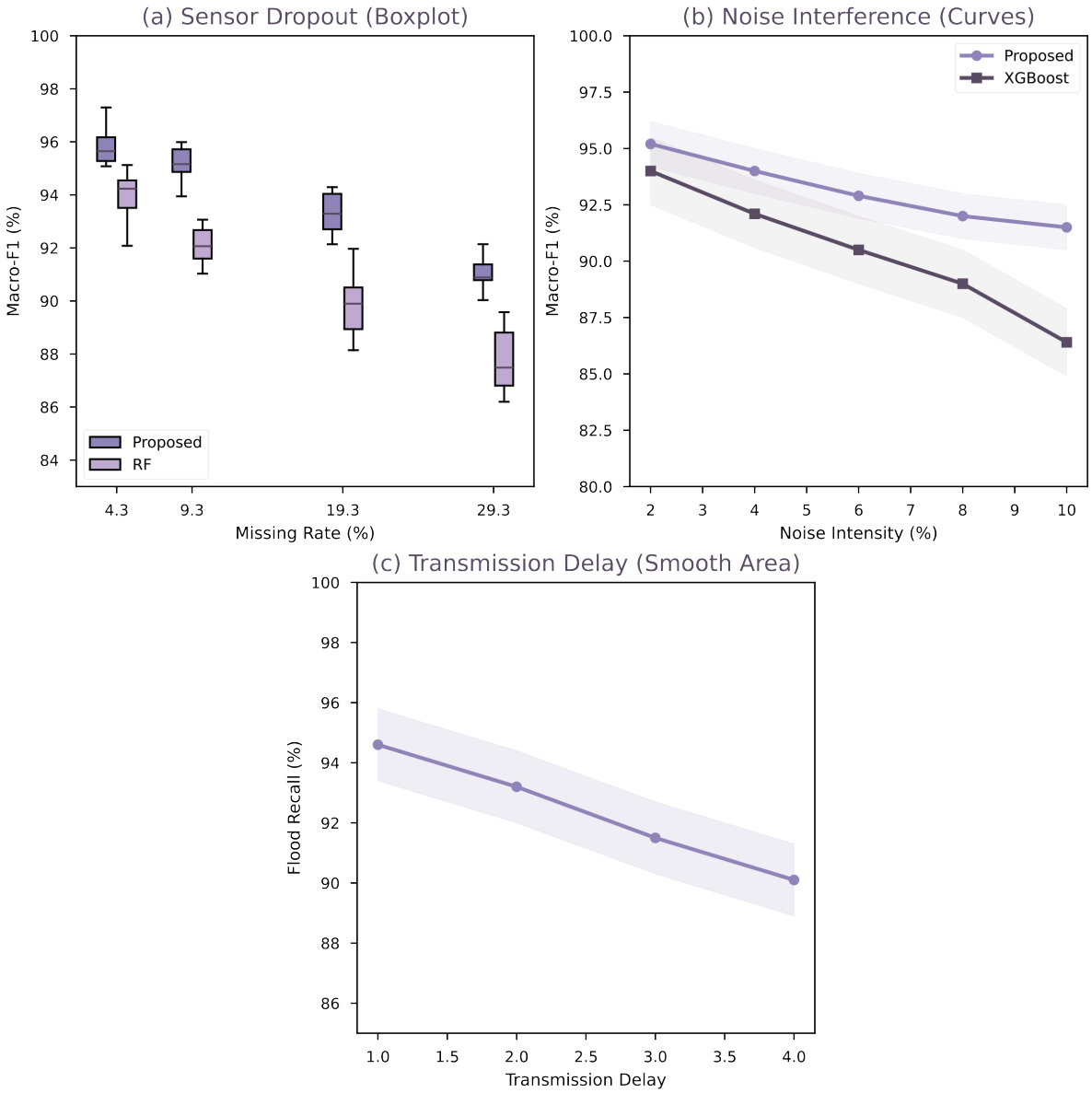

Figure 6. Model robustness under sensor disturbance. (a) Sensor dropout. (b) Noise interference. (c) Data transmission delay.

