IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Multimodal Sentinel-1/2 Data Fusion with a Dual-Path ViT-Kolmogorov-Arnold Network Pipeline for AI-driven Flood Event Classification

JAYAKRISHNAN ANANDAKRISHNAN[1], ALKHA MOHAN[2], VIPIN VENUGOPAL[3], WEIWEI JIANG[4], MOHD AMIRUDDIN ABD RAHMAN[5], and JAWAD KHAN[6]

[1]Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore, Amrita Vishwa Vidyapeetham, India (e-mail: a_jayakrishnan@cb.amrita.edu)
[2]Department of Computer Science and Engineering, Indian Institute of Information Technology (IIIT) Kottayam, Kerala, India (e-mail: alkha@iiitkottayam.ac.in)
[3]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India (e-mail: v_vipin@cb.amrita.edu)
[4]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: jww@bupt.edu.cn)
[5]Department of Physics, Faculty of Science, Universiti Putra Malaysia (UPM), Malaysia (e-mail: mohdamir@upm.edu.my)
[6]School of Computing, Gachon University, Seongnam-si 13120, Republic of Korea (e-mail: jkhanbk1@gachon.ac.kr)

Corresponding authors: Alkha Mohan, Vipin Venugopal, and Weiwei Jiang (e-mails: mohan.alkha@gmail.com; v_vipin@cb.amrita.edu; jww@bupt.edu.cn).

Alkha Mohan and Vipin Venugopal are co-first authors.

ABSTRACT Timely identification of flood events is a challenging problem, as they are devastating natural disasters and cause severe damage to lives and livelihoods. Satellite-based flood identification is widely adopted because it offers extended spatio-temporal coverage compared to traditional ground and Unmanned Aerial Vehicle (UAV)-based methods. However, satellite-based flood identification is heavily hindered by adverse weather and cloud conditions. Utilizing Synthetic Aperture Radar (SAR) data helps address sensing challenges under adverse weather conditions; however, SAR provides limited spectral information. Therefore, fusing SAR with optical satellite data becomes an ideal choice, although multimodal learning with SAR–optical fusion remains a challenging task. This research proposes FloodSense, a dual-path ViT–KAN pipeline for AI-driven flood event classification via Sentinel-1/2 multimodal fusion, introducing a dual-path learning strategy that jointly captures global and local characteristics and enhances nonlinear mapping. The FloodSense pipeline comprises a Vision Transformer Flood Feature Extractor (ViT-FloodX) and a Spatial Spectral Feature Extractor (3D2D-Extract), jointly capturing cross-domain global contextual semantics and fine-grained spatial-spectral patterns. Moreover, a Kolmogorov–Arnold Network Feature Refinement (Refine-KAN) module enhances nonlinear feature representation by combining KANs with Multi-Layer Perceptron (MLP) layers, improving multimodal feature learning. The empirical evaluations of FloodSense are conducted on the popular SEN12-FLOOD multimodal flood event classification dataset and demonstrate competitive performance against the state-of-the-art (SOTA) by achieving an accuracy, precision, recall, and F1 score of 94.87%, 94.74%, 96.48%, and 95.61%, respectively. FloodSense, with its multimodal fusion capability, achieves efficient large-scale flood event classification under adverse weather conditions, requiring only 19.5 GFLOPs and an inference time of 1.34s per image.

INDEX TERMS Multimodal data fusion, Flood classification, Vision Transformer (ViT), Kolmogorov–Arnold Network (KAN), Sentinel-1, Sentinel-2

I. INTRODUCTION

Flood events raise major concerns in the context of global disasters, which primarily affect human populations, infrastructure, economies, and environmental ecosystem balances [1]. Studies from the climate science bodies of the Intergovernmental Panel on Climate Change (IPCC) highlight that the frequency and gravity of such incidents have seen a gradual increase. In 2021, major floods occurred across the globe and particularly affected Central European regions, including countries like Germany, Belgium, and the Netherlands.

Climate research indicates that these extreme flood scenarios shall become more frequent and intense, and thus point to the need for a proactive flood management system that performs early identification of flood events to aid response rollout for minimizing the livelihood casualties and damages. Timely sensing of flood events is mandatory and has become more important than ever to facilitate the swift evacuation of dense populations, where satellite-based flood detection plays a critical role [2]. These challenges further highlight the growing need for robust sensing and analysis frameworks capable of operating under adverse environmental conditions, where image degradation, reduced visibility, and atmospheric disturbances may affect reliable scene interpretation [3]–[5]. RS provides a crucial vantage point for monitoring the dynamics of a flood and allows continuous tracking of floodwater and affected land covers.

ML techniques struggled to generalize beyond trained regions as they mostly depended on locally engineered features, and hence failed to extract discriminative features of flood and non-flooded areas with a reliable precision rate. The integration of deep learning (DL) with satellite-based remote sensing (RS) has significantly advanced numerous Earth observation activities [6]. The expansions on vision-based learning approaches, such as convolutional neural networks (CNN) and their 3D variants, made flood identification from a large amount of satellite data a reliable task [7]. Further, the expansion of techniques such as recurrent neural networks (RNNs), long short-term memory (LSTMs), and combining them with convolutional modules with ConvLSTMs enabled multi-temporal analysis for flood detection [8], [9]. However, the multi-temporal analysis for flood detection is computationally intensive due to the curse of dimensionality from the rich spectral and temporal resolution of satellite data [10]. Recent advancements in transformers and their extensions to vision transformers (ViT) with attention have enabled new opportunities in flood identification [11].

Optical satellite observations from hyperspectral and multispectral sensors often suffer from challenges of unfavorable climatic conditions and cloud cover. This is because of the lower wavelength ranges of such observation. On the other hand, SAR images are immune to adverse conditions due to their larger wavelength and provide ground observation during day and night [12]. This led to the idea of fusing co-located data from multiple sources, thereby bringing complementary information to support one another and serve a common goal [13]–[15]. Although several advancements have been made in satellite-based flood classification, they still face shortcomings in learning spatial spectral dependencies for flood detection. Traditional ML models fail to generalize to larger spatial contexts, and multi-temporal models such as RNNs and ConvLSTMs struggle with the computational overhead imposed by the curse of dimensionality. Although Vision Transformer (ViT)-based models have demonstrated effectiveness in capturing global spatial dependencies, their application to multimodal SAR–optical fusion learning remains limited. In particular, standalone ViT architectures may not effectively model complementary spectral and cross-modal contextual information without dedicated fusion strategies. Above all, the final decision-making layers are generally implemented using MLPs, which rely on fixed activation functions for nonlinear mapping; however, newer alternatives to MLPs that can perform superior channel-wise nonlinear mapping are not utilized. Furthermore, existing methods often model global contextual semantics and local spatial–spectral dependencies independently, limiting effective cross-modal feature interactions in SAR-optical fusion. The proposed FloodSense addresses these shortcomings by introducing a dual-path complementary learning strategy that combines the global semantic learning power of ViTs, the fine-grained spatial-spectral extraction of the 3D2D-Extract CNN module for cross-modal and complementary feature learning, and the adaptive nonlinear refinement capabilities of KANs combined with MLPs. The combined operation of ViTs, Spatial-Spectral CNNs, and KANs+MLPs is designed to improve cross-sensor feature fusion, nonlinear representation, and robustness in flood classification under diverse conditions. The major contribution of the proposed FloodSense framework is given as follows:

1) We propose FloodSense, a novel dual-path multimodal learning framework for Sentinel-1/2 flood classification that explicitly couples global contextual learning and local spatial–spectral modeling, addressing the limitations of standalone CNN or ViT paradigms for SAR–optical fusion.
2) We introduce ViT-FloodX, a transformer-driven global modeling branch that captures long-range contextual flood semantics from fused multimodal imagery, complementing the limitations of local feature extraction in CNN-based approaches.
3) We design 3D2D-Extract, a hybrid spatial–spectral feature learning mechanism that leverages 3D–2D convolutions to model cross-modal spectral interactions between SAR and optical modalities.
4) We propose Refine-KAN, an adaptive nonlinear feature refinement strategy that integrates KANs with MLPs to improve channel-wise nonlinear mapping and decision-making beyond fixed-activation MLP classifiers.
5) Extensive evaluations on the SEN12-FLOOD benchmark demonstrate that FloodSense achieves competitive classification performance while maintaining computational feasibility for flood monitoring.

The rest of the article is organized as follows. Section *II* provides an overview of the major related works in flood classification. The proposed FloodSense framework is elaborated in Section *III*, whereas Section *IV* and *V* cover the experimental settings, results, and discussion. Finally, Section *VI* concludes the manuscript with insightful comments and directions for future research.

II. RELATED WORKS

Several techniques have been used to classify floods and non-floods from RS data; among these, data-driven and

image-processing methods were the first automated methods. Researchers used Coupled Routing and Excess Storage (CREST) hydrological modeling [16], backscatter-based change analysis [17], and hierarchical split-based thresholding approaches [18] to classify floods and non-floods. These methods generally fail in statistical and image processing tasks and fail to learn from large amounts of spatiotemporal data. ML techniques have evolved to provide a more generalized solution to large-scale flood classification problems. A hybrid thresholding–Markov modeling framework with a Markov Random Field (MRF) is proposed for a computationally efficient flood identification [19]. Fuzzy-based classification and its extended version towards spatiotemporal time series analysis for flood classification have become popular [20]. Random Under-Sampling Boosted (RUSBoost) classification, ensemble-based classification, and Support Vector Machines (SVM) classification were developed and showcase improved performance [21]–[23]

Advancements in DL techniques showcase improved flood sensing compared to ML variants. A neural network-based optical flow estimation is proposed, combining an attention mechanism with positional encoding to surpass standard optical velocimetry [24]. Convolutional GRU, Long Short-Term Memory (LSTM), and U-Net-based models for flood classification have become popular [25]–[27]. Recently, DC4Flood was introduced, utilizing a DL-based clustering approach with multiscale feature integration and dilated convolutions to address issues with standard linear-based clustering approaches [28]. A spatiotemporal variational autoencoder-based unsupervised change analysis model was proposed for flood extent mapping, leveraging reconstruction and contrastive learning to achieve improved generalization across diverse flood events [29]. These techniques and their results indicated notable improvements during the transition from conventional ML to advanced DL frameworks. However, the real-time applicability of these models is questionable, as they are trained and tested in a near-perfect environment with no atmospheric disturbances or clouds.

