Information Fusion 129 (2026) 104005

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Mitigating class imbalance in forest fire prediction with GAN-Augmented data fusion

Vishal Krishna Singh [a,*], Deepshikha Agarwal [b], Vivek Kumar Gediya [b,c], Rajkumar Singh Rathore [d], Weiwei Jiang [e]

[a] School of Computer Science and Electronics Engineering, University of Essex, Colchester, CO43SQ, Essex, UK
[b] Indian Institute of Information Technology, Lucknow, C.G. City, Lucknow, 226002, Uttar Pradesh, India
[c] Advanced Micro Devices (AMD), Xilinx India Technology Services Pvt Ltd, Hyderabad, Telangana, 500032, India
[d] Department of Computer Science, School of Technologies, Cardiff Metropolitan University, Cardiff, CF5 2YB, UK
[e] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China

ARTICLE INFO

Keywords:
Bayesian network
Bias mitigation
Conditional tabular generative adversarial network
Forest fire
Synthetic data

ABSTRACT

Imbalanced data sets exacerbate recognition biases in forest fire prediction models, as disproportionate representation of class instances leads to skewed results. Existing work on bias mitigation has limited ability to generalize and extract features specific to forest fires. Internet of Things (IoT)-based sensor networks can provide real-time, granular data on environmental factors such as temperature, humidity, and soil moisture, helping to capture the dynamic nature of forest conditions and alleviate data imbalance. To address these challenges, this work introduces a novel hybrid approach that explores complex probabilistic relationships among environmental factors, incorporating IoT-driven data, and using a generative adversarial network (GAN) to synthetically augment minority classes. The proposed model is validated on publicly available datasets, and the performance is reported on evaluation metrics such as accuracy, precision, recall, F1-score, computational efficiency and training cost. The results show that the proposed hybrid model is able to achieve a significant improvement over the exiting methods achieving classification accuracy of 95.08 %, a precision of 93.03 %, a recall of 92.80 %, and an F1-score of 92.91 %.

1. Introduction

1.1. Background and motivation

Forest fires are a critical danger to wildlife survival and results in significant loss to the local economy and environmental conditions. As such, considerable efforts have been made in the field of accurate detection and early prediction of the nature of the forest fires [1,2]. Technological interventions, lead by Internet of Things (IoT) driven predictive modeling, is one of the major research frontier and helps timely human intervention [3]. Similarly, Satellite Remote Sensing, together with IoT networks have shown diverse applications including crop health, ground-level temperature, and humidity levels [4,5].

However, due to limited accuracy and significantly high computation requirements, existing methods fail to achieve the desired objective. The harsh terrain and frequently changing meteorological variables, further impose severe constraints on the model prediction accuracy. This, along with the highly imbalanced and skewed datasets, require advanced methods of machine learning (ML) and data analytics (DA) to achieve accurate real-time inference. The literature identifies data imbalance as one of the main reasons for the failure of ML methods, owing to the low representation of certain classes. This may happen due to oversampling the minority class, under-sampling the majority class, or a combination of both, and impacts the model's performance.

Recent years have seen a major growth in the use of resampling techniques to address the issues imposed due to skewed class distributions. However, with the inherent constraints, imposed due to under-sampling the majority class or oversampling the minority class, several optimizations have been proposed to address this issue. A suitable example of this is presented in the *Synthetic Minor Oversampling Technique* (SMOTE) [6], which selectively extracts subsets from underrepresented classes to prevent overfitting caused by mere replication of rare instances. It subsequently generates new instances that resemble the existing ones. Nonetheless, a drawback is the potential for class overlap or noise introduction, as it does not consider the context of the general data in proximity to the underrepresented class data.

* Corresponding author.
E-mail address: v.k.singh@essex.ac.uk (V.K. Singh).

https://doi.org/10.1016/j.inffus.2025.104005
Received 11 June 2025; Received in revised form 6 November 2025; Accepted 24 November 2025
Available online 1 December 2025

More examples of enhancing the performance by resampling includes the method proposed in [7], which introduced two novel minority over-sampling techniques, *borderline-SMOTE1*, and *borderline-SMOTE2*, derived from the established SMOTE method. These methods specifically target minority instances located near the data boundary, offering a refined approach to addressing imbalanced datasets. The study broadly categorizes solutions for imbalanced datasets into data-level and algorithmic-level methods. While data-level methods aim to modify the distribution of imbalanced datasets, algorithmic-level methods propose new data mining algorithms or modify existing ones to effectively handle the imbalance issue. The authors address the issue of the two-class problem in imbalanced domains, but fail to address the limitation imposed due to multiple classes. A different approach, proposed in [8], presents a forest fire prediction method that utilizes a sparse autoencoder-based deep neural network and a novel data balancing procedure. This method outperforms other state-of-the-art methods in predicting large-scale forest fires, reducing mean absolute error by 19.3 % and root mean squared error by 0.95 %. The proposed method is evaluated using a real-world dataset that contains the environmental conditions of forest fires and the corresponding burned size. In a different method, the authors in [9], utilize *Random Forest Regression* along with Hyperparameter tuning through RandomizedSearchCV to forecast forest fires, taking into account meteorological factors like temperature, rain, wind, and humidity. Various models are evaluated for predicting forest fires, including *Decision Tree*, *Random Forest*, *Support Vector Machine* (SVM), and *Artificial Neural Networks* (ANN).

1.2. Related work

One of the seminal works addressing the data imbalance problem in recommendation systems, is the hybrid Generative Adversarial Network (GAN) proposed in [10,11]. The proposed method implements a conditional Wasserstein GAN (W-GAN) with gradient penalty to generate tabular data containing both numerical and categorical values. The proposed scheme is able to augment auxiliary classifier loss to enforce the model to explicitly generate data belonging to the minority class. Moreover, to address the data imbalance problem, the author in [12] proposed anomaly detection method based on message importance measure (MIM)-based GAN. Further, to address the mode collapse problem, the discriminator architecture was designed using the concept of Pac-GAN, capable of receiving *m-packed* samples as input instead of a single input. Other important methods, addressing the issue of data imbalance, include the method proposed in [13] which combines GAN with the Random Forest algorithm for anomaly detection and the autoencoder based method proposed in [14]. The technique incorporates an autoencoder along with two distinct discriminators to enhance its effectiveness in identifying anomalies. Following a similar approach, the authors in [15] proposed a novel Autoencoder-GAN architecture to detect operational anomalies.

However, the inability of these approaches to capture the complex dependencies and intricate patterns present in the data, prohibits high prediction accuracy [18]. Traditional methods have limited capacity to generalize and extract specific features of forest fires. Existing methods have limited ability to address the frequently changing environmental conditions in forest scenario and ignore the critical characteristics unique to forest fires. The inherent drawbacks of these approaches to address the complex relationships between the environmental factors, specific to the forest fires, leads to inaccurate labeling of different types of fire and their unique characteristics in diverse environments. Furthermore, the computational complexity and resources required to execute the methods, add to the network requirements and failure to address the issues imposed due to data imbalance [20].

1.3. Contributions of this work

This work is aimed at addressing the existing challenges of inaccurate prediction and inability to deal with imbalanced datasets through a Bayesian Network (BN) based method. The proposed method is able to capture the intricate probabilistic dependencies among various environmental features and a *Conditional Tabular Generative Adversarial Networks (CTGAN)*, is used to refine and generate minority classes by enhancing their quality and realism. The contributions of the work are summarized as follows:

- A novel BN is proposed and is trained on the original imbalanced dataset, capturing the intricate probabilistic dependencies among various environmental features. Using the proposed BN, the initial conditional samples are generated to encapsulate the underlying distribution of the dataset. The samples produced by the BN are fed into a novel, *CTGAN*, to undergo refinement and to generate minority class by enhancing their quality and realism.
- The proposed *CTGAN* is trained to focus on minority class and minimize distinguishability between real and synthetic samples, ensuring that the generated data closely mirrors the distribution of the real dataset.
- *Statistical summaries*, *correlation plots*, and *feature importance (FI) scores* are employed to assess the synthetic data, while the performance of the proposed predictive model is validated through metrics such as *accuracy*, *precision*, *recall*, and *F1-score*.

1.4. Paper organization

The remainder of this paper is organized as follows: Section 2 outlines the problem statement, while Section 3 provides a detailed description of the dataset. In Section 4, we introduce the proposed methodology, followed by a description of the simulation environment in Section 5. The Section 6 presents a comprehensive evaluation of the proposed method through extensive simulations. Finally, Section 7 concludes the paper.

2. Problem statement

The primary challenge at the core of this research lies in the inherent data imbalance within the context of forest fire prediction. This imbalance gives rise to a recognition bias, often referred to as a *Type-II error*. This bias manifests as the predictive model fails to identify the real class that carries the utmost significance in forest fires. Specifically, the model frequently misclassifies instances of *large-scale* fire as *small-scale* fire or even as *no-fire*. The critical nature of this issue stems from the fact that large-scale fires pose the most substantial threat to both the forest ecosystem and human safety.

The mathematical essence of this challenge can be distilled into the equation for Type-II error, or False Negative Rate (FNR) in Eq. (1), derived from representing the False Negative (FN) and True Positive (TP). This metric underscores the model's oversight in recognizing the critical 'large-scale' fire class amidst the imbalance.

$$FNR = \frac{FN}{FN + TP}. \tag{1}$$

In the case of accurate forest fire detection, the accuracy and appropriate resource allocation for large-scale fires is of paramount importance. However, the limited number of instances, representing large-scale fires, exacerbates the recognition bias problem. Thus, addressing recognition bias and effectively predicting large-scale fires becomes an urgent and vital objective.

3. Dataset description

The forest fire data as used and presented in [17] is used for this study. The data was collected in the *Montesinho Natural Park*, in Portugal which is a region of high flora and fauna with diverse vegetation types. The region is marked by the *Supra-Mediterranean* climate and the average annual temperature is recorded in the range 8 to 12 °C. The data collection started in *January* 2000 to *December* 2003 at two different sources. Features such as *time*, *date* and *spatial location* were collected at

Table 1
Features Describing the Occurring Condition of the Forest Fire Data [17].

Features	Description	Value
X	X-coordinate Montesinho park's Map	1 to 9
Y	Y-coordinate Montesinho park's Map	2 to 9
Day	Days of Data Collection	*Monday* to *Sunday*
FFMC	Surface Litter Moisture Levels	18.7 to 96.20
DMC	Shallow Organic Layer Moisture Levels	1.1 to 291.3
DC	Deep Organic Layer Moisture Levels	7.9 to 860.6
ISI	Index reflecting the Rate of Fire Spread	0.0 to 56.10
Temp	Temperature	2.2 to 33.30 (in °C)
RH	Relative Air Humidity	15.0 to 100 (%)
Wind	Wind Velocity	0.40 to 9.40 (km/h)
Rain	Precipitation Amounts Outside	0.0 to 6.4 (mm/m^2)
Area	Affected Area by Fire	0.00 to 1090.84 (hectare)
Month	Duration of Data Collection	January December 2000 to 2003

Table 2
Target Class Distribution.

Labels	Count
no_fire	247
small_fire	255
large_fire	15

a daily basis (during the period of forest fires) in a 9×9 grid. The second collection was made by the *Braganca Polytechnic Institute*, and consisted of various metrological variables such as wind speed, temperature etc. The observations were manually integrated to form a single dataset containing 517 observations. Other details of the used dataset are presented in Table 1.

The dataset exhibits a significant skewness in the target variable, 'area', which represents the burned area of the forest. The majority of instances have a value of 0.00, making the distribution highly skewed towards zero. Recognizing the challenges posed by right-skewed distribution, a shift from a regression task to a classification task is advocated. Through an empirical examination of the dataset, Eq. (2) is used to take into account the threshold value and enables the transformation of regression into classification. Table 2 shows the class distribution of each instance of the dataset.

$$\text{fire\_scale} = \begin{cases} \text{no\_fire}, & \text{if } x = 0, \\ \text{small\_fire}, & \text{if } 0 < x < 74.2, \\ \text{large\_fire}, & \text{otherwise.} \end{cases} \tag{2}$$

To define the *large_fire* category, a cutoff of 74.2 hectares was applied. This choice of cutoff is motivated by two factors: (i) it corresponds approximately to the upper quantiles of our dataset's distribution of burned areas. This allows the *large_fire* class to be adequately represented and (ii) it lies within the range of thresholds commonly used in wildfire management practice around the globe. For example, in Portugal, official statistics classify wildfires larger than 100 hectares as "large fires", while the European Forest Fire Information System (EFFIS) routinely tracks fires starting from approximately 30–40 hectares [33,34] and [35]. Further to this, we performed a comprehensive *sensitivity analysis*, by varying the cutoff between 30 and 150 hectares in increments of 10. For a fair analysis, the dataset was relabeled and the model was re-evaluated using the same leakage-preventive strategy for each threshold. Fig. 1 presents the performance curves for *macro-F1* and *PR-AUC* which represents the outcome for large_fire vs. rest. It can be observed from the plot that the curve remains stable for a wide range of the threshold proving that our conclusions are robust and not critically dependent on the exact cutoff chosen.