The deployment-oriented efficiency is shown in Figure 7. The suggested model's inference time is 7.8 ms per sample, compared to 13.6 ms for the traditional Random Forest and 21.4 ms for the LSTM, as shown in Figure

7(a). Feature dimension reduction is shown in Figure 7(b), and 68 of the 164 candidate descriptors were retained by the model. The selected-feature model uses 8.9MB of RAM, which is less than the whole Random Forest's 15.2MB, as seen in Figure 7(c). The benefits are appropriate for edge stations with limited processing power that must process several upstream windows and provide warnings [36].

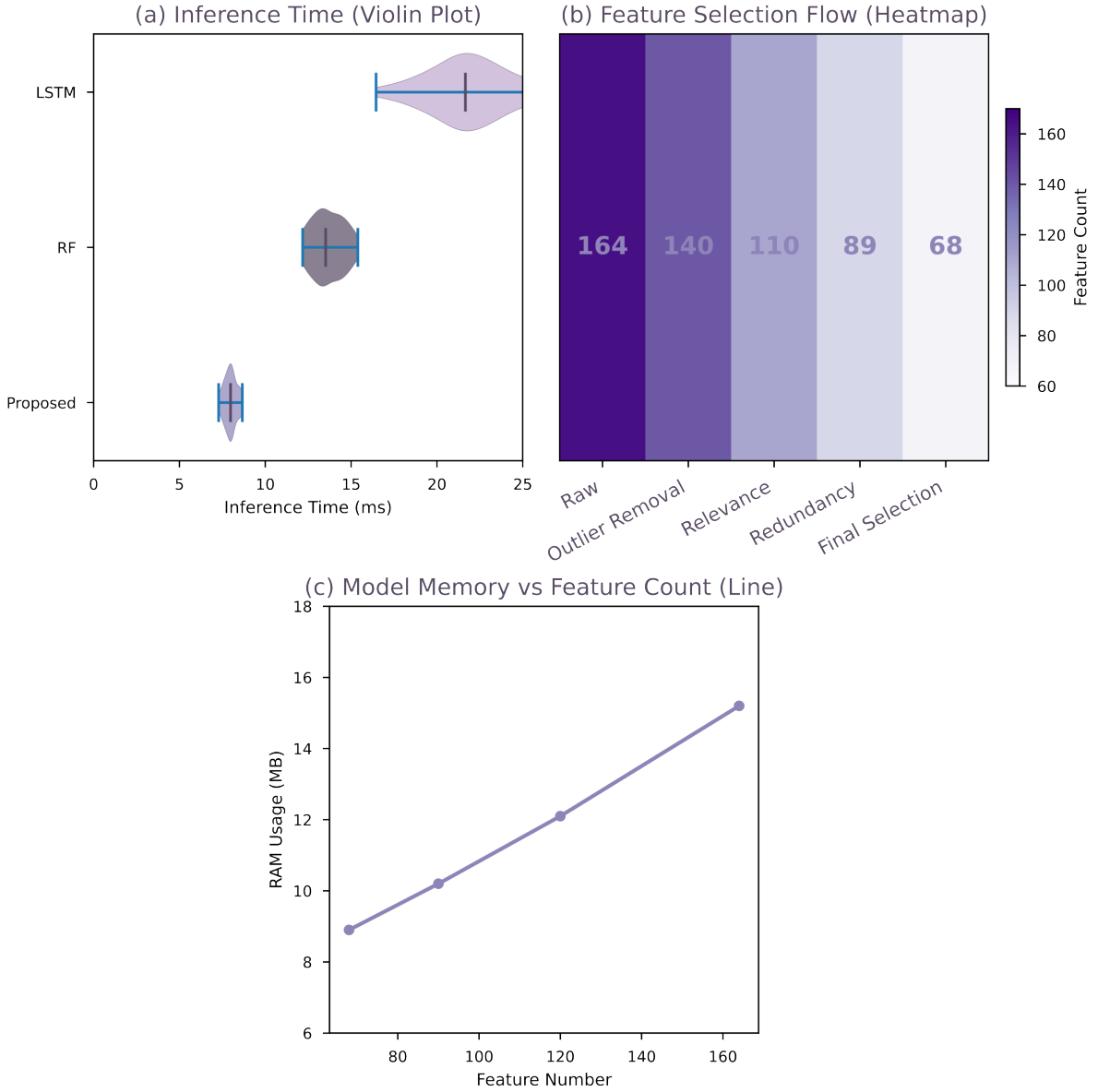

Figure 7. Deployment-oriented efficiency comparison. (a) Inference time. (b) Feature dimension reduction. (c) Memory usage.