Although sensor fusion techniques were less explored, they enhance flood mapping by integrating optical multispectral data with SAR structural backscatter under adverse climates and cloud cover. A probabilistic Bayesian-based graphical model learns from fused ASTER Depth Elevation Map (DEM) and RapidEye RS data [30]. SAR and Airborne LiDAR were fused to predict the delineation of the flood extent [31]. The 2009-2010 Data Fusion Contest by the IEEE Geoscience and Remote Sensing Society was aimed at flood event classification from multimodal multi-temporal data [32]. Integration of AVHRR and MODIS RST-FLOOD with contextual flood rules improved revisit durations and enabled continuous monitoring of flood extent [33]. The STFCD methodology combined spatial clustering and temporal fusion to perform change analysis and identification of floods from multisource heterogeneous (MSH) satellite image time series (SITS) [34]. The fusion of Sentinel-1 data with OpenStreetMap, tested on the 2017 Houston flood, confirms the importance of multi-source data for fine-grained urban flood identification [35]. The introduction of the SEN12-FLOOD dataset with co-registered optical and SAR SITS catalyzed the development towards multimodal flood identification models [35], [36]. Recent advancements in cross-domain learning and multimodal fusion have demonstrated the effectiveness of attention-driven feature integration, semantic learning, and large language model (LLM)-assisted multimodal representations for improving cross-domain generalization and complementary feature learning [37]–[40]. Although several fusion-based strategies are in place, they are not sophisticated enough to extract the rich spatial, spectral, and temporal context of cross-sensor data. Furthermore, previous techniques have overlooked the potential of MLPs for final decision-making, without considering the advantages of hybrid MLP variants or advanced adaptive nonlinear alternatives that could enhance decision performance.

III. METHODOLOGY

This section details the methodologies integrated into the proposed FloodSense for enhanced flood classification using dual-path multimodal satellite data fusion. It elaborates on the theoretical foundations of Vision Transformers (ViTs) and Kolmogorov-Arnold Networks (KANs) and further describes the ViT-FloodX, 3D2D-Extract, and Refine-KAN blocks along with their integrated operations.

A. VISION TRANSFORMER (VIT)

The introduction of the self-attention mechanism for natural language processing paved the way for the development of advanced transformer architectures for vision-based applications [41]. Transformers process the input data in parallel, enabling flexibility in handling large data. and various applications. The input data I of size $C \times H \times W$ is initially split into non-overlapping patches of size $M \times N$, where H and W denote spatial extent, and C represents the number of channels. Following the patching operation, each patch undergoes embedding to transform it to a representative dimension of K, and the embedded patch P_i is given in Equation (1), where PE stands for patch embedding.

$$P_i = PE(I) \in \mathbb{R}^{\frac{HW}{M^2} \times K} \tag{1}$$

The transformer block of the ViT can have multiple layers l, each with h attention heads, normalization, and a feed-forward. Compared to the CNN designs, the transformer attention mechanism aids in recognizing and learning long-term dependencies in the images. The attention process initially generates Query ($Q = XW^q$), Key ($K = XW^k$), and Value ($V = XW^v$) vectors from an individual patch x_i, where W^q, W^k, and W^v are the weight matrices. The self-attention is then computed using Equation (2). However, in multi-head attention mode, the self-attention is performed simultaneously across multiple heads, with each head capturing distinct aspects of the data. The resulting outputs are then concatenated and linearly transformed back to the original dimension. The multi-head attention is then given by Equation (3), where

W^o is the projection weight. The residual additions and layer normalizations help to stabilize the outcome of the multi-head attention, whereas the feed-forward layers with MLP perform feature transformations and uplift the representation power of the transformer module.

$$Attention(Q,K,V) = SoftMax(\frac{QK^T}{\sqrt{d}})V \tag{2}$$

$$MultiHead(Q,K,V) = Concat(h_0,h_1...h_n)W^o \tag{3}$$

B. KOLMOGOROV-ARNOLD NETWORKS (KANS)

KANs are developed based on the Kolmogorov-Arnold Representation (KAR) theorem, which identifies all multivariate functions h as a bounded mix of univariate functions with a binary addition operation [42]. Compared to standard MLPs, KANs have shown promising efficiency and adaptability. Conventional MLPs learn fixed nonlinearities using various activation functions, whereas KANs implement a univariate learnable function, such as B-spline activation at the node, which adaptively adjusts activation behaviours in line with the data. The function learning at each node can depict the finer-grained, richer characteristics required for highly nonlinear or irregular distributions in data. Therefore, compared to conventional MLPs, KANs provide enhanced feature extraction and generalisation. The generalized equation for KAR is given by Equation (4), where Γ_k denotes the additive function of the univariate components $\gamma_{k,r}(z_r)$

$$h(\mathbf{z}) = h(z_1,\cdots,z_n) = \sum_{m=1}^{2n+1}\Gamma_k\left(\sum_{r=1}^{d}\gamma_{k,r}(z_r)\right) \tag{4}$$

The B-spline activation makes every connection of the neural

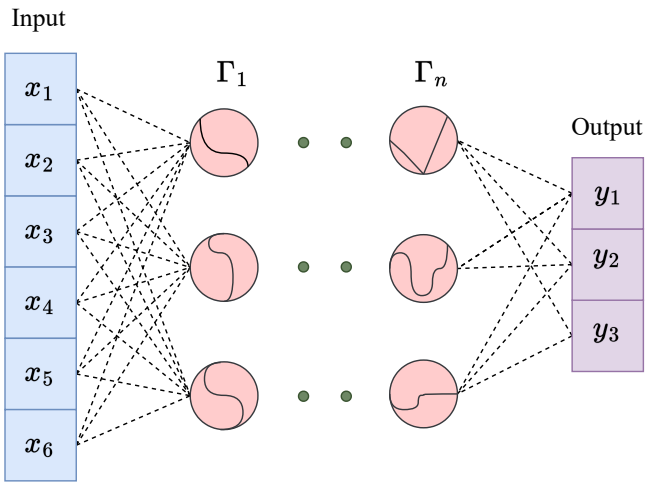

FIGURE 1. The architecture of Kolmogorov-Arnold Network (KAN). `₁ and `ₙ denote the learnable basis activations on neurons

network learnable and enables a smooth and controlled learning of the complex functions Γ_k and $\gamma_{k,r}(z_r)$. The mathematical approximations of these functions are conducted with a weighted sum of multiple B-spline basis functions and are given by Equation (5). The Weight and spline coefficient updates during training are carried out using backpropagation with B-splines, which also determines the typical gradient loss with respect to the spline coefficient.

$$\begin{aligned}\Gamma_k(x) &= \sum_j v_{k,j}B_j^m(x)\\ \gamma_{k,r}(z_r) &= \sum_i w_{k,r,i}B_i^k(z_r)\end{aligned} \tag{5}$$

C. FLOODSENSE ARCHITECTURE

The FloodSense takes fused Sentinel-1 and Sentinel-2 patches $I \in \mathbb{R}^{C\times H\times W}$, where $C = 14$ and $H = W = 85$, as input. Input I is then processed in parallel by the Vision Transformer Flood Feature Extractor (ViT-FloodX) and the Spatial Spectral Feature Extractor (3D2D-Extract). The resulting feature map undergoes hybrid MLP-KAN transformations for effective mapping and classification.

1) ViT-FloodX Branch

ViT-FloodX begins the transformer operation by splitting the fused input into $N_P = \frac{HW}{MN} = 289$ non-overlapping patches of size P_{size} and then projecting individual patches to a latent space of $K = 64$ before applying positional encoding. The augmented positional embeddings patches Z_i are given in (6), where P_i denotes the embedded representation of the i-th patch, $E_{pos,i}$ denotes the positional embedding corresponding to the i-th patch.

$$Z_i^{(0)} = P_i + E_{pos,i} \tag{6}$$

These resulting patches $Z^{(0)} = \{Z_1^{(0)},\ldots,Z_N^{(0)}\}$ are then sequentially processed through L transformer layers, each consisting of layer normalization, a 4-head multi-head attention, residual connections, and a feed-forward network with an MLP head of size $[256, 128]$. At the l-th transformer layer, a 4-head multi-head self-attention is computed and is given by (7)

$$H^{(l)} = \text{MultiHead}_{h=4}\left(Z^{(l-1)}W^q,\ Z^{(l-1)}W^k,\ Z^{(l-1)}W^v\right) \tag{7}$$

The attention output is combined with the input using residual connections and layer normalization, and then processed by a feed-forward network with an MLP head of size $[256, 128]$. This operation produces an intermediate feature vector $Z^{(l)}$ and is given in (8). $Z^{(l)}$ further undergoes a final transformation with another MLP transformation of $[2048, 1024]$ to produce final ViT branch feature vector F_{ViT}, which has global information from multimodal fused S1/S2 patches.

$$Z^{(l)} = \text{MLP}(\text{LN}\left(Z^{(l-1)} + H^{(l)}\right)) \tag{8}$$

2) 3D2D-Extract Branch

The 3D2D-Extract branch specifically extracts spatial-spectral features from the fused data, utilizing 3D-2D convolutions with MLP transformations. First, the input is reshaped into a 5D tensor for 3D convolutions, the transformed input