3.1. Data splitting strategy and leakage analysis

To ensure robust model evaluation, temporal and spatial splitting and leakage prevention strategies were applied to the dataset.

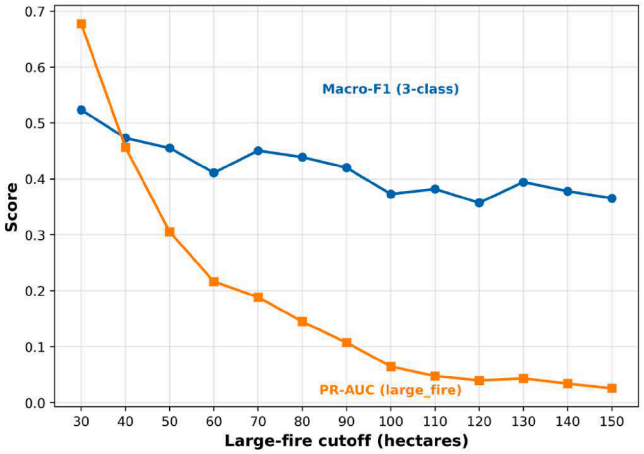

Fig. 1. Sensitivity of Classification Performance to the *large_fire* Cutoff.

3.1.1. Temporal split

The dataset, as described in Section 3, consists of observations starting from January 2000 and ending in December 2003. A strict chronological split was adopted to prevent temporal leakage such that the range for *Training set* was considered from January 2000 to December 2002 and the range for *Testing set* was considered from January 2003 to December 2003.

3.1.2. Spatial split (Grouped Cross-Validation)

The spatial locations were represented as a 9×9 grid, with each cell treated as a distinct group. Grouped cross-validation was performed to prevent spatial leakage. In each iteration, entire grid cells were withheld for testing, while the remaining cells were used exclusively for training.

3.1.3. Leakage analysis

To assess the impact of different data partitioning schemes, three splitting strategies were compared:

1. **Random Split:** Baseline scenario with a random 80/20 partition.
2. **Temporal Split:** Chronological partition as described in Section 3.1.1.
3. **Spatial Split:** Grouped cross-validation as described in Section 3.1.2.

3.1.4. Out-of-distribution (OOD) testing

To evaluate the model's generalization and performance on unknown samples, OOD experiments were also conducted. For the *Temporal OOD*, the training sample was the data between 2000–2002 and the data in 2003 was exclusively used for testing. Additionally, for the *Spatial OOD*, the training on a subset of grid cells was performed and was tested on entirely unseen locations, validating its robustness in real-world deployments.

Fig. 2 presents the outcome of the temporal and spatial splitting strategies and OOD testing. It can be clearly observed in Fig. 2(a) and (b) that the training (blue) and test (red) sets do not overlap in time and space. This reflects the successful execution of the strategy to ensure that the model is unable to see future observations or neighboring grid cells during training. Fig. 2(c) and (d) presents the outcome of the OOD testing on unseen temporal and spatial data (orange). The plot confirms that the model is capable of generalizing beyond the training distribution. Furthermore, the clear separation supports the robustness of the splitting strategy proving that the strategy works efficiently and there is no data leakage.

4. Proposed methodology

The structure and representation of the data are comprehended, and essential tasks are applied to ensure suitability for model input. The data

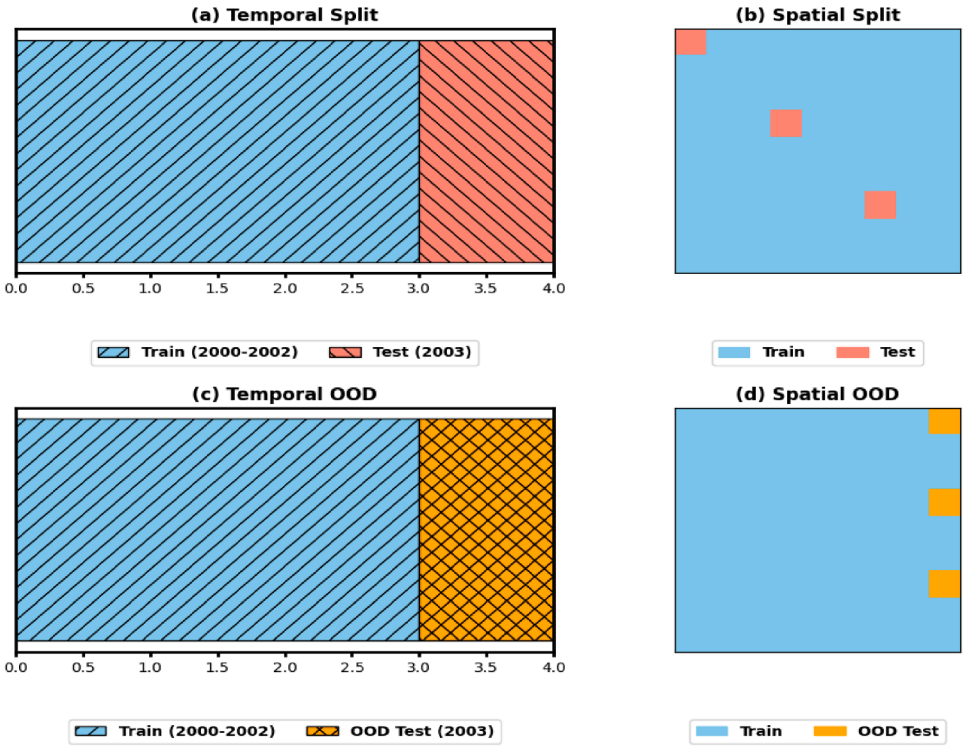

Fig. 2. Splitting with OOD Testing.

undergoes preprocessing, encompassing cleaning, missing data imputation, and formatting as necessary. Although the dataset does not require preprocessing, further exploratory data analysis (EDA) proves valuable in elucidating data behavior and patterns. Utilizing correlation analysis, statistical summaries, and data distribution facilitates understanding a real-time, scalable, and adaptable approach for efficiently processing commands and replies. Following the completion of data preprocessing and EDA, the refined data is channeled into the hybrid methodology, which uses the proposed *BN based method* and the proposed *CTGAN*. The detailed representation of the proposed method is shown in Fig. 3 and the algorithm is presented as Algorithm 1.

Algorithm 1 Proposed Method.

1: $D_{BN\_con} \leftarrow$ BAYESIAN_NETWORK(BNs, D_{org}, N_{BN})
2: $D_{synth} \leftarrow$ CTGAN_OVERSAMPLING$(D_{org}, BN_con, E, B, C_{\min}, V_{target})$
3: $D_{aug} \leftarrow$ MERGE(D_{org}, D_{synth})
4: $(X, y) \leftarrow$ TRAIN_TEST_SPLIT(D_{aug}, V_{target})
5: $Alg_{trained} \leftarrow$ TRAIN(Alg_{class}, X, y)
6: **return** $D_{synth}, Alg_{trained}$

4.1. Theoretical foundation of the BN-guided CTGAN

The proposed BN-guided CTGAN can be interpreted as a two-stage generative process that approximates the empirical joint distribution of the data. The Bayesian Network (BN) first estimates structured conditional dependencies among features, expressed as $P_{\text{BN}}(X \mid Y)$, while the Conditional Tabular GAN (CTGAN) subsequently refines these dependencies through adversarial learning to produce realistic and diverse samples $P_G(X \mid Y)$. The joint target distribution of the real data can therefore be represented as

$$P_{\text{data}}(X, Y) \approx P_G(X \mid Y)\, P_{\text{data}}(Y), \tag{3}$$

Here, $P_{\text{data}}(Y)$ denotes the empirical class prior (often imbalanced), and $P_G(X \mid Y)$ represents the conditional distribution learned by the generator. In this framework, the BN provides structured priors that guide the CTGAN toward the underlying data manifold, particularly for underrepresented classes.

To formalize this interaction, the overall generator objective can be expressed as an adversarial loss regularized by a divergence between the BN-generated and CTGAN-generated distributions:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda\, D_{\text{KL}}\big(P_{\text{BN}}(X \mid Y) \,\|\, P_G(X \mid Y)\big), \tag{4}$$

where $\mathcal{L}_{\text{GAN}}$ is the standard conditional adversarial loss and $D_{\text{KL}}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. The regularization term constrains the generator to remain close to the BN's learned conditional distribution, stabilizing adversarial training and mitigating mode collapse, especially when minority-class data are sparse.

Expanding the conditional adversarial component, the min-max optimization can be written as

$$\min_G \max_D \; \mathbb{E}_{(x,y)\sim P_{\text{data}}}[\log D(x \mid y)] + \mathbb{E}_{z\sim P_z,\, y\sim P_{\text{BN}}(y)}[\log(1 - D(G(z, y) \mid y))], \tag{5}$$

where z is the latent noise vector and $P_{\text{BN}}(y)$ represents class-conditioned sampling guided by the BN. This modification ensures that minority-class conditions are sampled more frequently and with meaningful contextual dependencies, allowing the generator to better approximate the true conditional distributions. The resulting alignment between generated and real minority distributions can be related to a reduction in expected misclassification risk. Let R denote the expected classification risk and y_{minor} the minority class; then, the approximate difference in risk can be expressed as:

$$\Delta R \approx \int \left[P_{\text{data}}(x \mid y_{\text{minor}}) - P_G(x \mid y_{\text{minor}})\right]^2 dx, \tag{6}$$

indicating that as the generated distribution $P_G(x \mid y_{\text{minor}})$ approaches the true minority distribution $P_{\text{data}}(x \mid y_{\text{minor}})$, the expected false-negative risk decreases. Therefore, the BN-guided regularization not only improves the fidelity of synthetic data but also enhances downstream classifier performance by achieving better minority-class coverage.

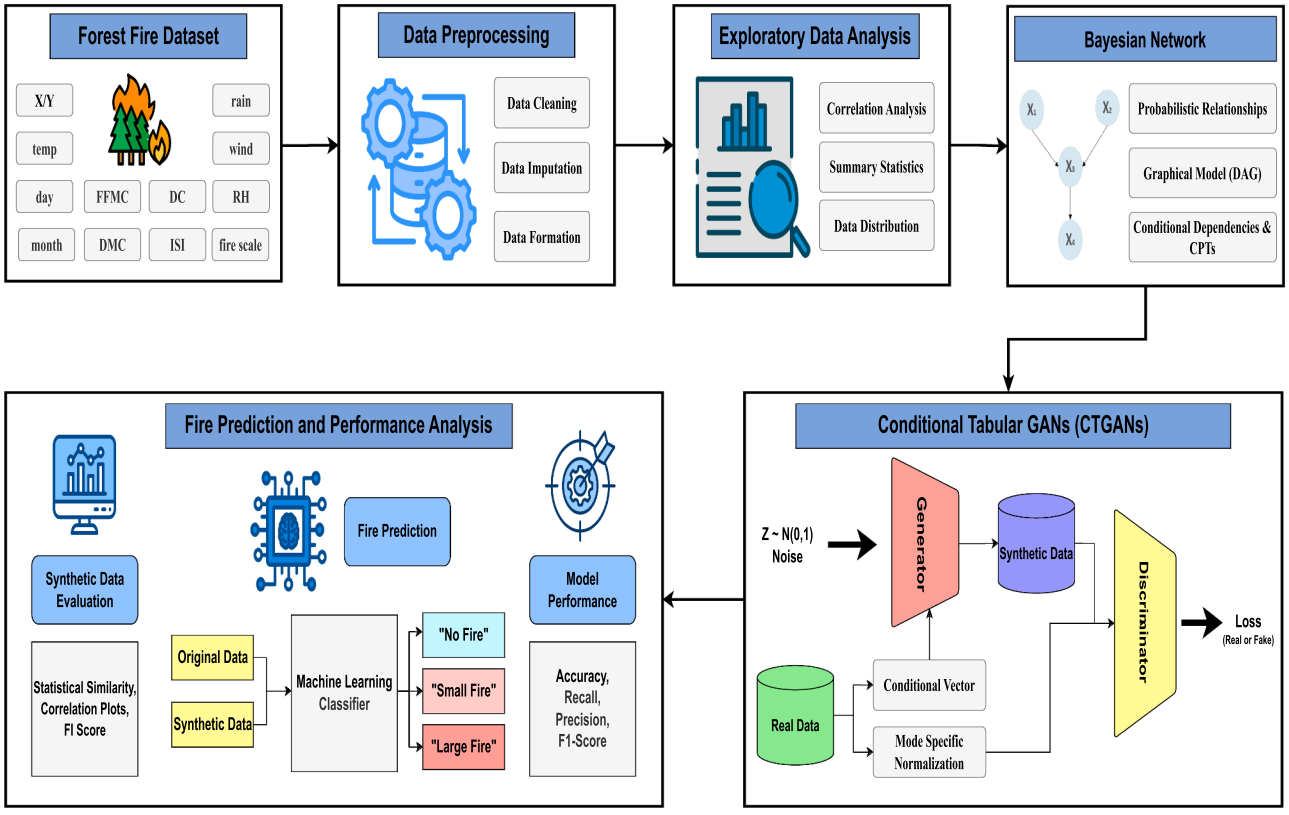

Fig. 3. An Overview of the Proposed Methodology.