The ablation findings demonstrate the necessity of every component. The macro-F1 value is 95.4% without lowering redundancy suppression, which raises the false-alarm rate to 8.1%. The flood recall with missing data decreases from 94.8% to 92.2% when the reliability weight is removed. While 76 features are retained when using simply mutual information, the entire relevance-redundancy-importance technique performs better. According to the findings, the model is not enhanced by merely decreasing dimensions; rather, the variables that are kept are stable, complimentary, and connected to the physical process of flood creation [37]. Analysis of reasons can also be applied in operational inspection to establish if the source of the alarm is excess rainfall collection, a quick increase in upstream water level, or soil saturation. The above transparent evidence fits the present requirement for explainable tree models in risk-sensitive prediction [38]. Predictor-effect inspection shows a non-linear threshold of roughly 31 mm for 2-h cumulative rainfall and 0.29 m/h for upstream rise rate; consequently, it can discern between warning and flood states more effectively than a linear baseline [39]. Tree-

based local explanations also reveal that a high flood probability is not driven by a single element, but rather by the combination of rainfall burst, rising upstream stage and saturated antecedent circumstances [40].

Conclusion

In this paper, a Random Forest feature selection model based on watershed sensor networks was developed for flood prediction. The way multi-source rainfall, water level, outflow, soil moisture, upstream regulation and sensor reliability observations are translated into temporal hydrological descriptors, and then how compact variables are picked by relevance, redundancy, embedded importance and event recall criteria. Based on the preceding results, a relatively modest feature set can be selected to reduce the dependence on many sensors while keeping the non-linear discriminatory power of Random Forest.

In both clear and cluttered sensing settings, the model is capable of accurately classifying floods at the risk level. Compared to the baseline Random Forest and other baselines, it has improved the macro-F1 score, flood-event recall, false-alarm control, and inference speed. Together, the chosen characteristics—cumulative rainfall, upstream rise rate, discharge acceleration, soil-moisture memory, and downstream residual—can explain how short-term flood risk is formed. They are also hydrologically interpretable. It can assist in reaching the objective of high predictability and interpretability for the practical advantages of watershed warning systems.

In the future, one of the extensions to this framework will be to integrate online feature updates, radar rainfall prediction, remote-sensing precipitation products, and physically limited rainfall-runoff knowledge. In the event of real communication failures, sensor deterioration, and significant network congestion, the model can be utilized at the edge to test for a long time. Automatic alarms can be directly linked to basin management and evacuation action by combining Random Forest explanations with decision-level emergency rules.

Author Contributions

Adéla Svoboda contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. Jakub Dvořák contributes to conceptualization, methodology, software, validation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Schoppa, L., Disse, M., & Bachmair, S. (2020). Evaluating the performance of random forest for large-scale flood discharge simulation. Journal of Hydrology, 590, 125531. https://doi.org/10.1016/j.jhydrol.2020.125531

[2] Nguyen, D. T., & Chen, S. T. (2020). Real-time probabilistic flood forecasting using multiple machine learning methods. Water, 12(3), 787. https://doi.org/10.3390/w12030787

[3] Noymanee, J., & Theeramunkong, T. (2019). Flood forecasting with machine learning technique on hydrological modeling. Procedia Computer Science, 156, 377-386. https://doi.org/10.1016/j.procs.2019.08.214

[4] Ranit, A. B., & Durge, P. V. (2019). Flood forecasting by using machine learning. 2019 International Conference on Communication and Electronics Systems, 166-169. https://doi.org/10.1109/icces45898.2019.9002579

[5] Shamsi, S. (2019). Flood forecasting review using wireless sensor network. Global Sci-Tech, 11(1). https://doi.org/10.5958/2455-7110.2019.00003.x

[6] Suresh, S. (2020). Flood prediction using IoT enabled sensor network and machine learning. International Journal of Computing, Communications and Networking, 9(2), 25-30. https://doi.org/10.30534/ijccn/2020/06922019