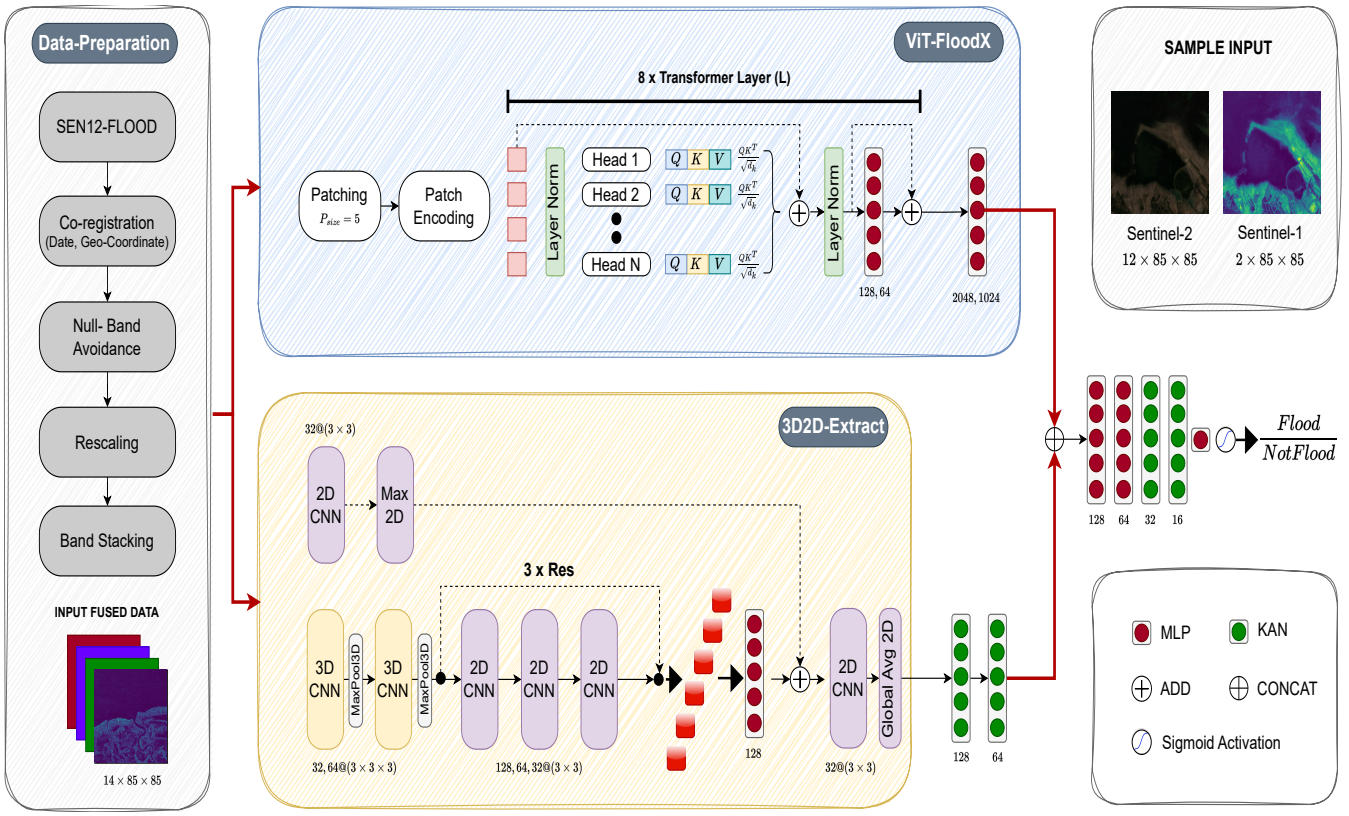

FIGURE 2. Architecture of the proposed FloodSense model detailing Data preparation operations, ViT-FloodX module, 3D-2D extract and Refine-KAN operation with MLP-KAN hybrid structure

is then $I_{3D} \in \mathbb{R}^{1\times C\times H\times W\times 1}$, where the last dimension is an auxiliary singleton dimension introduced to satisfy the input format requirement of the 3D convolution operation. A two-stage 3D convolution is then applied with 32 and 64 filters, using a $3 \times 3 \times 3$ kernel and ReLU activation, followed by max-pooling with a $3 \times 3 \times 3$ window. The convolutions are performed with 'same' padding to preserve spatial resolution. H_1 and H_2 stands for the convolution and the 3DMaxPooled output is denoted as H_p. Following this, the resulting 3D feature map is reshaped into a 2D tensor and processed by 2D CNN blocks with 128, 64, and 32 filters, each using a 3×3 kernel, ReLU activation, and same padding. These 2D CNN blocks are repeated three times in a residual fashion, and then flattened and fed to an MLP of size 128. The resulting feature is subsequently fused with another feature $FCNN$ extracted directly from the original input using a 2D CNN layer with 128 filters, followed by a 2×2 max-pooling operation, Finally, the features are pooled with global average pooling and passed through KAN Layers of size 128 and 64, producing a feature vector $F_{3D\text{-}2D}$. $F_{3D\text{-}2D}$ is given by (9). The MLP units of both ViT-FloodX and 3D2D-Extract implement a dropout of 0.3 to deal with regularization.

$$F_{3D\text{-}2D} = \mathrm{KAN}_{64}(\mathrm{KAN}_{128}(\mathrm{GAP}([F_{\mathrm{cnn}}, F_{\mathrm{direct}}]))) \tag{9}$$

3) Feature Fusion and Classification

The outputs from both branches, F_{ViT} and $F_{3D\text{-}2D}$, are concatenated to form a unified feature representation,F_{fusion}, which is then processed by a fully connected Refine-KAN fusion network. The Refine-KAN fusion network consists of dense layers of 128 and 64 units, each implementing ReLU nonlinearity and a dropout of 0.3. The fusion operation is expressed in (10), where $\sigma(\cdot)$ denotes the ReLU activation function.

$$H_{\mathrm{fusion}} = \mathrm{Dropout}(\sigma(F_{\mathrm{fusion}} W_f), 0.3) \tag{10}$$

Following this, two successive KAN layers of sizes 32 and 16 are employed to capture enhanced nonlinear transformations, thereby enabling effective integration of global attention-driven features with local spatio-spectral features through a hybrid MLP–KAN representation. This process is denoted by (11). Finally, the fused representation is passed through a sigmoid layer for binary classification to produce a predicted flood probability $\hat{y}$, where $\hat{y}$ is given in (12). The overall architecture of the proposed FloodSense framework is illustrated in Fig. 2.

$$H_{\mathrm{KAN}} = \mathrm{KAN}_{16}(\mathrm{KAN}_{32}(H_{\mathrm{fusion}})), \tag{11}$$

$$\hat{y} = \sigma(H_{\mathrm{KAN}} W_c + b_c), \tag{12}$$

IV. RESULTS AND DISCUSSION

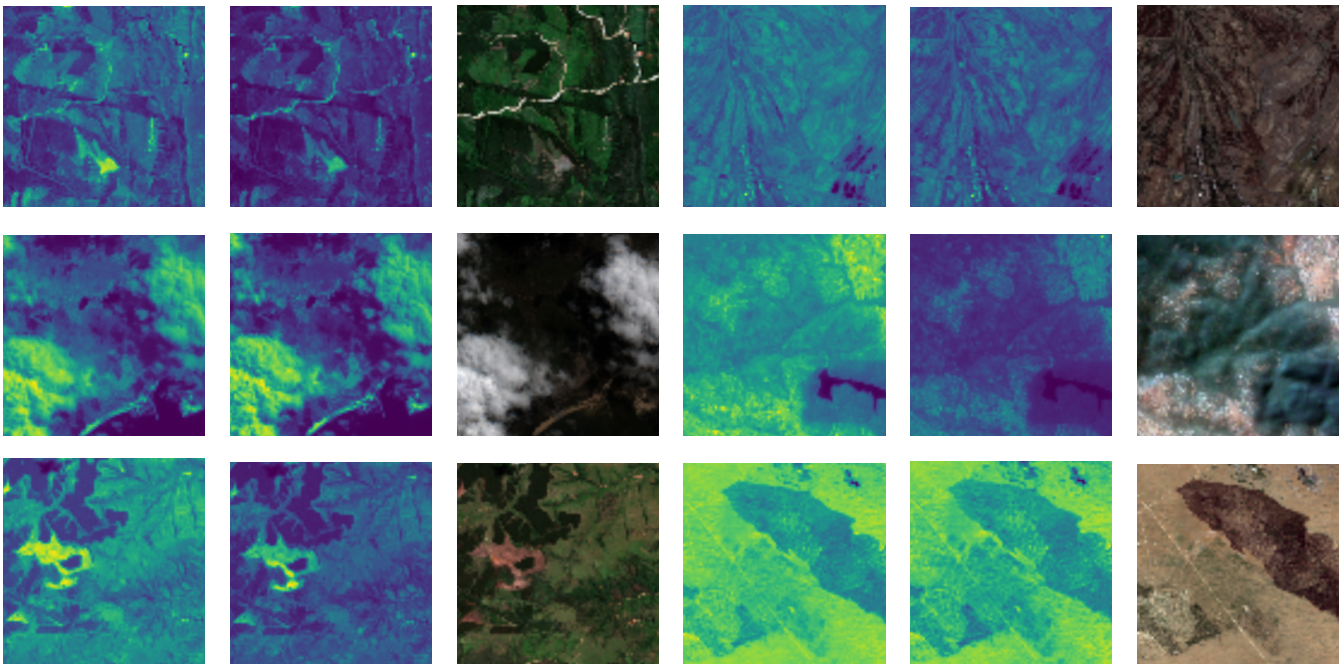

FIGURE 3. Sample flood and non-flood patches from Sentinel-1 and Sentinel-2. (a, b, c) show Sentinel-1 VV, Sentinel-1 VH, and Sentinel-2 RGB composite of flood samples, while (d, e, f) show the corresponding non-flood samples.