Synthetic data is used to feed as a classification algorithm to make predictions of forest fires. Various classification algorithms such as *Decision Tree, Random Forest, Extra Tree, LightGBM, CatBoost, SVM,* and *ANN* were used, and the best classifier model was selected for obtaining the results (explained in detail in Section 6).

4.2. Bayesian network

BNs are adept at encapsulating the complex probabilistic dependencies within the data. By modeling the intricate interplay of variables that influence forest fire occurrences, this work aims to derive conditional samplings that not only provide a more representative sample of minority classes but also preserve the integrity of the underlying data distribution. The algorithm for the proposed BN based method is given as Algorithm 2.

Initially, the data is passed through the BN, i.e. a probabilistic graphical model that represents the joint probability distribution over a set of variables. The idea is to model the conditional dependencies, found between the meteorological variables, to identify the inter-dependencies within the data. Therefore, to model the probability of a given variable to be considered as a risk is evaluated in the context of current meteorological conditions. Mathematically, this is represented as:

$$P(X_i|\text{Parents}(X_i)) = \frac{P(X_i \cap \text{Parents}(X_i))}{P(\text{Parents}(X_i))}. \tag{7}$$

Furthermore, the joint probability distribution of all the identified variables in the given network can be represented using the chain rule as (8):

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1) \\ P(X_3|X_1, X_2) \ldots P(X_n|X_{1:n-1}) \tag{8}$$

Algorithm 2 Proposed Bayesian Method.

Require: BN_{cpd}: Bayesian Network with learned CPDs, N: number of conditional samples
Ensure: $samples$: list of generated samples

```
Model ← BN_cpd
Nodes ← TOPOLOGICALSORT(Model)
samples ← LIST(N)
for s ← 1 to N do
    sample ← {}
    for all node ∈ Nodes do
        if HASNOPARENTS(node) then
            Dist ← PRIOR(node)
        else
            V_parents ← VALUES(sample, PARENTS(node))
            Dist ← CONDITIONAL(node, V_parents)
        end if
        sample[node] ← SAMPLE(Dist)
    end for
    samples[s] ← sample
end for
return samples
```

and,

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)). \tag{9}$$

The Eq. (9), therefore is the building block of the probabilistic model, allowing it to capture the complex relations among various environmental and fire-related factors. Using the Bayes' rules, the probability of a given variable to be true, given that the occurrence of its parents is already verified, we work with the Eq. (10).

$$P(X_i|\text{Parents}(X_i)) = \frac{P(\text{Parents}(X_i)|X_i)P(X_i)}{P(\text{Parents}(X_i))}. \quad (10)$$

and calculate the Maximum Likelihood Estimation (MLE) using Eq. (11).

$$L(\theta|X) = \prod_{i=1}^{n} P(x_i|\theta). \quad (11)$$

Using the CPTs within the BN, the proposed approach uses the probabilistic aggregation as Eq. (12):

$$P(X_i) = \sum_j P(X_i|\text{Parents}(X_i)_j)P(\text{Parents}(X_i)_j). \quad (12)$$

Eq. (12) ensures that the marginal probability of each variable X_i in a network is correctly determined by summing over all possible configurations of its parent variables, denoted as $Parents(X_i)_j$. In practical terms, it guarantees that the computed probabilities in each CPTs accurately reflects the combined effect of the parent variables on X_i, thereby preserving the integrity of the probabilistic relationships within the proposed model. For instance, in estimating the likelihood of specific fire scales, this approach allows us to comprehensively consider the influence of various environmental and temporal conditions, thus ensuring a holistic and accurate representation of the factors contributing to forest fire risks. Using the proposed BN model to generate conditional sampling, ensures that it has strong probabilistic dependencies with every feature.

4.3. Conditional tabular generative adversarial networks (CTGAN)

A combination of the original data along with conditional sampling is now fed to the proposed CTGAN. By using the proposed CTGAN, resampling the minority class to re-balance the instances of large_fire class with the other classes is described in Algorithm 3.

Conditional GANs [16] are an extension of GANs where the generation process is conditioned on additional information, typically in the form of class labels or some other categorical information. WGAN-GP [21] and PacGAN, generate synthetic data and oversample minority classes to improve the effectiveness of recommendation models. CWGAN-GP-PACGAN [22] is used to create synthetic data and increase the representation of minority classes, aiming to improve the efficacy of classification models. The proposed CTGAN architecture enables explicit conditioning on the minority class and helps to generate the synthetic data as accurately as the original dataset. To handle the problem of class imbalances, the proposed CTGAN is conditioned to focus on underrepresented classes, helping to generate more data points for these classes and thus creating a more balanced dataset for model training and analysis. A probabilistic understanding of the relationships between various variables can be used to guide the CTGAN in generating more accurate and representative synthetic data.

The proposed CTGAN uses the concepts of the generator and the discriminator as a core of its functioning. Generator (G) in the proposed CTGAN is a neural network that learns to generate synthetic data that resembles the real data. The goal of the generator is to produce data that is indistinguishable from actual data by the discriminator. Generator takes a random noise vector (z) as input. This vector is sampled from a standard normal distribution. Noise vector is passed through multiple layers of the neural network, which transforms it into data. These transformations involve matrix multiplications, biases, and non-linear activation functions and generate synthetic data point that mimics the real data distribution. $G(z;\theta_g)$, where G represents the generator, z is the input noise vector, and θ_g is the parameters (weights and biases) of the generator.

Algorithm 3 Proposed CTGAN.

Require: D_{org}: Original data, BN_{sample}: Bayesian Network, E: Number of epochs, B: Batch size, s: Step size, $cond_i$: Condition Vector, m: Mask vector, Φ_G: Conditional generator parameter, $C_{\min}$: Minority class, V_{target}: Target variable

Ensure: D_{synth}: Synthetic data

$cond_i$, FOR $1 \le i \le m$
$Data \leftarrow$ MERGE(BN_{sample}, D_{org})
$G \leftarrow$ GENERATOR($Data, cond_i$)
$D \leftarrow$ DISCRIMINATOR()
$Nodes \leftarrow$ TOPOLOGICALSORT($Data$)
$D_{\min} \leftarrow$ FILTER($Data, C_{\min}, V_{target}$)
for $e \leftarrow 1$ **to** E **do**
  for $batch \leftarrow$ GETBATCHES($Data, B$) **do**
    $Batch_{\min} \leftarrow$ GETBATCHES($D_{\min}, B$)
    $C_{samples} \leftarrow \{\}$
    for $node \in Nodes$ **do**
      $Dist \leftarrow$ NODEDISTRIBUTION($Data, node, C_{samples}$)
      $C_{samples}[node] \leftarrow$ SAMPLE($Dist$)
    end for
    $Noise \leftarrow$ SAMPLENOISE(B)
    $S_{batch} \leftarrow$ GENERATE($G, Noise, C_{samples}$)
    UPDATE($D, batch, S_{batch}$)
    $L_g \leftarrow \frac{1}{B}\sum_{i=1}^{B}$ CrossEntropy$(m_{(i^*,j)}, B_{i^*})$
      $-\frac{1}{B/s}\sum_{k=1}^{B/s}$ Critic$({S_{batch}}_j^k, cond_i^k)$
    $\Phi_G \leftarrow \Phi_G - 0.0002 \times$ Adam$(\nabla_{\Phi_G} L_g)$
    UPDATE($G, D, L_g, Noise, C_{samples}$)
  end for
end for
$D_{synth} \leftarrow$ GENERATE($G, BN_{sample}, Nodes, E, B$)
return D_{synth}

Discriminator (D) in the proposed CTGAN is another neural network that evaluates data and attempts to distinguish between real and synthetic data. It receives either real data or synthetic data generated by the generator and processes this data through multiple layers (similar to the generator), trying to classify if the data is real or fake. $D(x;\theta_d)$, where D represents the discriminator, x represents the input data, and θ_d the parameters (weights and biases) of the discriminator.The training of the proposed CTGAN involves an adversarial process where the generator and discriminator are trained simultaneously in a zero-sum game. The generator's objective is to maximize the probability of the discriminator making a mistake (i.e., mistaking synthetic data as real data), while the discriminator is implemented to classify real and synthetic data correctly.

5. Simulation environment

Simulations were performed using HP ProDesk 600 *G5-SFF* workstation equipped with a 12th Gen Intel, Core $i7-8700$ 3.2 GHz. The models were implemented within the *Jupyter* integrated development environment using *scikit-learn*, *TensorFlow*, and various ML frameworks, chosen for their robust support for ML and DL techniques. Multiple scenarios were tested by modifying the imbalance levels of the dataset, with class ratios ranging from highly skewed distributions involving minority classes (representing large fires) to more balanced ones. The results are presented for 5000 epochs with a specific focus on enhancing the generalizability for minority classes by using the proposed method with 3500 samples at the model level for the synthetic data generation. For classification, *CatBoost* [26] was employed with *GridSearchCV* [29] (a hyperparameter tuning technique), to get optimal model parameters. The best performing model was configured with a depth of 10, at 1000 iterations with $L2$ leaf regularization of 1 and learning rate of 0.1 and utilized a *MultiClass* loss function.

Table 3
Importance Scores for Features with the Target Variable.

Features	Importance Score (in %)
temp	19.42
RH	17.48
DC	11.55
DMC	11.23
wind	11.23
RSI	9.69
FFMC	9.69
day	6.25
month	6.25
rain	0.57

6. Performance evaluation

The performance of the proposed method is validated on several parameters and the results are presented under three parts. In the first part, we present the statistical analysis by running statistical analysis & modeling on the dataset to derive and present the conditional dependencies. In the second part, the synthetic data from the proposed model is used to evaluate the quality of the data as compared to the original dataset. Finally, in the third part, we conduct an in-depth comparison to evaluate the performance of the proposed method with the existing state-of-the-art methods. The synthetic dataset is generated with the defined techniques and is fed to the classifier, and a detailed comparative analysis is presented.

6.1. Statistical summary

6.1.1. Random forest

To analyze the impact of input features on each class in the 'fire_scale' column ('no_fire', 'small_fire', 'large_fire') more effectively, employing ML models provides feature importance metrics. One suitable approach is to use a Random Forest algorithm such as in [23], which is well-suited for handling both numerical and categorical data and is robust against overfitting, outlined as follows:

1. **Preprocessing the Data:** Encoding the categorical variables and scaling the numerical variables if necessary.
2. **Training the Model:** Fitting a Random Forest Classifier to the data.
3. **Feature Importance:** Analyzing the feature importance, provided by the model to determine which features most strongly influence the classification into 'no_fire', 'small_fire', and 'large_fire'.

By performing the above analysis, the *Importance Score* of each column with the target variable is derived and is presented in Table 3.