[7] Rana, M. S., & Mahanta, C. (2022). Spatial prediction of flash flood susceptible areas using novel ensemble of bivariate statistics and machine learning techniques for ungauged region. Natural Hazards, 115(1), 947-969. https://doi.org/10.1007/s11069-022-05580-9

[8] Janizadeh, S., Avand, M., Jaafari, A., Phong, T. V., Bayat, M., Ahmadisharaf, E., et al. (2019). Prediction success of machine learning methods for flash flood susceptibility mapping in the Tafresh Watershed, Iran. Sustainability, 11(19), 5426. https://doi.org/10.3390/su11195426

[9] Elkhrachy, I. (2022). Flash flood water depth estimation using SAR images, digital elevation models, and machine learning algorithms. Remote Sensing, 14(3), 440. https://doi.org/10.3390/rs14030440

[10] Alipour, A., Ahmadalipour, A., Abbaszadeh, P., & Moradkhani, H. (2020). Leveraging machine learning for predicting flash flood damage in the Southeast US. Environmental Research Letters, 15(2), 024011. https://doi.org/10.1088/1748-9326/ab6edd

[11] El-Magd, S. A. A., Pradhan, B., & Alamri, A. (2021). Machine learning algorithm for flash flood prediction mapping in Wadi El-Laqeita and surroundings, Central Eastern Desert, Egypt. Arabian Journal of Geosciences, 14(4), 323. https://doi.org/10.1007/s12517-021-06466-z

[12] Kinage, C., Kalgutkar, A., Parab, A., Mandora, S., & Sahu, S. (2019). Performance evaluation of different machine learning based algorithms for flood prediction and model for real time flood prediction. 2019 5th International Conference on Computing, Communication, Control and Automation, 1-7. https://doi.org/10.1109/iccubea47591.2019.9128379

[13] Prasad, P., Loveson, V. J., Das, B., & Kotha, M. (2021). Novel ensemble machine learning models in flood susceptibility mapping. Geocarto International, 37(16), 4571-4593. https://doi.org/10.1080/10106049.2021.1892209

[14] Farhadi, H., & Najafzadeh, M. (2021). Flood risk mapping by remote sensing data and random forest technique. Water, 13(21), 3115. https://doi.org/10.3390/w13213115

[15] McGrath, H., & Gohl, P. N. (2022). Accessing the impact of meteorological variables on machine learning flood susceptibility mapping. Remote Sensing, 14(7), 1656. https://doi.org/10.3390/rs14071656

[16] Vafakhah, M., Mohammad Hasani Loor, S., Pourghasemi, H., & Katebikord, A. (2020). Comparing performance of random forest and adaptive neuro-fuzzy inference system data mining models for flood susceptibility mapping. Arabian Journal of Geosciences, 13(11), 417. https://doi.org/10.1007/s12517-020-05363-1

[17] Ghosh, S., Saha, S., & Bera, B. (2022). Flood susceptibility zonation using advanced ensemble machine learning models within Himalayan foreland basin. Natural Hazards Research, 2(4), 363-374. https://doi.org/10.1016/j.nhres.2022.06.003

[18] Aydin, H. E., & Iban, M. C. (2022). Predicting and analyzing flood susceptibility using boosting-based ensemble machine learning algorithms with SHapley Additive exPlanations. Natural Hazards, 116(3), 2957-2991. https://doi.org/10.1007/s11069-022-05793-y

[19] Gudiyangada Nachappa, T., & Meena, S. R. (2020). A novel per pixel and object-based ensemble approach for flood susceptibility mapping. Geomatics, Natural Hazards and Risk, 11(1), 2147-2175. https://doi.org/10.1080/19475705.2020.1833990

[20] Abedi, R., Costache, R., Shafizadeh-Moghadam, H., & Pham, Q. B. (2021). Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. Geocarto International, 37(19), 5479-5496. https://doi.org/10.1080/10106049.2021.1920636

[21] Esfandiari, M., Abdi, G., Jabari, S., McGrath, H., & Coleman, D. (2020). Flood hazard risk mapping using a pseudo supervised random forest. Remote Sensing, 12(19), 3206. https://doi.org/10.3390/rs12193206