A. DATASET

SEN12-FLOOD is the largest available co-registered optical-SAR multimodal fusion benchmark dataset for flood classification [43]. The dataset comprises Sentinel-1 and Sentinel-2 observations collected from the regions of West and South-East Africa, the Middle-East, and Australia. Sentinel-2 provides 12-band multispectral atmospherically corrected imagery, whereas Sentinel-1 offers 2-band radiometrically calibrated and terrain-corrected SAR data consisting of VV and VH polarimetric channels. The co-registered nature of the dataset and the pre-processing applied to Sentinel-1 and Sentinel-2 data help mitigate acquisition geometry and atmospheric correction issues. The dataset is totally comprised of 3332 and 1950 patches of Sentinel-1 and Sentinel-2. The dataset consists of bands with varying spatial resolutions. Sentinel-1 VV and VH bands are mostly of size 521×518, whereas Sentinel-2 bands have a large disparity among band resolutions. B1 and B9 have a size of 85×85, B2, B3, B4, and B8 are of 512×512, and B5, B6, B7, B8A, B11, and B12 are 256×256. To ensure uniformity across sensor modalities, all bands were first resampled to a common 10m spatial resolution and subsequently rescaled to the lowest resolution of 85×85, then stacked band-wise to produce the final fused data of $14 \times 85 \times 85$. These patches are then split into training, validation, and testing patches in a 70:20:10 ratio. The samples from the SEN12-FLOOD dataset are given in Fig. 3.

B. EXPERIMENTAL SETUP

The FloodSense model is trained and tested on a 40GB NVIDIA A100 GPU from Google Cloud Platform (GCP). All experiments, including SOTA comparisons, were conducted under the same hardware environment to ensure fair computational evaluation, with the GPU operating in a dedicated mode without resource sharing. The inference evaluation was performed using standard FP32 precision, without applying any model compression, pruning, or quantization techniques. All implementations were carried out in Python, using Keras as the backend. Inference measurements were recorded after an initial warm-up phase, excluding preprocessing overheads, as they are common across all comparative strategies, while GPU prefetching was enabled to facilitate seamless data loading.

The pre-processed SEN12-FLOOD dataset is rescaled to a spatial resolution of 85×85 with 14 spectral bands. This $14 \times 85 \times 85$ fusion data is then divided into 5×5 patches, producing 289 patches per channel, and then projected to a latent space of 64. The ViTs in the ViT-FloodX module are designed with 8-layer transformer blocks, each performing multi-head attention with 4 heads and implementing a feed-forward MLP head of size $[256, 128]$. Following multi-head attention, concatenation, and layer normalization, an additional hidden MLP mapping of size $[128, 64]$ is employed for feature refinement. Feed forwarding from the last ViT layer is mapped to $[2048, 1024]$ MLP units. The Refine-KAN module has two KAN configurations, one with latent spaces of $[128, 64]$ and the other with hybrid-MLP-KAN configurations of $[128, 64]$

TABLE 1. Comparison of FloodSense with SOTA methods based on performance metrics

Method	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
Baseline CNN (2020) [43]	0.31	88.45	87.22	88.55	87.87	90.51
Early Fusion CNN (2021) [43]	0.28	90.12	89.27	90.16	89.71	91.83
DSE (2023) [44]	0.24	91.65	90.84	92.21	91.51	93.48
ViT Fusion (2023) [45]	0.21	92.84	92.31	93.51	92.90	94.72
ProCANet (2025) [46]	0.20	93.26	93.14	94.22	93.67	95.21
FloodSense (Proposed)	**0.16**	**94.87**	**94.74**	**96.48**	**95.61**	**94.91**

TABLE 2. Comparison of FloodSense with SOTA methods based on model complexity and efficiency

Method	Accuracy (%)	Params (mP)	FLOPs (G)	Time (s)
Baseline CNN (2020) [43]	88.45	8.2	14.6	2.10
Early Fusion CNN (2021) [43]	90.12	12.4	18.3	1.95
DSE (2023) [44]	91.65	18.9	22.7	1.78
ViT Fusion (2023) [45]	92.84	28.6	32.4	1.62
ProCANet (2025) [46]	93.26	35.1	38.9	1.55
FloodSense (Proposed)	**94.87**	**43.48**	**19.5**	**1.34**

MLP, and $[32, 16]$ KAN. The standard Keras Efficient KAN implementation from PyPI uses 3rd-order B-spline functions with a grid size of 3. Furthermore, the model is trained with a fixed learning rate of 0.001, a weight decay of 0.0001, and a batch size selected from the range $[16, 32, 64, 128, 256]$. For fair comparison, all comparative methods were reproduced under a unified experimental setting with a common learning rate of 0.001, weight decay of 0.0001, an optimal batch size of 128, and Adam as the optimizer. The DSE framework was reproduced using a VQGAN encoder–decoder, 1000 diffusion time steps, 200 sampling steps, and a 5-stage encoder–decoder U-Net. The ViT Fusion model employed a similar 5×5 patching strategy to the proposed architecture and shared the common training setup. The ProCANet model was reproduced using the same optimization settings and implemented Cosine Annealing with Warm Restarts consisting of 10 cycles. The test evaluations are conducted on a random test sample of size 125. The performance analysis and feasibility of the proposed FloodSense are conducted through empirical evaluations based on metrics including accuracy, precision, recall, F1 score, AUC, loss, number of parameters, GFLOPs, and inference time.

C. RESULT ANALYSIS

The performance of the proposed FloodSense is tested against state-of-the-art flood classification frameworks, including Baseline CNN [43], Early Fusion CNN [43], DSE [44], ViT Fusion [45], and ProCANet [46]. The Baseline CNN framework employs standard convolutional layers; the Early Fusion CNN integrates optical and SAR data at the input level and fuses them for flood classification. The DSE framework, based on diffusion principles, leverages SAR-to-EO transformation for improved feature representation and flood identification, whereas the ViT Fusion model applies transformer-based global attention for multimodal fusion. Finally, the latest ProCANet combines progressive cross-attention with deep feature refinement. All these framework understudies have been tested and evaluated on the same SEN12-FLOOD dataset for uniformity in test conditions. The characteristics of loss during the training and validation phase of model design are visualized in Figure 4 (a). The graph indicates a stabilized training–validation loss profile. Convergence between training and validation loss is observed at epoch 82, which was selected as the best candidate for testing. The resulting confusion matrix from FloodSense is shown in Figure 4 (b), and read as [54 TP, 3 FP, 65 TN, 3 FN].

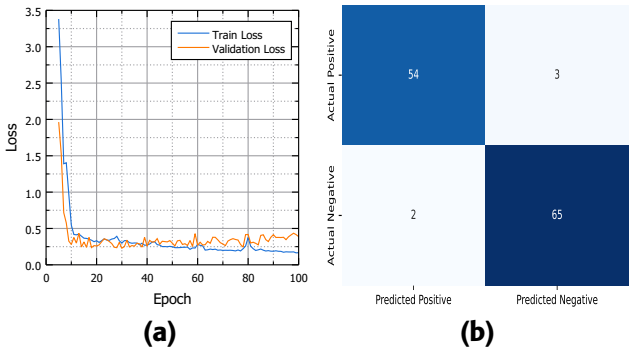

FIGURE 4. (a) Characteristic curve showing the training and validation loss trends (b) Confusion matrices of the full configuration of the FloodSense.

1) SOTA Performance Comparison

Table 1 provides a performance comparison of the proposed FloodSense against SOTA techniques in terms of various evaluation metrics. The Baseline CNN achieves an accuracy of 88.45%, precision of 87.22%, recall of 88.55%, F1 score of 87.87%, and AUC of 90.51% with a loss of 0.31 and inference time of 2.10 seconds. Early Fusion CNN improves performance to an accuracy of 90.12%, precision of 89.27%, recall of 90.16%, F1 score of 89.71%, and AUC of 91.83%, with a reduced loss of 0.28 and time of 1.95 seconds. This is due to the effective integration of the cross-sensor features. However, the diffusion-based DSE model further enhanced flood classification performance and showcases a reliable 91.65% accuracy, 90.84% precision, 92.21% recall, 91.51% F1 score, and 93.48% AUC, along with a loss of 0.24 and time of 1.78 seconds. ViT Fusion provides improved outcomes with 92.84% accuracy, 92.31% precision, 93.51% recall, 92.90% F1 score, and 94.72% AUC, using a loss of 0.21 and inference time of 1.62 seconds. ProCANet further improves performance with 93.26% accuracy, 93.14% precision, 94.22% recall, 93.67% F1 score, and 95.21% AUC at a loss of 0.20 and time of 1.55 seconds. However, the proposed FloodSense shows competing performance with a minimal loss of 0.16 and inference time of 1.34 seconds, while attaining 94.87% accuracy, 94.74% precision, 96.48% recall, 95.61% F1 score, and second-best AUC of 94.91%. FloodSense demonstrates competitive performance compared with previous fusion and transformer-based methods, but it also offers a favorable trade-off between accuracy and efficiency in terms of flood classification. This is possible due to the synergistic integration and learning ability of FloodSense in integrating complex cross-sensor local spatial–spectral features, along with the global semantic context of flood information.