6.1.2. Pearson correlation matrix

The *Pearson Correlation Matrix* is calculated for the various environmental variables, to identify the inter-dependencies among them. The inter-dependencies are crucial for understanding the wildfire risk and other meteorological phenomena and is presented in Fig. 4. Reviewing Fig. 4, it is evident that *DMC* and *DC* have a strong positive correlation ($r = 0.68$), suggesting that as one increases, the other does as well. This relationship is pivotal, indicating that both metrics might respond similarly under the same environmental conditions. Conversely, the analysis reveals a notable negative correlation between *temperature* and *RH* as the values is marked as ($r = -0.53$). This inverse relationship is typical in environmental studies, reflecting the physical principle that warmer air can hold more moisture, thus decreasing humidity. This correlation is particularly significant, as it underlines the thermodynamic interactions that might influence fire behavior through moisture availability and temperature variations. Furthermore, the correlation between the *ISI* and *FFMC* is moderately positive ($r = 0.53$), pointing to a potential predictive linkage between these indices in forecasting the initial spread of wildfires.

Table 4
The Relationship Between Meteorological Variables.

Labels	Features
no_fire	↓ humidity
small_fire	↑ temperature
large_fire	↑ temperature and humidity

Based on *random forest modeling* and *pearson correlation analysis*, it is concluded that environmental conditions like *temperature*, *humidity*, and *fuel moisture*, play pivotal roles in determining the scale and intensity of forest fires. The relationship among these meteorological variables is shown in Table 4.

6.1.3. Conditional dependencies

Fig. 5 represents the proposed BN model, detailing the conditional dependencies among key environmental and temporal variables. Considerable dependencies include the direct effects of month and day on all other variables, emphasizing the significant role of temporal factors in predicting fire behavior. Meteorological variables like *temperature*, *RH*, *wind*, and *rain* are also shown to directly influence fire specific indices such as *FFMC*, *DMC*, *DC*, and *ISI*. This, in turn, impacts the fire scale which illustrates the complexity of interactions and the probabilistic influences that these variables exert on the likelihood and scale of forest fires.

6.2. Synthetic data evaluation

6.2.1. Data distribution parameters

Comparative analysis of the original and synthetic data, based on various data distribution parameters such as *mean*, *standard deviation*, *cumulative sum per feature*, *distribution of data per feature*, and *correlation plot* based on the pearson correlation, is used to evaluate the quality of the data as compared to the original dataset.

The graphs presented in Fig. 6 compares the *log-transformed means* and *standard deviations (STDs)* of real and synthetic data, crucial for assessing the fidelity of the synthetic data generated by the proposed approach. The first plot displays a scatter plot where each point represents a variable from the dataset. The x-axis shows the log-transformed means of the real data, while the y-axis corresponds to the synthetic data. The near-perfect alignment of data points along the diagonal line indicates that the synthetic data closely mirrors the central tendency of the real data, demonstrating the model's accuracy in reproducing the statistical properties of the dataset. The second graph presents the log-transformed standard deviations of the real and synthetic data. Similar to the mean, the data points align closely with the diagonal line, illustrating that the variability of the synthetic data is consistent with that of the real data. This alignment confirms that the proposed model not only captures the average trends of the input data but also effectively reproduces the variability inherent in the real dataset. This analysis validates the synthetic data's statistical similarity to the real dataset, affirming the effectiveness of the proposed hybrid model in generating high-fidelity synthetic data.

6.2.2. Cumulative distribution functions

To extend the validation further, Figs. 7 and 8 provide a visual comparison of *cumulative distribution functions* for both *real* and *synthetic* data, showing a close alignment in their trend lines. It presents a detailed evaluation of the synthetic data generated by the proposed approach, comparing it against real data. It is evident from Fig. 7 that cumulative sums for 'month' and 'day' indicate that the synthetic data effectively captures the cyclic nature and distribution. Continuous environmental feature alignment suggests that the hybrid model is adept at capturing and reproducing the underlying distributions that govern forest fire behavior. The cumulative sums for the categorical fire scale feature shows

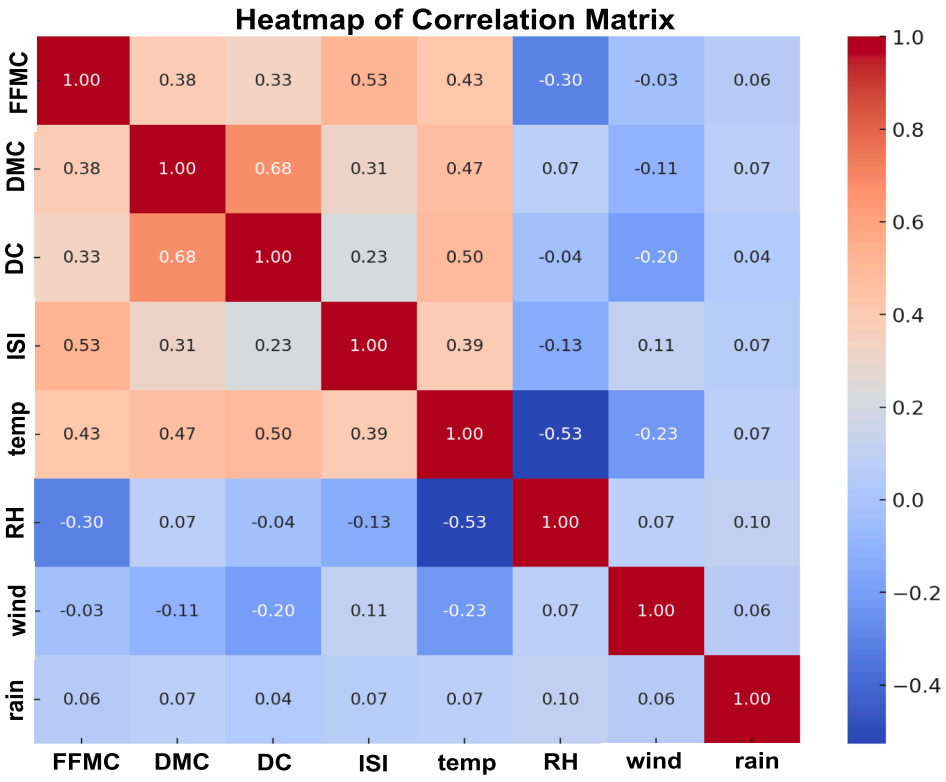

Fig. 4. Pearson Correlation Analysis of Numerical Features.

distinct steps corresponding to the classification threshold, underscoring the model's ability to respect the proportional representation of fire severity categories. In Fig. 8, histograms and density plots compare the frequency distributions of real and synthetic data for the respective features. The histograms for continuous features show that the synthetic data approximates the distribution of real data. The bar charts for features like 'rain' and 'fire scale' illustrate that while there is a reasonable mimicry of the distribution shapes, some discrepancies in exact proportions suggest areas for model refinement, especially in underrepresented categories.

6.2.3. Correlation matrix

Fig. 9 presents the 'heatmaps' comparing the correlation metrics for real and synthetic data and their differences across all data features. The first heatmap displays the correlation among variables in the real dataset, with darker shades indicating stronger relationships mapping the analysis shown in Fig. 4. The second heatmap represents correlations in the synthetic dataset. It mimics the real data's structure well, reflecting the model's capability to understand and replicate the underlying interactions among all forest fire dynamics variables. The third heatmap highlights the differences between the real and synthetic data correlations. The lighter shades across most of the matrix suggest minor differences, indicating that the model effectively captures the essence of the real data's inter-variable relationships. Notable discrepancies in some correlations, particularly involving the fire scale variable, indicate areas where the model could be further refined to enhance the accuracy. This outcome affirms the effectiveness of the hybrid modeling approach in generating synthetic data that preserves critical statistical properties required for accurate and reliable forest fire prediction.

6.2.4. Principal component analysis

Two scatter plots representing the first two principal components (PCs) of a PCA for both real and synthetic data are shown in Fig. 10. The first and the second plot show the PCA results for the real and synthetic datasets respectively. Both plots exhibit almost similar distribution of data points, suggesting that the synthetic data generation process, successfully managed to capture the underlying structure of the real dataset. However, there are subtle differences in the spread and density of points which are dependent on the requirement and number of data points generated which is not related to the performance of the hybrid model.

6.3. Model performance and evaluation

A comprehensive assessment of the proposed hybrid approach, that integrates BN with CTGAN to handle highly imbalanced datasets for predictive modeling, is presented. The performance of the proposed hybrid model is evaluated using standard classification metrics, including *accuracy*, *precision*, *recall*, and *F1-score*.

Fig. 11 represents a comparative analysis of various classification models evaluated on the forest fire dataset, using metrics discussed above. Ensemble methods like - *Random Forest* [23], *Extra Tree* [24], *LightGBM* [25], and *CatBoost* [26] consistently show high performance across all metrics, typically due to their ability to handle complex datasets with imbalanced classes and a large number of features. These models benefit from multiple learning algorithms, reducing the likelihood of overfitting and improving prediction accuracy. *Decision Tree* and *ANN* [28] provide moderate results, where decision trees are generally more interpretable but might suffer from overfitting. *SVMs* [27] are typically robust at handling high-dimensional data but can be sensitive to the choice of kernel and regularization parameters.

Fig. 14 presents a comparative analysis of all the previously used classification models, optimized with *GridSearchCV* and evaluated on the same dataset using the same performance metrics. *GridSearchCV* was employed as a systematic method to optimize hyperparameters through cross-validated grid-search over a parameter grid. This en-

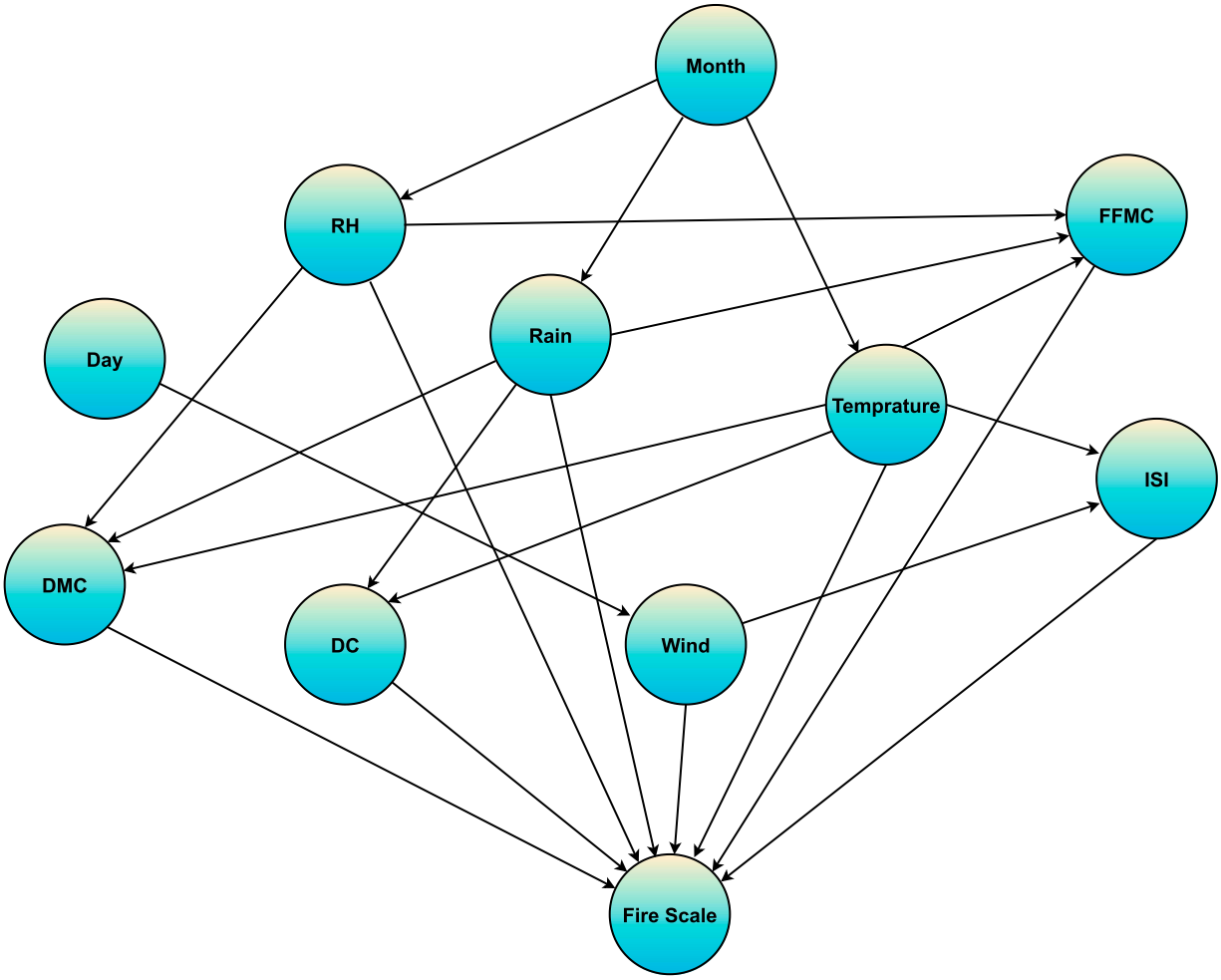

Fig. 5. Conditional Dependency Graph with the Proposed Bayesian Method.