[22] Motta, M., de Castro Neto, M., & Sarmento, P. (2021). A mixed approach for urban flood prediction using machine learning and GIS. International Journal of Disaster Risk Reduction, 56, 102154. https://doi.org/10.1016/j.ijdrr.2021.102154

[23] Ekwueme, B. N. (2022). Machine learning based prediction of urban flood susceptibility from selected rivers in a tropical catchment area. Civil Engineering Journal, 8(9), 1857-1878. https://doi.org/10.28991/cej-2022-08-09-08

[24] N, M., & A U, A. (2021). Cloud-based flood prediction using IoT devices and machine learning algorithms. 2021 Second International Conference on Electronics and Sustainable Communication Systems, 754-762. https://doi.org/10.1109/icesc51422.2021.9532916

[25] Yan, X., Xu, K., Feng, W., & Chen, J. (2021). A rapid prediction model of urban flood inundation in a high-risk area coupling machine learning and numerical simulation approaches. International Journal of Disaster Risk Science, 12(6), 903-918. https://doi.org/10.1007/s13753-021-00384-0

[26] Norallahi, M., & Seyed Kaboli, H. (2021). Urban flood hazard mapping using machine learning models: GARP, RF, MaxEnt and NB. Natural Hazards, 106(1), 119-137. https://doi.org/10.1007/s11069-020-04453-3

[27] Dazzi, S., Vacondio, R., & Mignosa, P. (2021). Flood stage forecasting using machine-learning methods: A case study on the Parma River (Italy). Water, 13(12), 1612. https://doi.org/10.3390/w13121612

[28] Faruq, A., Hussein, S. F. M., Marto, A., & Abdullah, S. S. (2022). Flood river water level forecasting using ensemble machine learning for early warning systems. IOP Conference Series: Earth and Environmental Science, 1091(1), 012041. https://doi.org/10.1088/1755-1315/1091/1/012041

[29] Phan, T. T. H., & Nguyen, X. H. (2020). Combining statistical machine learning models with ARIMA for water level forecasting: The case of the red river. Advances in Water Resources, 142, 103656. https://doi.org/10.1016/j.advwatres.2020.103656

[30] Wang, Q., & Wang, S. (2020). Machine learning-based water level prediction in Lake Erie. Water, 12(10), 2654. https://doi.org/10.3390/w12102654

[31] Atashi, V., Gorji, H. T., Shahabi, S. M., Kardan, R., & Lim, Y. H. (2022). Water level forecasting using deep learning time-series analysis: A case study of Red River of the North. Water, 14(12), 1971. https://doi.org/10.3390/w14121971

[32] Wang, G., Yang, J., Hu, Y., Li, J., & Yin, Z. (2022). Application of a novel artificial neural network model in flood forecasting. Environmental Monitoring and Assessment, 194(2), 125. https://doi.org/10.1007/s10661-022-09752-9

[33] Okkan, U., Ersoy, Z. B., Ali Kumanlioglu, A., & Fistikoglu, O. (2021). Embedding machine learning techniques into a conceptual model to improve monthly runoff simulation: A nested hybrid rainfall-runoff modeling. Journal of Hydrology, 598, 126433. https://doi.org/10.1016/j.jhydrol.2021.126433

[34] Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing, 86, 105936. https://doi.org/10.1016/j.asoc.2019.105936

[35] Khoder, A., & Dornaika, F. (2021). Ensemble learning via feature selection and multiple transformed subsets: Application to image classification. Applied Soft Computing, 113, 108006. https://doi.org/10.1016/j.asoc.2021.108006

[36] Xuan, Y., Si, W., Zhu, J., Sun, Z., Zhao, J., Xu, M., et al. (2021). Multi-model fusion short-term load forecasting based on random forest feature selection and hybrid neural network. IEEE Access, 9, 69002-69009. https://doi.org/10.1109/access.2021.3051337

[37] Fryer, D., Strumke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. IEEE Access, 9, 144352-144360. https://doi.org/10.1109/access.2021.3119110

[38] Barredo Arrieta, A., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[39] Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(4), 1059-1086. https://doi.org/10.1111/rssb.12377

[40] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9