2) Cost-Complexity Evaluation

The complexity analysis of FloodSense is also included in Table 2, which indicates that the baseline CNN has a minimal architecture with only 8.2M parameters and 14.6 GFLOPs. However, this directly affects the limited representation capacity and thus model accuracy. Early Fusion CNN increases complexity to 12.4M parameters and 18.3 GFLOPs due to the added layers at the early sensor fusion stages. The spatial encoding operations of the DSE model again increase the computational complexity to 18.9 M parameters and 22.7 GFLOPs; however, an increase in accuracy is observed. The transformer operations on ViT Fusion model introduced a drastic increase in the model complexity, raising the parameter requirements to 28.6M and GFLOPs to 32.4. ProCANet is the most FLOP-intensive model, with 35.1M parameters and 38.9 GFLOPs. The additional cost incurred from cross-attention mechanisms was partially compensated by improved accuracy. FloodSense utilizes a larger training parameter of 43.48M due to its hybrid CNN-attention architecture; however, it requires far fewer FLOPs, at 19.5 GFLOPs, due to the optimized dual-branch extraction and lightweight fusion. FloodSense achieves better performance while being the fastest among the available multimodal architectures.

D. ABLATION ANALYSIS: IMPACT OF MODEL CONFIGURATIONS

The Table 3 and 4 provides an in-depth comparison of various FloodSense configurations, showing how exclusion of various key modules such as 3D CNN, ViT, MaxPool3D, KAN, Long Skip Connection (L-Skip), and Residual Networks (ResNet) influences the classification performance. The confusion matrices in Figure 5 summarize the performance of FloodSense under different ablation settings. From the figure, it is clear that the absence of 3D CNN or its combination with MaxPool3D resulted in a confusion matrix of [50 TP, 8 FP, 61 TN, 6 FN] and [51 TP, 9 FP, 60 TN, 5 FN] leads to a degradation in classification consistency. Further, the removal of ResNet and ViT moderately affected the classification performance and produced a confusion matrix of [51 TP, 5 FP, 64 TN, 5 FN] and [51 TP, 5 FP, 64 TN, 5 FN], showing only a moderate performance. Excluding KAN or L-Skip also impacts the distribution of misclassifications, producing a confusion matrix of [50 TP, 4 FP, 65 TN, 6 FN] and [51 TP, 7 FP, 62 TN, 5 FN]. Among these models, the exclusion of 3D CNN produced classification imbalance with higher FP, and FN of 8 and 6. A detailed analysis of performance and cost-complexity aspects is presented subsequently.

1) Performance impact on model

Table 3 provides the performance ablation of the proposed model. The configuration with all modules FloodSense original with 43.48 mP, achieved improved performance with a minimal loss of 0.16, and showcased superior accuracy of 94.87%, precision of 94.74%, recall of 96.48%, F1 score of 95.61%, and AUC of 94.91%. This confirms the capability of each module and its interoperability in maximizing classification performance. The table reveals that the 3D CNNs in the 3D2D-Extract are fundamental to FloodSense's promising performance, as their removal from the configuration results in a notable drop in accuracy to 88.46%. This reveals the capability of 3D CNNs in extracting fine-graded spatial-spectral details from rich fused multimodal data for contextual mapping of flood. The effect of 3D CNNs on model performance is tested in two sub-configurations, one with maxpooling and the other without maxpooling. The performance drop was similar for both sub-configurations; however, the sub-configuration without maxpooling increased the parameter count to 71.03% million from 48.66% million in the configuration with maxpooling. This reveals the capability of 3D CNNs in extracting rich spatial spectral details from rich cross-fusion multimodal data for contextual mapping of flood. ResNet modules in the 3D2D-Extract are the next important component contributing to the classification performance of the FloodSense, as their removal dropped accuracy to 89.74%, F1 score, and AUC dropped to 89.56% and 89.71%. Introduction of KAN and L-Skip further improved accuracy, surpassing 90% mark, also all the metrics such as precision, recall, F1 score, and AUC surpassed 90%.

TABLE 3. Accuracy-oriented performance comparison of FloodSense configurations

FloodSense Configurations						Accuracy Metrics					
3D CNN	ViT	MaxPool3D	KAN	L-Skip	ResNet	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
✗	✓	✓	✓	✓	✓	0.21	88.46	86.27	89.39	87.71	88.59
✗	✓	✗	✓	✓	✓	0.18	88.46	85.01	91.16	87.95	88.67
✓	✓	✓	✓	✓	✗	0.20	89.74	87.98	91.19	89.56	89.71
✓	✓	✓	✗	✓	✓	0.23	90.01	90.25	91.14	90.68	91.02
✓	✓	✓	✓	✗	✓	0.20	90.11	91.13	89.35	90.26	91.08
✓	✗	✓	✓	✓	✓	0.25	91.03	90.28	91.12	90.65	91.05
✓	✓	✓	✓	✓	✓	**0.16**	**94.87**	**94.74**	**96.48**	**95.61**	**94.91**

TABLE 4. Cost-complexity and accuracy comparison of FloodSense configurations

FloodSense Configurations						Cost, Complexity & Accuracy Metrics		
3D CNN	ViT	MaxPool3D	KAN	L-Skip	ResNet	Accuracy (%)	Params (mP)	Time (s)
✗	✓	✓	✓	✓	✓	88.46	48.66	0.42
✗	✓	✗	✓	✓	✓	88.46	71.03	0.64
✓	✓	✓	✓	✓	✗	89.74	43.22	1.29
✓	✓	✓	✗	✓	✓	90.01	43.60	1.33
✓	✓	✓	✓	✗	✓	90.11	43.16	1.27
✓	✗	✓	✓	✓	✓	91.03	16.85	1.33
✓	✓	✓	✓	✓	✓	**94.87**	**43.48**	**1.34**

TABLE 5. Performance comparison of single-modality and multi-modality configurations of FloodSense

Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)	Params (mP)	FLOPs (G)	Time (s)
S1-Only (SAR)	90.42	89.65	91.22	90.42	92.58	27.82	12.3	1.12
S2-Only (Optical)	92.16	91.84	93.02	92.42	93.87	29.65	13.1	1.18
FloodSense (SAR+Optical)	**94.87**	**94.74**	**96.48**	**95.61**	**94.91**	**43.48**	**19.5**	**1.34**

TABLE 6. KAN-specific ablation study in FloodSense, revealing the significance of the KAN module over conventional MLPs

Configuration	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)	Params (mP)	Time (s)
FloodSense w/o KAN (MLP Classifier)	0.23	90.01	90.25	91.14	90.68	91.02	43.60	1.33
FloodSense w/ KAN (Proposed)	**0.16**	**94.87**	**94.74**	**96.48**	**95.61**	**94.91**	**43.48**	**1.34**
Gain/Reduction (%)	**30.43↓**	**5.40↑**	**4.97↑**	**5.86↑**	**5.44↑**	**4.27↑**	**0.28↓**	**0.75↑**

However, our empirical evaluations revealed that increasing the KAN order and grid size increased the complexity and inference time required, while only minimally impacting the accuracy metrics. The ViT was moderately critical; however, the combined performance of 3D CNN, ViT, and KANs improved performance to achieve an accuracy of 94.87%.

2) Cost-Complexity impact on model

Table 3 provides the cost-performance aspect of the proposed model. The comparison of the number of trainable parameters, test time against performance, reveals a clear trade-off. The configurations without 3D CNN were the lightest in terms of test time and required only a test time of 0.42, but were affected by undesirable performance. As each module was added to the configuration, the computational complexity in terms of test time gradually increased from a minimum 0.42 to a maximum 1.34 seconds. This gradual increase is evident in the increase in the number of parameters and performance. This reveals the direct relationship between model complexity and performance. However, the complexity has to be added thoughtfully, for example, the second configuration without 3D CNNs and maxpooling resulted in a model with 71.03 mP; however, the increase in parameters did not directly translate to model performance and showcased only an accuracy of 88.46%.

3) Interaction Analysis of ViT, 3D2D-Extract, and Refine-KAN

This section discusses the complementary interactions among the modules ViT-FloodX, 3D2D-Extract, and Refine-KAN, highlighting their significance beyond the observed metric improvements. The ViT-FloodX branch captures global contextual information and large-scale flood semantics derived

TABLE 7. Performance comparison of FloodSense across different batch size configurations

Batch	Loss	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)	Time (s)
B=16	0.39	87.18	84.53	87.57	86.02	87.28	1.43
B=32	0.18	88.46	85.04	91.13	87.97	88.61	0.51
B=64	0.21	89.74	87.98	91.15	89.53	89.72	1.36
B=128	**0.16**	**94.87**	**94.74**	**96.48**	**95.61**	**94.91**	**1.34**
B=256	0.18	88.46	87.58	87.52	87.54	88.49	1.28

from fused Sentinel-1 and Sentinel-2 inputs. This global context is particularly important, as water spread continuity and surrounding land-cover interactions often extend beyond local receptive fields. However, while ViT-FloodX effectively handles global dependencies, minute local spatial–spectral dependencies may not be sufficiently captured. The 3D convolutions of 3D2D-Extract explicitly learn cross-band interactions between Sentinel-1 SAR backscatter and Sentinel-2 spectral information, and their removal results in a loss of fine-grained spatial-spectral learning, reflected in a drop in accuracy from 94.87% to 88.46%. Although ViT and 3D2D-Extract captured the necessary features, their transformations toward the final classification layer are critical. Refine-KAN addresses this challenge by applying adaptive nonlinear transformations, enabling effective feature refinement beyond conventional fixed-activation MLPs. This interaction between Refine-KAN is reflected in the reduced performance after KAN removal, where accuracy dropped to 90.01%, indicating contributions of KANs. This reveals the performance of the FloodSense configuration, complemented by the benefits from the critical modules and their synergistic interactions.