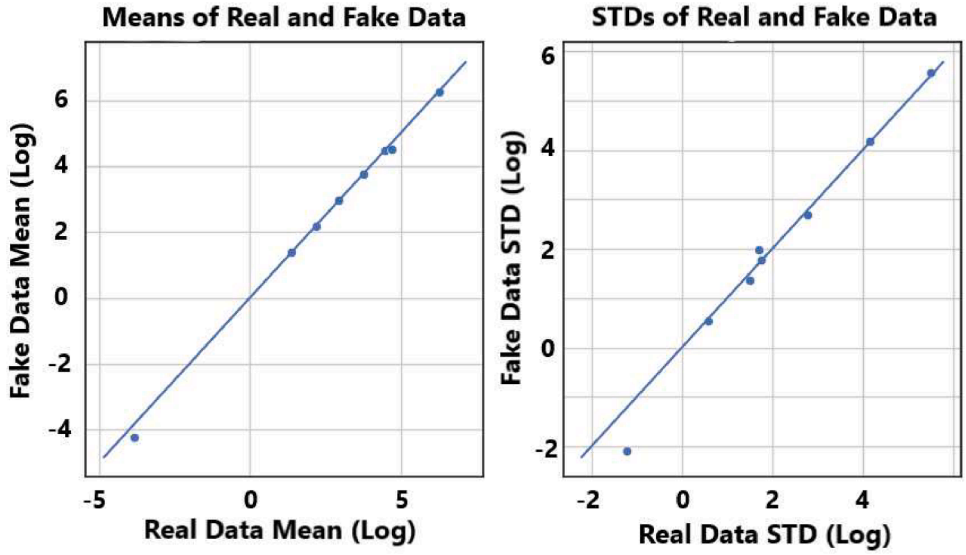

Fig. 6. Comparison of Mean and Standard Deviation of Original and Synthetic Data.

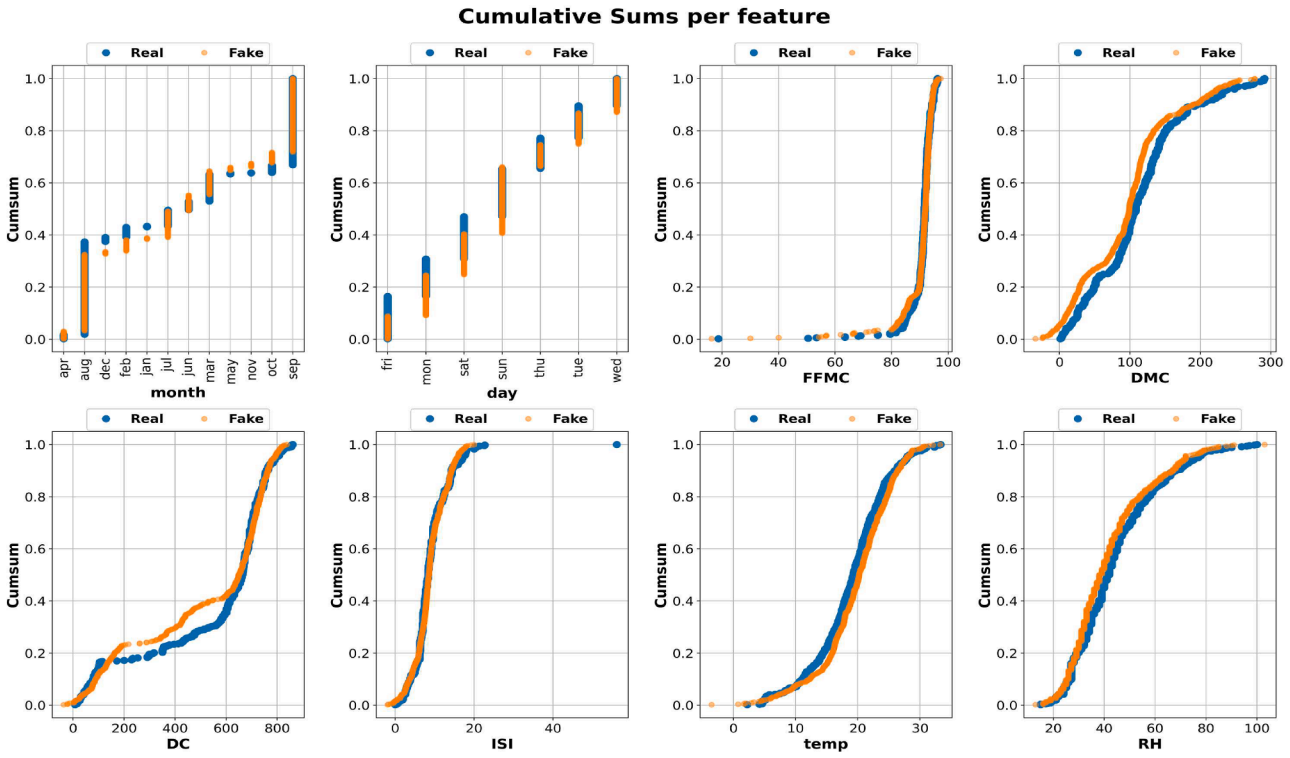

Fig. 7. Comparison of Cumulative Sum per Feature of Original and Synthetic Data.

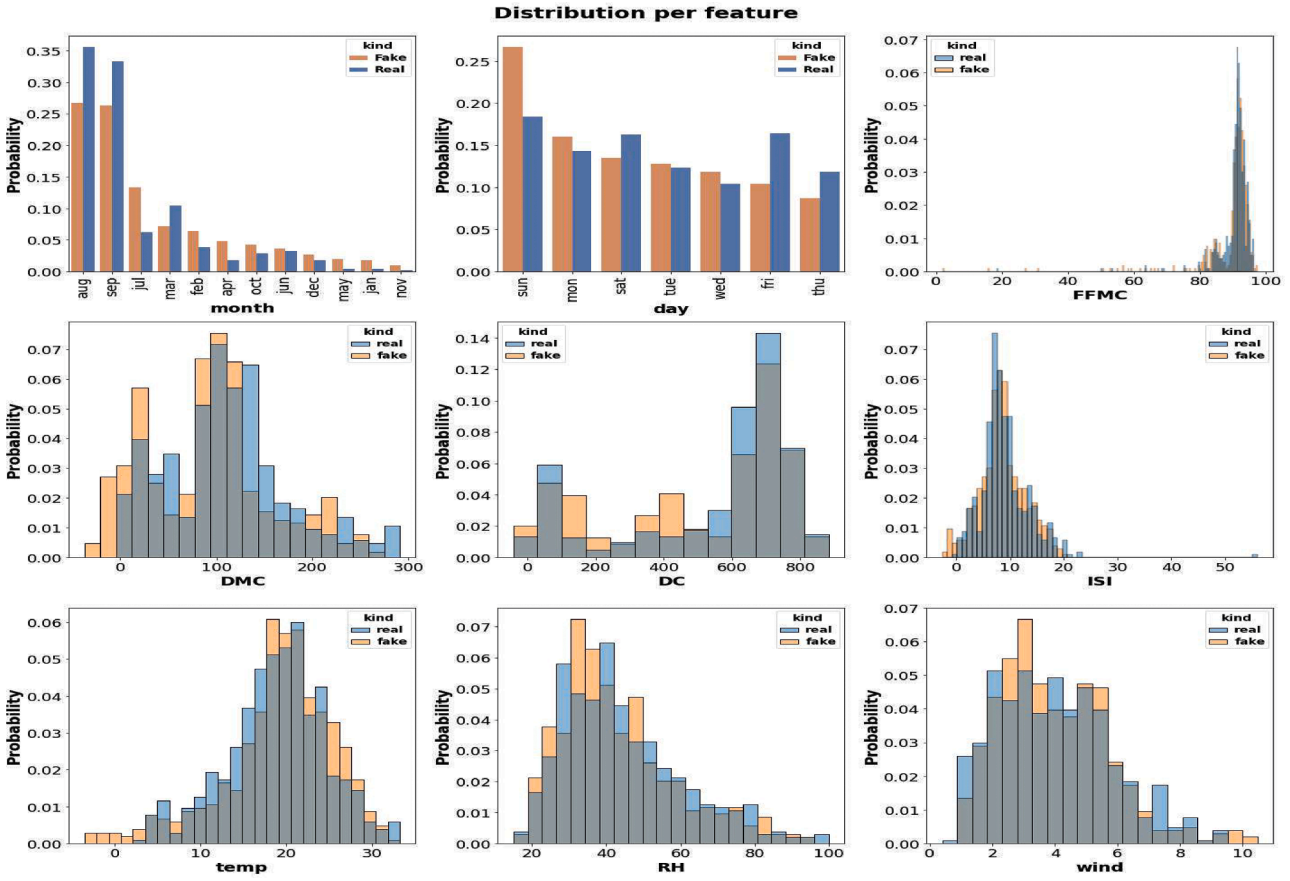

Fig. 8. Data Distribution & Kernel Density Estimation of Original and Synthetic Data.

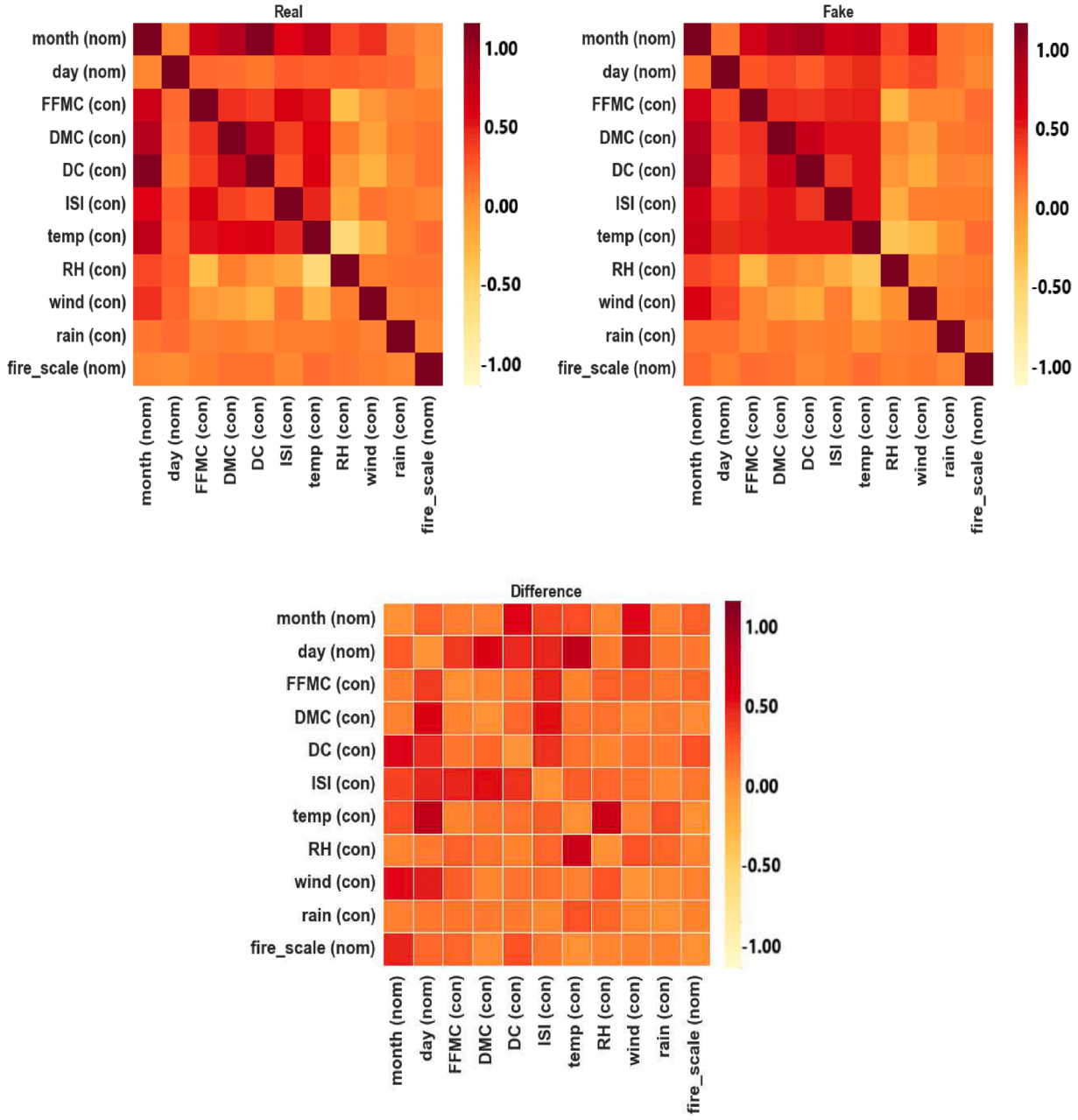

Fig. 9. Correlation plot based on Pearson correlation of original and synthetic data.

hances the generalization abilities of the classification models by meticulously searching for the optimal set of parameters that yield the best predictive performance. The use of *GridSearchCV* allowed for a comprehensive evaluation of various hyperparameter combinations, ensuring that each model configuration was rigorously assessed against the forest fire dataset. Optimization through *GridSearchCV* effectively improves the predictive metrics. Random Forest, Extra Tree, LightGBM, and CatBoost, show very high performance, generally achieving the highest across all metrics. This suggests that ensemble methods and boosting algorithms are particularly effective for this dataset. While SVM [27] and ANN [28] are still not good as compared to other models. In the given context, CatBoost being a more generalized model, outperforms all other models not only in accuracy (95.08 %) but also having the highest and nearer values to the accuracy for precision (93.03 %), recall (92.80 %), and f1-score (92.91 %) as well (Fig. 12).