4) KAN-Specific Ablation

A dedicated KAN-specific ablation study is provided in Table 6, which reveals the improvement in classification performance when KAN-based components are integrated in place of conventional MLP neural networks. The analysis shows that incorporating KAN improved the accuracy from 90.01% to 94.87%, corresponding to an increase of 4.86 percentage points ($+5.40\%$ relative improvement). Although the percentage improvement may appear modest, flood detection is performed over large geographical regions, where even small improvements can have meaningful practical significance, particularly in scenarios involving risks to human life and infrastructure. Similar relative gains were observed across other evaluation metrics, including precision ($+4.97\%$), recall ($+5.86\%$), F1-score ($+5.44\%$), and AUC ($+4.27\%$). The analysis also shows a 0.28% reduction in parameter count, while the inference time increased only marginally by 0.75%. These findings suggest that KAN provides more effective nonlinear feature approximation than a conventional MLP, thereby improving flood classification performance without substantially increasing computational complexity.

Apart from the quantitative improvements, the performance gains are due to the architectural characteristics of KANs, which explain why FloodSense with KAN performed better. Conventional MLPs rely on fixed activation functions and linear weight transformations; however, KANs employ learnable nonlinear basis functions, enabling adaptive feature transformations. While distinguishing flooded and non-flooded terrain across heterogeneous land types, the adaptive nonlinear representation capability of KANs allows the model to better capture complex feature relationships and improve classification performance. Furthermore, our KAN implementation utilizes the efficient KAN implementation from PyPI, employing only a 3rd-order B-spline function with a grid size of 3. This relatively lightweight configuration preserves computational efficiency by avoiding excessive spline complexity, thereby maintaining FLOPs and inference time comparable MLPs.

E. ABLATION ANALYSIS: EFFECT OF SAR FUSION

The analysis of the FloodSense architecture with inputs of SAR only, Optical only, and SAR+Optical fusion (Proposed) is given in the Table 5 for ablations of input data and the importance of SAR. From the table, it is clear that SAR auxiliary information is crucial for the proposed framework and enhances its accuracy. Architectures using a single source of data for classification are lighter and have faster inference. The SAR-only input configuration required only 27.82 mP, 12.3 GFLOPs, and 1.12 seconds, whereas the Optical-only configuration required 29.65 mP, 13.10 GFLOPs, and 1.18 seconds. SAR-only configurations were slightly faster, primarily because they have fewer channels compared to the optical-only configurations. However, they have a lower prediction accuracy of 90.42% and 92.16%, due to the limited feature representation compared to the fused SAR-Optical configuration, which has an accuracy of 94.87%. It is also worth noting that, although the fusion configuration can provide superior accuracy, it demands more computational resources in terms of the number of parameters, GFLOPs, and Inference time.

F. ABLATION ANALYSIS: EFFECT OF BATCH SIZE

Table 7 tabulates the FloodSense performance across various batch sizes of $16, 32, 64, 128, 256$. Data from the table clearly shows that the FloodSense with a batch size of 128 demonstrated the best performance by achieving the lowest training loss of 0.16, the highest accuracy of 94.87%, precision of

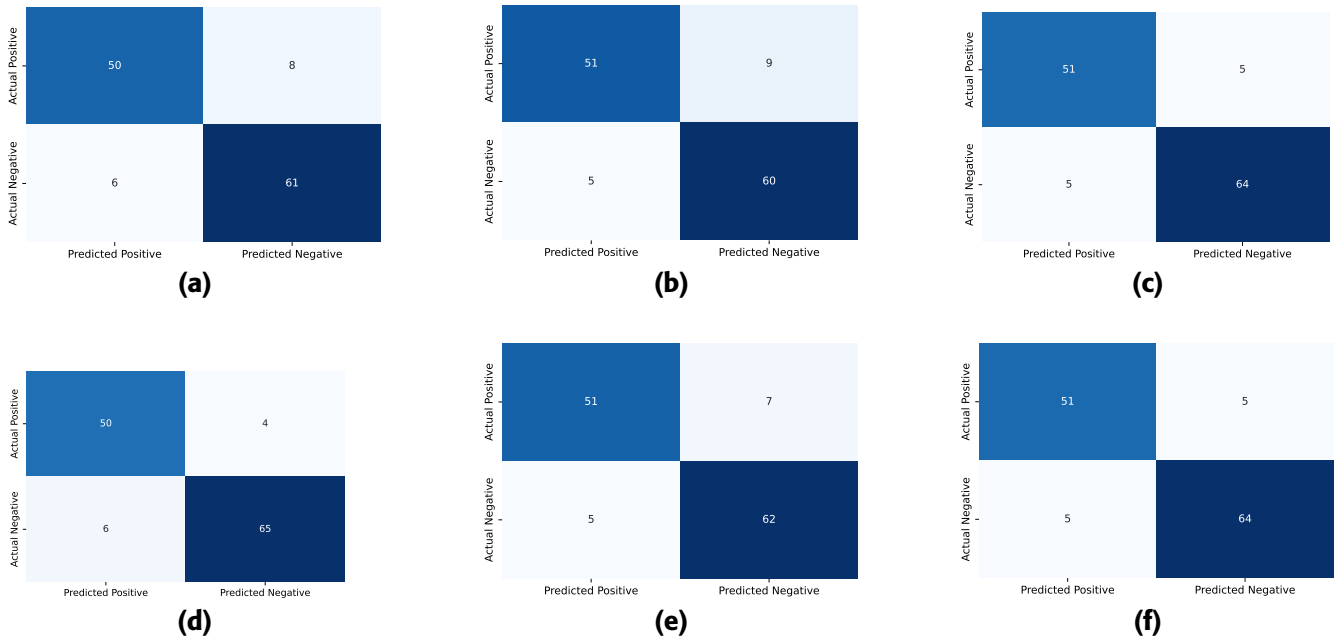

FIGURE 5. Confusion matrices of FloodSense ablation study: (a) Without 3D CNN, (b) Without 3D CNN and MaxPool3D, (c) Without ResNet, (d) Without KAN, (e) Without L-Skip, (f) Without ViT.

TABLE 8. Three-fold cross-validation performance comparison (mean ± standard deviation)

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
Baseline CNN (2020) [43]	88.12 ± 1.43	87.01 ± 1.51	88.14 ± 1.38	87.54 ± 1.42	90.23 ± 1.19
Early Fusion CNN (2021) [43]	89.84 ± 1.16	89.03 ± 1.11	89.92 ± 1.25	89.45 ± 1.18	91.62 ± 1.03
DSE (2023) [44]	91.37 ± 0.98	90.52 ± 1.04	91.94 ± 0.95	91.20 ± 0.99	93.14 ± 0.84
ViT Fusion (2023) [45]	92.46 ± 0.81	91.96 ± 0.89	93.08 ± 0.85	92.51 ± 0.87	94.31 ± 0.76
ProCANet (2025) [46]	92.91 ± 0.72	92.73 ± 0.77	93.85 ± 0.71	93.28 ± 0.75	94.88 ± 0.69
FloodSense (Proposed)	**94.31 ± 0.58**	**94.12 ± 0.63**	**95.84 ± 0.55**	**94.97 ± 0.57**	**95.34 ± 0.52**

TABLE 9. Statistical significance analysis of FloodSense against ProCANet using the Wilcoxon signed-rank test

Metric	FloodSense	ProCANet	p-value
Accuracy	94.31 ± 0.64	92.91 ± 0.72	0.014
Precision	94.12 ± 0.73	92.73 ± 0.77	0.019
Recall	95.84 ± 0.66	93.85 ± 0.71	0.010
F1 Score	94.97 ± 0.67	93.28 ± 0.75	0.013
AUC	95.34 ± 0.46	94.88 ± 0.69	0.042

94.74%, recall of 96.48%, F1 score of 95.61%, and AUC of 94.91%. This enhanced performance can be credited to the clear balance between gradient stability from batches and computational efficiency. Considering the smaller batch sizes of 16 and 32, they failed to showcase a considerable performance and resulted in lower Accuracy, Precision, Recall, F1 score, and AUC values. This shall be due to the effect of more frequent weight updates, limiting the optimization to achieve a generalizable minimum during training. Conversely, as the batch size increased to 256, the performance of the model is reduced and the accuracy dropped to 88.46%. This performance drop may be due to the fact that the optimizer may converge toward sharp minima that fit the training data well but fail to extend to unseen test samples. Although the accuracy is reduced, the test time is lowered to 1.28 seconds compared to 1.34 seconds of batch size 128; however, this is not desirable as the most important metric under consideration is accuracy.

An interesting observation is that the training time for batch size 32 appears slightly lower than expected relative to batch sizes 16 and 64. This can be attributed to hardware-level optimizations such as memory alignment, GPU parallelization efficiency, and how the training workload is distributed across compute cores. The minimal test time observed for batch size 32 is 0.51, which is lower than any other batch configuration; also, they had a minimal loss value of 0.18 seconds to 0.16 loss of batch size 128. For all configurations, the number of trainable parameters was consistent at 43.48 mP.

TABLE 10. Cross-dataset generalization performance of FloodSense on UrbanSARFloods and Sen1Floods11 dataset

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
SEN12-FLOOD	94.87	94.74	96.48	95.61	94.91
SEN12-FLOOD (S1-Only)	90.42	89.65	91.22	90.42	92.58
UrbanSARFloods (Cross-Dataset)	89.42	87.95	89.56	88.73	91.64
Sen1Floods11 (Cross-Dataset)	88.16	87.42	88.91	88.15	90.87

G. CROSS-VALIDATION AND STATISTICAL ANALYSIS

The FloodSense framework undergoes 3-fold cross-validation and reports its average accuracy across folds inTable 8. The table shows that FloodSense outperformed all competing methods, achieving 94.31% accuracy, 94.12% precision, 95.84% recall, 94.97% F1-score, and 95.34% AUC. These results demonstrate cross-validation capability, and we also found that the standard deviation is relatively low across folds, ensuring stable behavior across varying test samples. In comparison, ProCANet was the second-best performer; however, FloodSense led with a 1.40% improvement in accuracy. A 3-fold cross-validation with 3 repeated runs was used to assess the statistical significance of the model relative to the second-best-performing ProCANet. FloodSense consistently achieved superior performance across all evaluation metrics with statistically significant improvements ($p < 0.05$). These findings further support the reliability of the proposed framework under repeated experimental conditions.

H. CROSS-DATASET GENERALIZATION ANALYSIS

To evaluate the generalizability of FloodSense beyond the SEN12-FLOOD dataset, additional cross-dataset experiments were conducted using the UrbanSARFloods benchmark dataset [47] and Sen1Floods11 [48]. The FloodSense model was directly evaluated on UrbanSARFloods and Sen1Floods11 without retraining to assess its cross-region and cross-domain adaptability. Since both datasets primarily focus on SAR-based flood mapping, only the SAR branch of FloodSense was used for inference. The evaluation was performed using the official train-validation-test split provided by the dataset authors. Table 10 presents the cross-dataset evaluation results. Although a moderate performance drop is observed compared to the SEN12-FLOOD benchmark, FloodSense consistently maintains strong segmentation capability under complex flood conditions. In particular, the model achieves an F1-score of 88.73% and an AUC of 91.64% on UrbanSARFloods, while obtaining an F1-score of 88.15% and an AUC of 90.87% on Sen1Floods11, demonstrating that the learned multimodal feature representations generalize effectively across geographically diverse flood events, heterogeneous land-cover conditions, and varying SAR imaging scenarios.

V. DISCUSSION

The experimental evaluation demonstrates that FloodSense achieves consistent performance improvements over state-of-the-art flood classification models on the SEN12-FLOOD dataset. The superior performance across accuracy, recall, F1 score, and loss highlights the effectiveness of the proposed hybrid architecture in capturing both local spatial–spectral cues and global semantic context. In particular, integrating 3D CNN–based feature extraction with ViT-driven global attention enables FloodSense to effectively exploit the inter-band and cross-modal correlations inherent in SAR–optical data. The ablation analysis further confirms that no single module alone is responsible for the observed gains; rather, performance improvements emerge from the synergistic interaction of the 3D CNN backbone, ViT-based contextual modeling, KAN-based nonlinear refinement, and long skip connections. This architectural synergy leads to more stable convergence behavior, reduced misclassification rates, and a notable improvement in recall, which is especially critical for flood detection tasks where false negatives can have severe real-world implications. Beyond accuracy, FloodSense offers a favorable balance between computational complexity and inference efficiency, making it suitable for near-real-time flood monitoring applications. Although the proposed model has a higher parameter count than earlier CNN-based approaches, its optimized dual-branch design reduces FLOPs and inference time compared to other transformer-heavy models. The cost–complexity analysis reveals that thoughtful architectural design is more impactful than merely reducing parameters, as evidenced by lighter configurations that failed to translate reduced complexity into better performance. Furthermore, the modality ablation study underscores the critical role of SAR–optical fusion, where SAR contributes robustness to atmospheric disturbances and optical data provides rich spectral detail, together yielding a substantial accuracy gain over single-modality setups. The batch-size analysis further reinforces the importance of training stability, with batch size 128 offering the best trade-off between gradient reliability and generalization. Overall, these findings suggest that FloodSense is not only a high-performing model but also a computationally efficient and deployable solution for large-scale, operational flood classification systems.

VI. CONCLUSION

Flood classification remains a critical challenge due to the limitations of conventional ground, UAV, and optical satellite-

based approaches, especially under adverse weather conditions. In this work, we proposed FloodSense, a novel dual-path multimodal data fusion framework that integrates Sentinel-1 SAR and Sentinel-2 optical data for reliable flood event identification. The ViT-FloodX, 3D2D-Extractfor, and Refine-KAN components of the FloodSense handles cross-domain spatial-spectral feature extraction, semantic flood pattern mapping, and nonlinear feature refinement. Experimental evaluation on the SEN12-FLOOD dataset demonstrates that FloodSense surpasses SOTA methods, showcasing its superiority among all evaluation metrics under consideration. These results highlight its robustness, scalability, and adaptability for large-scale, real-world flood monitoring and emergency response, even under cloud-covered and challenging atmospheric conditions.

REFERENCES

[1] S. Manes, M. M. Vale, and A. P. F. Pires, "Nature-based solutions potential for flood risk reduction under extreme rainfall events," *Ambio*, vol. 53, no. 8, pp. 1168–1181, Aug. 2024.

[2] M. Nadella, G. V. K. Rayalu, M. S. Akshay, S. K. Eswar Sudhan, V. V. Sajith Variyar, V. Sowmya, and R. Sivanpillai, "Semi supervised flood damage detection using satellite images," in *Computational Intelligence and Data Analytics*, A. C. Frery, R. Buyya, R. M. R. Kovvur, and T. H. Sarma, Eds. Singapore: Springer Nature Singapore, 2025, pp. 163–174.

[3] S. Zhang, X. Zhang, L. Shen, S. Wan, and W. Ren, "Wavelet-based physically guided normalization network for real-time traffic dehazing," *Pattern Recognition*, vol. 172, p. 112451, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320325011136

[4] Y. Liu, X. Wang, E. Hu, A. Wang, B. Shiri, and W. Lin, "Vndhr: Variational single nighttime image dehazing for enhancing visibility in intelligent transportation systems via hybrid regularization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 7, pp. 10 189–10 203, 2025.

[5] S. Zhang, X. Zhang, S. Wan, W. Ren, L. Zhao, and L. Shen, "Generative adversarial and self-supervised dehazing network," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 3, pp. 4187–4197, 2024.

[6] T. N. Prabhakar and P. Geetha, "Two-dimensional empirical wavelet transform based supervised hyperspectral image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 133, pp. 37–45, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271617303039

[7] D. Sudiana, I. Riyanto, M. Rizkinia, R. Arief, A. S. Prabuwono, J. T. S. Sumantyo, and K. Wikantika, "Performance evaluation of 3-d convolutional neural network for multitemporal flood classification framework with synthetic aperture radar image data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 3198–3207, 2025.

[8] M. Moishin, R. C. Deo, R. Prasad, N. Raj, and S. Abdulla, "Designing deep-based learning flood forecast model with convlstm hybrid algorithm," *IEEE Access*, vol. 9, pp. 50 982–50 993, 2021.

[9] A. Mohan, V. M., P. P., and J. A., "Temporal convolutional network based rice crop yield prediction using multispectral satellite data," *Infrared Physics & Technology*, vol. 135, p. 104960, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1350449523004188

[10] R. Yadav, A. Nascetti, and Y. Ban, "Attentive dual stream siamese u-net for flood detection on multi-temporal sentinel-1 data," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 5222–5225.

[11] G. I. Drakonakis, G. Tsagkatakis, K. Fotiadou, and P. Tsakalides, "Ombrianet—supervised flood mapping via convolutional neural networks using multitemporal sentinel-1 and sentinel-2 data fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2341–2356, 2022.

[12] Y. Aoki, M. Furuya, F. De Zan, M.-P. Doin, M. Eineder, M. Ohki, and T. J. Wright, "L-band synthetic aperture radar: Current and future applications to earth sciences," *Earth, Planets and Space*, vol. 73, no. 1, p. 56, Feb. 2021.

[13] G. Dong, C. Zhao, X. Pan, and A. Basu, "Learning temporal distribution and spatial correlation toward universal moving object segmentation," *IEEE Transactions on Image Processing*, vol. 33, pp. 2447–2461, 2024.

[14] S. Zhang, X. Zhang, W. Ren, L. Shen, and J. Zhang, "Shape-aware and feature fused power line detection network," *Engineering Applications of Artificial Intelligence*, vol. 170, p. 113981, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197626002629

[15] J. Anandakrishnan, V. M. Sundaram, and P. Paneer, "Cermf-net: A sar-optical feature fusion for cloud elimination from sentinel-2 imagery using residual multiscale dilated network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 11 741–11 749, 2024.

[16] S. I. Khan, Y. Hong, J. Wang, K. K. Yilmaz, J. J. Gourley, R. F. Adler, G. R. Brakenridge, F. Policelli, S. Habib, and D. Irwin, "Satellite remote sensing and hydrologic modeling for flood inundation mapping in lake victoria basin: Implications for hydrologic prediction in ungauged basins," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 85–95, 2011.

[17] L. Giustarini, R. Hostache, P. Matgen, G. J.-P. Schumann, P. D. Bates, and D. C. Mason, "A change detection approach to flood mapping in urban areas using terrasar-x," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 2417–2430, 2013.

[18] M. Chini, R. Hostache, L. Giustarini, and P. Matgen, "A hierarchical split-based approach for parametric thresholding of sar images: Flood inundation as a test case," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 6975–6988, 2017.

[19] S. Martinis, A. Twele, and S. Voigt, "Unsupervised extraction of flood-induced backscatter changes in sar data using markov image modeling on irregular graphs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 251–263, 2011.

[20] C. Li, J. Liu, X. Liu, X. Kang, and S. Li, "Combining time-series variation modeling and fuzzy spatiotemporal feature fusion: A novel approach for unsupervised flood mapping using dual-polarized sentinel-1 sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[21] P. Ghasemigoudarzi, W. Huang, O. De Silva, Q. Yan, and D. T. Power, "Flash flood detection from cygnss data using the rusboost algorithm," *IEEE Access*, vol. 8, pp. 171 864–171 881, 2020.

[22] C. Krullikowski, C. Chow, M. Wieland, S. Martinis, B. Bauer-Marschallinger, F. Roth, P. Matgen, M. Chini, R. Hostache, Y. Li, and P. Salamon, "Estimating ensemble likelihoods for the sentinel-1-based global flood monitoring product of the copernicus emergency management service," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 6917–6930, 2023.