Selecting *CatBoost* as the classification model, the Table 5 compares the performance of different data generation techniques, including the proposed approach, against traditional data generation methods such as BN [19], CTGAN [16], W-GAN + PacGAN [11], SMOTE [6], Borderline-SMOTE [7], and ADASYN [22]. The metrics used for comparison are *accuracy*, *precision*, *recall*, and *F1-Score*. Reviewing the Table 5, it becomes clear that the proposed approach is able to significantly outperform the traditional data generation methods. A descriptive study is presented below:

- **Accuracy:** The proposed approach achieves an accuracy of 95.08 %, which is significantly higher compared to CTGAN alone (92.35 %), and far exceeds the accuracy of traditional methods like SMOTE, Borderline-SMOTE, ADASYN, and the standalone BN.
- **Precision:** The analysis of the findings reveal that the precision of the proposed method is as high as 93.03 %. Compared to the CTGAN,

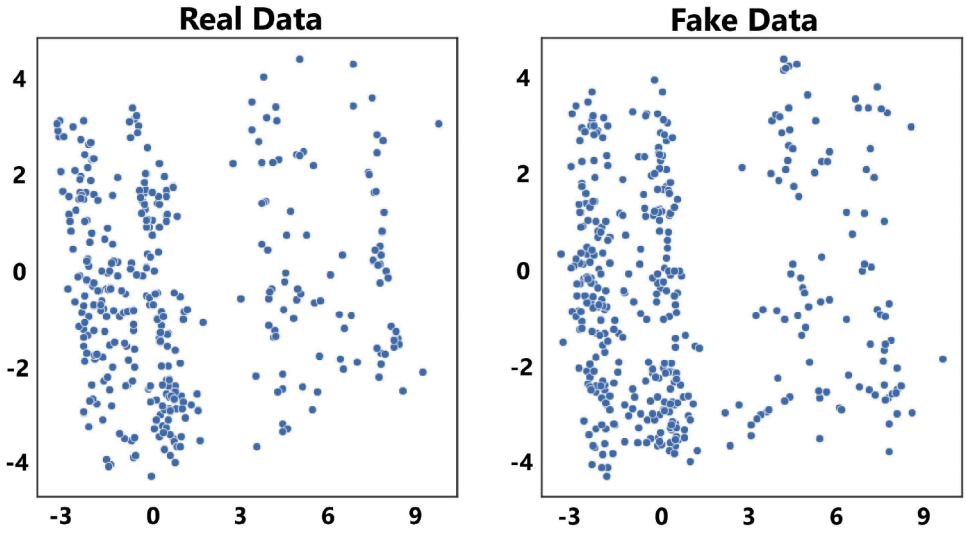

Fig. 10. Comparison of First Two Components of Original and Synthetic Data.

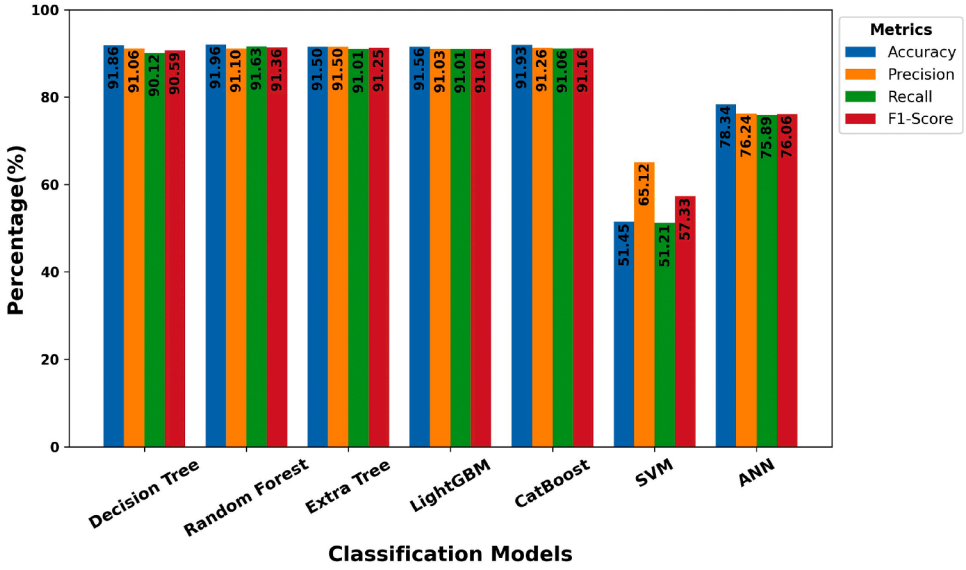

Fig. 11. Classification Model Performance Comparison with the Proposed Approach on Forest Fire Data.

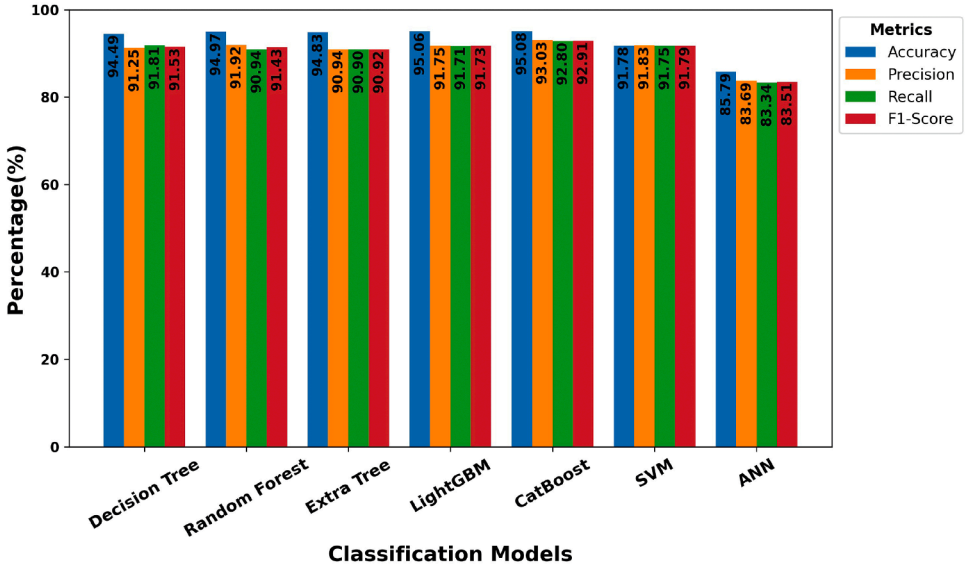

Fig. 12. Classification Model Performance Comparison with Proposed Approach on Forest Fire Data Using GridSearchCV.

Table 5
Model Evaluation with the Proposed Approach and SOTA Techniques on Forest Fire Data.

Data Generation Algorithms	Metrics			
Classification	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Proposed Approach	**95.08**	**0.9303**	**0.9280**	**0.9291**
W-GAN + PacGAN [11]	91.25	0.8820	0.8630	0.8723
CTGAN [16]	92.35	0.9098	0.8878	0.8987
BN [19]	76.48	0.7465	0.7568	0.7516
SMOTE [6]	54.89	0.5459	0.5450	0.5454
Borderline-SMOTE [7]	53.37	0.5213	0.5221	0.5217
ADASYN [22]	53.98	0.5302	0.5345	0.5323

which is at (90.98 %), and other state-of-the-art approaches SMOTE, Borderline-SMOTE, ADASYN, and BN, the proposed method is able to significantly outperform them with a clear margin of approximately 3.8 %. As obious from the results, the high precision indicates the performance supremacy of the proposed method.

- **Recall:** As verified through the results, the ability of the proposed method to achieve a high recall of 92.80 %, demonstrates its effectiveness in identifying all relevant instances of the target class without missing many actual positives. When compared to the other methods, identified for this analysis, the improvement percentage is more than 4 % over the compared approaches which are at a distant 88.78 % (CTGAN), 54.50 % (SMOTE), 52.21 % (Borderline-SMOTE), 53.45 % (ADASYN), and 75.68 % (BN). This results clearly highlights the methods ability to accurately identify the important events, in this case - fire, with significantly less number of missed instances.
- **F1-Score:** The proposed approach is able to achieve a maximum F1-Score of 92.91 %, highlits its ability achieving balanced levels of precision and recall. As compared to CTGAN, which reports a F1-score of 89.87 %, the proposed method is able to surpass the performance of SMOTE, Borderline-SMOTE, ADASYN, and BN.

6.4. Extended validation

6.4.1. FOUS Dataset [31] and california weather and fire prediction dataset [32]

To validate the model's performance, an extended analysis was conducted on the FOUS Dataset [31] and the California Weather and Fire Prediction Dataset [32]. The results establish the fact with more certainty that the proposed method is able to effectively handle the class imbalance issue in highly imbalanced datasets for wildfire prediction. The outcomes are presented in Table 7 and the observations clearly suggest that the proposed method achieves consistently high performance across all evaluation metrics, with precision, F1-Score, and accuracy values exceeding 0.98 for both datasets. The observations, as reported in Table 7, clearly highlight that the proposed method achieves a high precision of 0.9907, an F1-Score of 0.9888, and an accuracy of 0.9806. The results clearly indicate that the proposed hybrid method is not only able to minimize the false positives, but is also able to maintain a balanced performance with precision and recall. The slightly lower accuracy, relative to the other metrics, reflects the inherent complexity of large fire growth data and the variability introduced by extreme weather conditions.

6.4.2. Distributional validation using Kolmogorov-Smirnov statistic

To check how well the synthetic data resemble the real observations, we used the Kolmogorov-Smirnov (KS) test. This statistic looks at the largest difference between the cumulative distribution functions (CDFs) of the two datasets and provides a direct way to see how far the generated samples deviate from the original data. The KS distance is calculated as:

$$D_{\mathrm{KS}} = \sup_{x} \left| F_{\mathrm{real}}(x) - F_{\mathrm{fake}}(x) \right| \tag{13}$$

where, $F_{\mathrm{real}}(x)$ and $F_{\mathrm{fake}}(x)$ refer to the empirical CDFs of the real and synthetic data, respectively. A smaller value of D_{KS} means that the two distributions are more alike, and therefore the model has managed to capture the underlying structure of the data more accurately.

Table 6
Mean KS distance comparison between real and synthetic data distributions. Lower values indicate better distributional alignment.

Method	Mean KS Distance ↓
Proposed Approach	**0.081**
TabPFGen [36]	0.213

The comparison in Table 6 shows the mean KS distances for the proposed BN-guided CTGAN and the Transformer-based TabPFGen [36]. The proposed model achieves consistently lower KS values across all features, suggesting that it maintains better diversity and captures relationships among variables more effectively, even when the data are limited. A careful analysis of the obtained results prove that TabPFGen tends to produce overly smooth feature densities with less variation, which means some finer details of the data are lost. On the other hand, the proposed BN-guided CTGAN retains sharper and more realistic feature patterns, staying closer to how the actual data behave. This outcome indicates that transformer-based models like TabPFGen may struggle with small or imbalanced datasets, while the Bayesian conditioning in the proposed model helps stabilize learning and improves the realism of generated samples.

6.4.3. Additional parameters for performance validation

In addition to aggregate metrics reported in Tables 5 and 7, we further investigate the performance of the proposed method by calculating the per-class confusion matrices, precision-recall curves, and cost trade-offs. It can be observed in Fig. 13, that the per-class behavior due to the imbalanced distribution of fire events is well handled by the proposed approach. The results on the three target classes: *no_fire*, *small_fire*, and *large_fire*, show that the proposed approach achieves high recognition rates across all classes, with significant improvements in correctly identifying the rare but operationally critical *large_fire* cases compared to baseline methods. Furthermore, the precision-recall curve reported in Fig. 13, shows the ability of the proposed approach to maintain high precision even as recall increases, demonstrating resilience to class imbalance.

6.4.4. Computational efficiency and training cost analysis

1. *Complexity Comparison*
   Traditional oversampling algorithms, such as SMOTE and ADASYN achieve low computational time by performing direct feature-space interpolation without any learned parameters. Their complexity

Table 7
Extended Validation and Performance Comparison on FOUS Dataset [31] and Fire Prediction Dataset [32].

Method ↓	FOUS Dataset [31]			Fire Prediction Dataset [32]		
Parameter →	Precision	F1-Score	Accuracy	Precision	F1-Score	Accuracy
Proposed Approach	0.9907	0.9888	0.9806	0.9990	0.9992	0.9993

Table 8
Comparative Computational Cost vs. Performance Trade-off.

Method	Complexity	Training Time (m̃s)	KS	F1-Score
SMOTE [6]	$O(N \cdot d)$	2000	0.18	0.79
Borderline-SMOTE [7]	$O(N \cdot d)$	2200	0.19	0.79
ADASYN [22]	$O(N \cdot d)$	3700	0.19	0.78
BN [19]	$O(N \cdot d \cdot k)$	10,700	0.14	0.91
W-GAN + PacGAN [11]	$O(E \cdot N \cdot d)$	90,000	0.14	0.90
CTGAN [16]	$O(E \cdot N \cdot d)$	120,000	0.10	0.96
TabPFGen [36]	$O(L \cdot H \cdot d^2 \cdot T)$	450,000	0.21	0.63
Proposed Approach	$\mathbf{O(E \cdot N \cdot d + N \cdot k)}$	**150000**	**0.08**	**0.98**

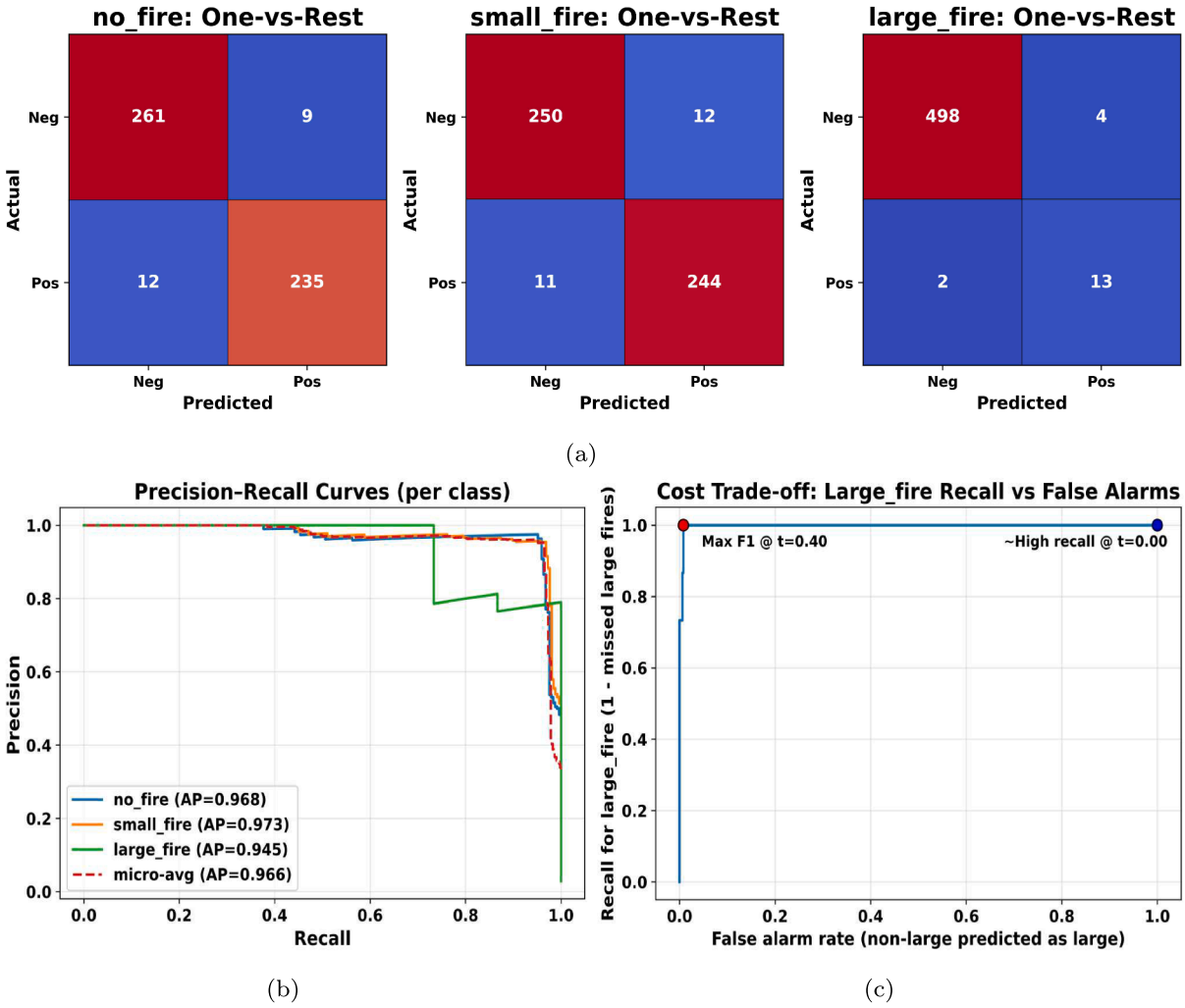

Fig. 13. Consolidated Evaluation: (a) Per-class Confusion Matrices, (b) Per-class Precision-Recall Curve, and (c) Cost Trade-offs.

scales linearly with the dataset size and feature dimension, $O(N \cdot d)$, but they fail to preserve inter-feature dependencies and often produce unrealistic synthetic samples. In contrast, deep generative models introduce a higher one-time training cost in exchange for better distributional realism.

The asymptotic complexities listed in Table 8 are expressed in terms of key parameters that influence computational cost. Here, N denotes the number of training samples and d the feature dimension of the dataset. The term E represents the number of training epochs used for adversarial optimization in GAN-based methods, while k corresponds to the average number of parent dependencies per node

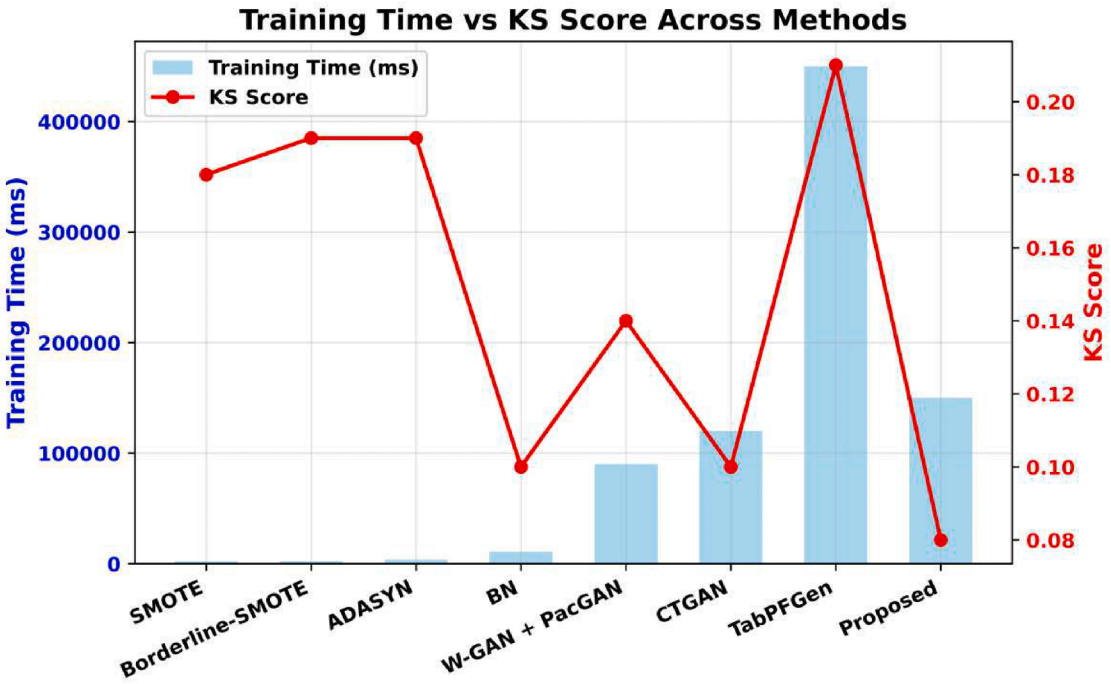

Fig. 14. Training Time vs KS Score across Methods.

in the Bayesian Network. For transformer-diffusion models such as TabPFGen, L indicates the number of stacked attention layers, H the number of self-attention heads, and T the number of diffusion steps required to reconstruct data distributions. Bayesian and adversarial models introduce additional training iterations or dependency estimation, increasing cost to $O(N \cdot d \cdot k)$ or $O(E \cdot N \cdot d)$. The proposed BN-guided CTGAN combines these components efficiently, requiring $O(E \cdot N \cdot d + N \cdot k)$ - a moderate increase over classical methods but substantially lower than the transformer-based TabPFGen whose complexity grows quadratically with d and scales linearly with both L and T.

2. *Computational Comparison*

While the proposed method incurs a modest increase in training time relative to traditional oversampling, this cost is offset by significant improvements in data fidelity and classification performance. The Bayesian conditioning effectively constrains the generator's search space, reducing adversarial instability and improving convergence efficiency. Once trained, the model can generate high-quality synthetic samples in seconds, making inference-time cost negligible.

Furthermore, for environmental datasets where data acquisition is costly and retraining occurs infrequently, the additional training effort is justified by the improvement in minority-class representation and reduced distributional divergence. In operational terms, the cost of slightly higher computation is substantially lower than the cost of performance degradation from simpler but statistically naive methods such as SMOTE variants or ADASYN. As such, the proposed approach provides a balanced trade-off between computational efficiency and data quality, remaining lightweight enough for deployment while achieving the highest distributional fidelity and predictive accuracy among evaluated techniques.

6.4.5. *Cross-domain robustness - credit card fraud dataset [30]*

To further validate the cross-domain robustness of the proposed model, the performance of the proposed approach is tested on another highly skewed dataset, the Credit Card Fraud Dataset [30]. This extended validation is aimed at determining the consistency of the proposed approach across domains where class imbalance is a significant challenge. (See A; these results do not affect the fire-domain conclusions.)

7. Conclusion

The proposed hybrid framework introduces a significant method in tackling the challenges of data imbalance and recognition bias in ML methods for forest fire detection. Targeting the drawbacks of existing methods, who are unable to capture the complex dependencies and intricate patterns present in the data, this work aims to achieve high accuracy and mitigate bias in fore fire detection. As such, this work introduces a novel hybrid approach that explores complex probabilistic relationships among environmental factors, incorporating IoT-driven data, and using a GAN to synthetically augment minority classes. The proposed model is validated on publicly available datasets, and the performance is reported on evaluation metrics such as accuracy, precision, recall, F1-score, Computational Efficiency, Training Cost, and Distributional Validation using Kolmogorov–Smirnov Statistic. The results show that the proposed hybrid model is able to achieve a significant improvement over the exiting methods with a significant margin. Validated on real world variables and real world deployment scenarios, the proposed method is able to maintain its performance across domains characterized by severe class imbalance. Beyond forestry, the adaptability and scalability of the proposed method make it well-suited to domains such as healthcare diagnostics, fraud detection, and security analytics, where rare yet critical events must be accurately detected in highly imbalanced settings.