[23] T. Igarashi and H. Wakabayashi, "Detection of flood area caused by typhoon hagibis using learning-based method with sentinel-1 data," in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 7202–7205.

[24] Y. Cao, Y. Wu, Q. Yao, J. Yu, D. Hou, Z. Wu, and Z. Wang, "River surface velocity estimation using optical flow velocimetry improved with attention mechanism and position encoding," *IEEE Sensors Journal*, vol. 22, no. 16, pp. 16 533–16 544, 2022.

[25] S. Miau and W.-H. Hung, "River flooding forecasting and anomaly detection based on deep learning," *IEEE Access*, vol. 8, pp. 198 384–198 402, 2020.

[26] C. Prakash, A. Barthwal, and D. Acharya, "Floodwall: A real-time flash flood monitoring and forecasting system using iot," *IEEE Sensors Journal*, vol. 23, no. 1, pp. 787–799, 2023.

[27] R. Jaturapitpornchai, R. Saito, N. Kanemoto, and S. Kuzuoka, "Flood area detection from a pair of sentinel-1 synthetic aperture radar images using res-unet based method," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 1023–1026.

[28] K. Rafiezadeh Shahi, A. Camero, J. Eudaric, and H. Kreibich, "Dc4flood: A deep clustering framework for rapid flood detection using sentinel-1 sar imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[29] R. Yadav, A. Nascetti, H. Azizpour, and Y. Ban, "Unsupervised flood detection on sar time series using variational autoencoder," *International Journal of Applied Earth Observation and Geoinformation*, vol. 126, p. 103635, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843223004594

[30] D. Frey, M. Butenuth, and D. Straub, "Probabilistic graphical models for flood state detection of roads combining imagery and dem," *IEEE*

Geoscience and Remote Sensing Letters, vol. 9, no. 6, pp. 1051–1055, 2012.
[31] D. C. Mason, M. S. Horritt, J. T. Dall'Amico, T. R. Scott, and P. D. Bates, "Improving river flood extent delineation from synthetic aperture radar using airborne laser altimetry," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 3932–3943, 2007.
[32] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, and Q. Du, "Multi-modal change detection, application to the detection of flooded areas: Outcome of the 2009–2010 data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 331–342, 2012.
[33] M. Faruolo, I. Coviello, T. Lacava, N. Pergola, and V. Tramutoli, "A multi-sensor exportable approach for automatic flooded areas detection and monitoring by a composite satellite constellation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 2136–2149, 2013.
[34] Z. Wang, X. Wang, W. Wu, and G. Li, "Continuous change detection of flood extents with multisource heterogeneous satellite image time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.
[35] Z. Liu, J. Li, L. Wang, and A. Plaza, "Integration of remote sensing and crowdsourced data for fine-grained urban flood detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 13 523–13 532, 2024.
[36] A. A, E. J. A, and S. N, "Sar-floodnet: A patch-based convolutional neural network for flood detection on sar images," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2022, pp. 195–200.
[37] W. Wu, S. Song, X. Qiu, X. Huang, F. Ma, and J. Xiao, "Image fusion for cross-domain sequential recommendation," 2025. [Online]. Available: https://arxiv.org/abs/2502.15694
[38] W. Wu, Z. Chen, X. Ma, W. Zhang, X. Qiu, S. Song, X. Huang, F. Ma, and J. Xiao, "Contrastive prompt clustering for weakly supervised semantic segmentation," *Expert Systems with Applications*, vol. 316, p. 131880, 2026.
[39] W. Wu, X. Chen, Z. Chen, J.-E. Jiang, K.-F. Tsang, X. Huang, F. Ma, and J. Xiao, "Tag-enriched multi-attention with large language models for cross-domain sequential recommendation," *IEEE Transactions on Consumer Electronics*, vol. 72, no. 1, pp. 1130–1137, 2026.
[40] W. Wu, Z. Chen, W. Zhang, S. Song, X. Qiu, X. Huang, F. Ma, and J. Xiao, "Llm-enhanced multimodal fusion for cross-domain sequential recommendation," *Expert Systems with Applications*, vol. 321, p. 132228, 2026.
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
[42] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljacic, T. Y. Hou, and M. Tegmark, "KAN: Kolmogorov–arnold networks," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=Ozo7qJ5vZi
[43] C. Rambour, A. Giros, M. Baetens, and P. Defourny, "Flood detection in time series of optical and sar images," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS)*, vol. XLIII-B2-2020, 2020, pp. 1343–1349.
[44] M. Seo, J. Jung, and D.-G. Choi, "Improved flood insights: Diffusion-based sar-to-eo image translation," *Remote Sensing*, vol. 17, no. 13, p. 2260, 2025. [Online]. Available: https://doi.org/10.3390/rs17132260
[45] I. Chamatidis, D. Istrati, and N. D. Lagaros, "Vision transformer for flood detection using satellite images from sentinel-1 and sentinel-2," *Water*, vol. 16, no. 12, 2024. [Online]. Available: https://www.mdpi.com/2073-4441/16/12/1670
[46] V. Feliren, F. Khikmah, I. Dwiki Bhaswara, B. I. Nasution, A. M. Lechner, and M. R. U. Saputra, "Progressive cross-attention network for flood segmentation using multispectral satellite imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 22, pp. 1–5, 2025.
[47] J. Zhao, Z. Xiong, and X. X. Zhu, "Urbansarfloods: Sentinel-1 slc-based benchmark dataset for urban and open-area flood mapping," in *CVPR Workshops*, 2024.
[48] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, "Sen1floods11: a georeferenced dataset to train and test deep learning flood algorithms for sentinel-1," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 835–845.

JAYAKRISHNAN ANANDAKRISHNAN received the bachelor's degree in Computer Science and Engineering from Kerala University, India, in 2011, and the master's degree in Computer Science with a specialization in Image Processing from the Cochin University of Science and Technology, India, in 2013. He received his Ph.D. in Computer Science and Engineering from the National Institute of Technology, Puducherry, Karaikal, India, in 2025. He expanded his expertise through an internship at the International Internship Pilot Program funded by the National Science and Technology Council of Taiwan, further enriching his knowledge and experience in his field. He is currently a full-time Assistant Professor with the School of Computing, Amrita Vishwa Vidyapeetham, Coimbatore Campus, India. His research interests encompass machine learning, remote sensing, geospatial analytics, IoT, and edge computing.

ALKHA MOHAN is a Postdoctoral Researcher at the International Institute of Artificial Intelligence and Management, National Yunlin University of Science and Technology, Taiwan. She earned her Ph.D. in 2022 from NIT Surathkal, India following an M.Tech. and B.Tech. from CUSAT, Kerala, India in 2013 and 2011, respectively. Dr. Mohan's research interests include Machine Learning, Deep Learning, Data Science, Remote Sensing, and Precision Agriculture. She has held Assistant Professor positions at IIIT Tiruchirappalli, VIT Vellore, GITAM Bangalore, and CE Karunagappally. Dr. Mohan has contributed to research in hyperspectral image classification, dimensionality reduction crop yield analysis, and spatio temporal analysis and satellite image processing etc.. Her research findings have been published in reputed high impact journals and supreme conferences.

VIPIN VENUGOPAL (Member, IEEE) received the master's degree from Amrita Vishwa Vidyapeetham, Coimbatore, India, and the Ph.D. degree from the National Institute of Technology Puducherry (NITPY), India. He is currently an Assistant Professor (Selection-Grade) with the School of Artificial Intelligence (AI), Amrita Vishwa Vidyapeetham. His academic journey is marked by a profound interest in cutting-edge technologies, particularly in the realms of AI and medical image analysis. His research interests include image processing, deep learning, machine learning, and medical image analysis. His expertise in these areas is demonstrated by his contributions to renowned international peer-reviewed journals and conferences, where he has published several impactful papers.

WEIWEI JIANG (Member, IEEE) received the B.Sc. degree in electronic engineering and Ph.D. degree in information and communication engineering from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2013 and 2018, respectively. He is currently an Assistant Professor with the School of Information and Communication Engineering, and the Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing. He is one of 2023 and 2024 Stanford's List of World's Top 2% Scientists. He has published more than 50 academic papers in IEEE Trans and other journals, with more than 3100 citations in Google Scholar. His current research interests include artificial intelligence, machine learning, big data, wireless communication, and edge computing.

MOHD AMIRUDDIN ABD RAHMAN (Member, IEEE) received the B.Sc. degree in electrical engineering from Purdue University and the M.Sc. degree in sensor and instrumentation from UPM. He is currently an Associate Professor of data science with the Department of Physics, Universiti Putra Malaysia (UPM). He is the Deputy Director of the University's Research Management Centre. He is a Pioneer in co-tutelle Ph.D. program, having earned the first such degree from both The University of Sheffield and UPM. His research experience extends globally, including stints as an Invited Researcher with Alcatel-Lucent Bell Labs and a Postdoctoral Fellow with Indian Institute of Technology Kanpur. He has also played a pivotal role in national research initiatives, contributing to the smart data and big data projects at the Ministry of Higher Education. He leads the Computational Intelligence Research Group at UPM and his research focuses on machine learning and artificial intelligence, with over 50 high-impact publications in the field. His areas of interest also include signal processing, pattern recognition, machine learning algorithms, electromagnetics-related computation and modeling, and RF and microwave-based sensor systems.

JAWAD KHAN received the master's degree in computer science from Kohat University of Science and Technology, Pakistan, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea. He worked for three and half years as a Postdoctoral Researcher with the Department of Robotics, Hanyang University, South Korea. He is currently an Assistant Professor with the School of Computing, Gachon University, South Korea. His research interests include natural language processing, information retrieval, sentiment analysis/opinion mining, text processing, social media mining, artificial intelligence, machine learning, deep learning, the Internet of Things, and computer vision.