CRediT authorship contribution statement

Vishal Krishna Singh: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis, Conceptualization; **Deepshikha Agarwal:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Investigation, Conceptualization; **Vivek Kumar Gediya:** Soft-

ware, Methodology, Investigation, Formal analysis; **Rajkumar Singh Rathore:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Formal analysis, Data curation; **Weiwei Jiang:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Formal analysis, Data curation.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Cross-domain robustness: Credit card fraud

A.1. Dataset and unified protocol

To stress-test class-imbalance behavior outside the wildfire domain, we evaluate our method on a public credit card fraud dataset [30]. We follow the same unified, leakage-safe protocol used for wildfire datasets: (i) preprocessing fitted only on training data, (ii) class-imbalance handled via the method under test, (iii) metrics include macro-F1, per-class PR-AUC, confusion matrices, and a cost trade-off (recall of the minority class vs. false alarms).

A.2. Results and interpretation

This extended validation is aimed at determining the consistency of the proposed approach across domains where class imbalance is a significant challenge. The Table A.1 showcases the performance evaluation of the *CatBoost* model using various data generation algorithms on the credit card fraud dataset, similar to the evaluation previously conducted with the forest fire dataset in Table 5. It compares the performance of the proposed approach against other SOTA techniques for data generation in terms of *Accuracy*, *Precision*, *Recall*, and *F1-score*. It is clear from the comparative analysis that the proposed approach significantly outperforms all the other techniques, achieving the highest score across all metrics, and accurately identifying fraudulent transactions. CTGAN alone also performs well, particularly in precision and F1-score, but does not surpass the proposed approach in any of the metrics.

Table A.1
Cross-domain Robustness through CatBoost Classification on Credit Card Fraud Data.

Data Generation Algorithms	Metrics			
Classification	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Proposed Approach	**98.72**	**0.9813**	**0.9796**	**0.9804**
W-GAN + PacGAN [11]	93.50	0.9070	0.8920	0.8995
TabPFGen [36]	63.00	0.6357	0.6312	0.6334
CTGAN [16]	97.26	0.9692	0.9647	0.9670
BN [19]	91.92	0.9158	0.9113	0.9136
SMOTE [6]	78.83	0.7825	0.7811	0.7818
Borderline-SMOTE [7]	79.08	0.7899	0.7824	0.7861
ADASYN [22]	78.96	0.7862	0.7817	0.7840

Appendix B. Practical deployment considerations

The practical deployment of the proposed method warrants few considerations, such as:

1. **Model Retraining Frequency:** The proposed method is able to downsize the long-term maintenance and computational costs by using the incremental learning process, achieved through the adaptive augmentation.
2. **Hardware Flexibility:** Considering the average computation requirements of the proposed approach, the method can be easily deployed on standard computing machines and edge devices.
3. **Latency and Real-Time Applicability:** One of the major achievements of the proposed method is it's ability to minimize the end-to-end delay due to the reduced sample redundancy and faster convergence characteristics. This makes it suitable for real/almost real time applications of data processing and inference.
4. **Scalability and Integration:** Asynchronous execution of augmentation and retraining processes, combined with simple light weight architecture of the proposed method allows it to be highly scalable.
5. **Robustness:** The ability to handle data drift and strong noise in wildly adverse conditions of a forest, makes the proposed method highly robust ensuring consistent performance in diverse scenarios.

References

[1] V.R.D. Dios, R.H. Nolan, Some challenges for forest fire risk predictions in the 21st century. Forests, 12, 2021
[2] L. Zhang, C. Lu, H. Xu, A. Chen, L. Li, G. Zhou, MMFNet: forest fire smoke detection using multiscale convergence coordinated pyramid network with mixed attention and fast-robust NMS, IEEE Internet Things J. 10 (20) (2025) 18168–18180.
[3] P. Cortez, A. Morais, A data mining approach to predict forest fires using meteorological data, in: J. Neves, M.F. Santos, J. Machado (Eds.), New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007-Portuguese Conference on Artificial Intelligence, Guimarães, Portugal, 2007, pp. 512–523.
[4] Y.O. Sayad, H. Mousannif, H.A. Moatassime, Predictive modeling of wildfires: a new dataset and machine learning approach, Fire Saf. J. 104 (2019) 130–146.
[5] M.B. Joseph, M.W. Rossi, N.P. Mietkiewicz, A.L. Mahood, M.E. Cattau, L.A.S. Denis, R.C. Nagy, V. Iglesias, J.T. Abatzoglou, J.K. Balch, Spatiotemporal prediction of wildfire size extremes with bayesian finite sample maxima, Ecol. Appl. 29 (2019) 1898.
[6] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[7] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-SMOTE: a new oversampling method in imbalanced data sets learning, in: Proc. Int. Conf. Intell. Comput, Berlin, Germany, Springer, 2005, pp. 878–887.
[8] C. Lai, S. Zeng, W. Guo, X. Liu, Y. Li, B. Liao, Forest fire prediction with imbalanced data using a deep neural network method, Forests 13, 2022.
[9] T. Preeti, S. Kanakaraddi, A. Beelagi, S. Malagi, A. Sudi, Forest fire prediction using machine learning techniques, in: 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1–6.
[10] L. Zinan, K. Ashish, F. Giulia, O. Sewoong, Pacgan: The power of two samples in generative adversarial networks, Advances in neural information processing systems, 31, 2018
[11] W. Shafqat, Y.C. Byun, A hybrid GAN-based approach to solve imbalanced data problem in recommendation systems, IEEE Access 10 (2022) 11036–11047.
[12] R. She, P. Fan, From MIM-Based GAN to anomaly detection: event probability influence on generative adversarial networks, IEEE Internet Things J. 9 (19) (2025) 18589–18606.
[13] J.H. Lee, K.H. Park, GAN-based imbalanced data intrusion detection system, Pers. Ubiquitous Comput 25 (2019) 121–128.
[14] J. Kim, K. Jeong, H. Choi, K. Seo, Gan-based anomaly detection in imbalance problems, in: Proc. Eur. Conf. Comput. Vis, Cham, Switzerland, Springer, 2020, pp. 128–145.
[15] I. Siniosoglou, P. Radoglou-Grammatikis, G. Efstathopoulos, P. Fouliras, P. Sarigiannidis, A unified deep learning anomaly detection and classification approach for smart grid environments, IEEE Trans. Netw. Serv. Manag. 18 (2) (2021) 1137–1151.
[16] S.S. Mullick, S. Datta, S. Das, Generative adversarial minority oversampling, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1695–1704.
[17] P. Cortez, A. Morais, Forest fires. UCI Machine Learning Repository, 2018. https://doi.org/10.24432/C5D88D
[18] V.K. Singh, C. Singh, H. Raza, Event classification and intensity discrimination for forest fire inference with IoT, IEEE Sens. J. 22 (9) (2022) 8869–8880. https://doi.org/10.1109/JSEN.2022.3163155
[19] E. Kyrimi, S. Mclachlan, K. Dube, M.R. Neves, A. Fahmi, N. Fenton, A comprehensive scoping review of Bayesian networks in healthcare: past, present and future, arxiv:2002.08627, 2020.
[20] T.A. Le, A.G. Baydin, R. Zinkov, F. Wood, Using synthetic data to train neural networks is model-based reasoning, in: 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 3514–3521.
[21] A. Koivu, M. Sairanen, A. Airola, T. Pahikkala, Synthetic minority oversampling of vital statistics data with generative adversarial networks, J. Amer. Med. Inform. Assoc. 27 (11) (2020) 1667–1674.
[22] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural

Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 1322–1328.
[23] L. Breiman, Random forests, Mach. Learn 45 (1) (2001) 5–32.
[24] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn 63 (1) (2006) 3–42.
[25] G. Ke, et al., LightGBM: a highly efficient gradient boosting decision tree, in: Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 3146–3154.
[26] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, in: Proc. Adv. Neural Inf. Process. Syst., null, 2018, pp. 6638–6648.
[27] I. Steinwart, A. Christmann, Support Vector Machines, Springer, New York, NY, USA, 2008.
[28] E. Grossi, M. Buscema, 'Introduction to artificial neural networks,", Eur. J. Gastroenterol. Hepatol. 19 (2007) 1046–1054.
[29] G.N. Ahmad, H. Fatima, S. Ullah, A. Salah, I. Saidi, Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV, IEEE Access 10 (2022) 80151–80173.
[30] P. Gupta, A. Varshney, M.R. Khan, R. Ahmed, M. Shuaib, S. Alam, Unbalanced credit card fraud detection data: a machine learning-oriented comparative study of balancing techniques, Proc. Comput. Sci 218 (2023) 2575–2584.
[31] B.E. Potter, Mcevoy, J. Daniel, Fire growth and associated weather data for selected fires of unusual size (FOUS) and other fires from 2004 to 2018, in: Forest Service Research Data Archive, Fort Collins, CO, 2022. https://doi.org/10.2737/RDS-2022-0040
[32] C.E. Yavas, C. Kadlec, J. Kim, L. Chen, California Weather and Fire Prediction Dataset, 1984. With Engineered Features [Data set, https://doi.org/10.5281/zenodo.14712845
[33] Rapid Damage Assessment, Copernicus Emergency Management Service, 2025. https://forest-fire.emergency.copernicus.eu/about-effis/technical-background/rapid-damage-assessment.
[34] Current Wildfire Situation in Europe, European Commission, Joint Research Centre, 2025. https://joint-research-centre.ec.europa.eu/projects-and-activities/natural-and-man-made-hazards/fires.
[35] Instituto da Conservação da Natureza e das Florestas (ICNF), 'Relatório de Incêndios Rurais em Portugal, 2025. https://www.icnf.pt/api/file/doc/504914cdd1a211bb.
[36] J. Ma, et al., TabPFGen-Tabular Data Generation with TabPFN, /arXiv:2406.05216, 2024.

Vishal Krishna Singh received his bachelor's degree in Information Technology in 2010, the master's degree in Computer Technology and Application in 2013, and the PhD degree in Information Technology from the Indian Institute of Information Technology, Allahabad, India, in 2018. He is currently a Lecturer and is associated with the Networks and Communications Research Group at the School of Computer Science and Electronics Engineering, University of Essex, Colchester, U.K. His research interests include the Internet of Things, wireless sensor networks, in-network inference, machine learning, and data analytics.

Deepshikha Agarwal is currently the Head of the Department of Information Technology at IIIT Lucknow. She completed her Ph.D. from MNNIT Allahabad in 2015 and her M.Tech from IIIT Allahabad in 2007. She has over 19 years of experience in teaching, research, and industry. She is a member of professional bodies including IEEE, IET, and Oxford Journals. She has authored numerous research papers, books, and chapters and actively serves as a reviewer, editor, mentor, and speaker. She holds three international patents and has received the Swami Vivekanand Changemaker Award and the Women Eduvisionary Award in 2021.

Vivekkumar D. Gediya received his Bachelor's degree in 2021 from GEC Gandhinagar and his Master's degree in 2024 from IIIT Lucknow. He currently works as a Machine Learning Engineer at Advanced Micro Devices (AMD). His research interests include machine learning, artificial intelligence, data science, and deep learning.

Rajkumar Singh Rathore (Senior Member, IEEE) is the Head of Cybersecurity for Connected and Autonomous Systems at CINC, and of Cyber Physical and Networked Systems at CeRISS. He is also Programme Director for MSc Computing and IT at Cardiff Metropolitan University, UK. He holds a PhD, dual master's degrees, and a bachelor's degree in Computer Science and Engineering. He is a Fellow of the HEA UK. His research has been supported by Nottingham Trent University and Manchester Metropolitan University. His interests include wireless communications, cyber-physical systems, cybersecurity, CAVs, drone networking, and AI/ML applications. He is a founding member of IEEE TRUST-IoT.

Weiwei Jiang received the BSc and PhD degrees in Electronic Engineering and Information and Communication Engineering, respectively, from Tsinghua University, Beijing, China, in 2013 and 2018. He is currently an Assistant Professor in the School of Information and Communication Engineering at the Beijing University of Posts and Telecommunications. His research interests include artificial intelligence, machine learning, big data, wireless communication, and edge computing. He has published over 60 academic papers, with more than 4700 citations on Google Scholar, and is listed in Stanford's World's Top 2 % Scientists (2023, 2024 and 2025).