Received: 21 April 2025 | Revised: 30 June 2025 | Accepted: 5 August 2025 | Published online: 13 October 2025

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2025, Vol. 00(00) 1-12
DOI: 10.47852/bonviewJCCE52025973

BON VIEW PUBLISHING

An Extreme Gradient Boosting Feature Selection–Based GAN-ELM for Classification of Imbalanced Big Data

Rithani Mohan[1] **and Prasanna Kumar Rangarajan**[1,*]

[1] *Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, India*

Abstract: The high-volume, velocity, and variety nature of Big Data introduces tremendous difficulties in accurate classification, especially when dealing with class imbalance. Traditional computational methods often fail to handle the imbalanced nature of datasets that may result in predictions biasing toward majority classes and poor model accuracies. This work presents a novel classification framework that combines state-of-the-art methodologies to provide a solution for the classic problem of class imbalance in Big Data environments. The proposed method is based on Generative Adversarial Network with Extreme Learning Machine for classification, Extreme Gradient Boosting with Bayesian Hyperparameter Optimization for feature selection, the Coati Optimization Algorithm for gradient optimization, and Fuzzy Adaptive SMOTE for oversampling. In addition to this, we put a Physics-Informed Policy Gradient to achieve interpretability of the model and classification decisions with respect to the domain classification rules. This framework provides better performance in terms of accuracy, robustness and scalability than other approaches for various types of imbalanced medical imaging datasets, such as histopathological images. The collaborative use of these state-of-the-art algorithms, taking into consideration common challenges like noisy data, redundant samples, and overfitting, will result in improved classification and provides a feasible solution for Big Data problems.

Keywords: Big Data, oversampling, feature selection, classification, vanishing gradients

1. Introduction

The steep growth of data (Internet of Things, eHealth, bioinformatics, social networks, and many others) resulted in the Big Data phenomenon, generating an enormous amount of structured and unstructured data at massive volumes and high velocity with structural heterogeneity [1]. Conventional data processing methods fail to process such huge and complex datasets, especially when labeled classification is aimed at high-dimensional or unstructured data. One of the biggest problems in Big Data classification is class imbalance because the minority class instances in such data are much less than majority class examples [2] leading to biased classification and less expressive generalization. In domains such as early disease diagnosis, fraud detection, and cybersecurity, the minority class often represents critical but rare events (presence of cancer, fraudulent transactions, or malware activity). Misclassifying these minority cases can lead to dire outcomes like delayed treatment, financial losses, or security breaches. Hence, accurate classification of minority instances in imbalanced datasets is not merely a statistical improvement but a functional necessity for high-impact decision-making. Techniques such as oversampling class imbalance have been widely accepted within the community to address class imbalance issues (Synthetic Minority Oversampling Technique [SMOTE]) [3]. In generating synthetic samples by interpolation, SMOTE has flaws that include noisy data, redundant samples, and class overlap. Better variants to overcome the imbalance of data have been proposed such as enhanced Center Point SMOTE (CP-SMOTE) and Inner–Outer SMOTE (IO-SMOTE). In CP-SMOTE, synthetic samples are generated around multiple centroids of the minority class to better preserve distribution diversity [4], while IO-SMOTE [5] differentiates between inner and outer minority samples to minimize overlap, noise, and related issues. Nevertheless, the variants are confronted with computational complexity and sample selection problems.

In Big Data classification, feature selection is essential for reducing computational complexity and improving performance. Extreme Gradient Boosting (XGB) works great for finding nonlinear features which are discriminative [6] but very expensive to hyperparameter tune manually. Incorporating Bayesian Hyperparameter Optimization (BHPO) to tune learning rates, tree depth, and regularization parameters by itself automates the process of XGB selection and boosts efficiency [7]. Deep learning (DL)–based methods are immensely powerful but often suffer from the vanishing gradient problem, which triggers suboptimal convergence in large-scale classification tasks. We propose rescaling the gradient using the Coati Optimization Algorithm(COA), which balances exploration and exploitation, thereby stabilizing weight updates and improving the convergence rate [8]. This is advantageous for training deep networks in imbalanced Big Data scenarios. A hybrid framework fusing Generative Adversarial Networks (GANs) with Extreme Learning Machines (ELMs) has proved effective for classification improvement under severe imbalance [9]. GANs yield good synthetic samples that help in class balancing, and ELMs have the advantage of fast learning and a single-layer structure that makes classification scalable as well as more efficient [10]. Moreover, the incorporation of Physics-Informed Policy Gradient Networks ensures that the model predictions adhere to domain constraints, thereby enhancing interpret ability and robustness [11].

This study introduces a novel integrated framework for Big Data classification that incorporates Fuzzy Adaptive SMOTE(FADA-SMOTE) for dynamic oversampling, Extreme Gradient Boosting with

***Corresponding author:** Prasanna Kumar Rangarajan, Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, India. Email: r_prasannakumar@ch.amrita.edu

Bayesian Hyperparameter Optimization (XGB-BHPO) for feature selection [7], COA [8] for optimization, and a robust Generative Adversarial Network with Extreme Learning Machine (GAN-ELM) for classification [9]. The framework was evaluated on three imbalanced datasets, primarily histopathology images, demonstrating high scalability, improved classification accuracy, and efficient convergence in Big Data environments.

Contribution. In this article, the Big Data classification approach within a DL framework uses oversampling, feature selection, and vanishing gradient problems. The tackling mechanism is proposed for time series Big Data.

- Development of FADA-SMOTE to overcome the limitations of existing oversampling techniques by generating statistically consistent synthetic samples while preventing noise and class overlap.
- Integration of XGB-BHPO for feature selection, ensuring optimal feature selection while reducing dimensionality and computational complexity.
- Implementation of GAN-ELM for robust classification, employing adversarial learning to enhance model generalization and classification accuracy.
- Optimization via COA to address the vanishing gradient problem, stabilizing DL training for Big Data classification.
- Incorporation of a Physics-Informed Policy Gradient Network, embedding domain knowledge into the classification process for improved interpretability and real-world applicability.

The article is organized as follows: The related works are reviewed in Section 2, followed by the proposed method in Section 3. Experimental details are given in Section 4. Results are discussed in Section 5, and finally, the conclusion is drawn in Section 6.

2. Literature Review

With the fast-developing field of information technology, intensive attention is given to intelligent medical data processes. However, imbalanced classification remains a major challenge, particularly in medical diagnostics, where class imbalance can lead to biased learning algorithms. In order to tackle the class imbalance issue, Vairetti et al. [12] work on SMOTENN [2] which is a hybrid resampling that uses SMOTE [1] and ENN [3] for more minority classes by using neighborhood structures to preserve the class margin. While effective, this approach struggles with high-dimensional data due to the generation of noisy synthetic samples. Similarly, Halim et al. [13] used ADASYN: adaptively generates synthetic minority classes in an epoch based on their learning difficulty. Although widely adopted, both ADASYN and SMOTE tend to create ambiguous samples near class boundaries, reducing classification performance. In order to reinforce the samples, Maldonado et al. [14] extended SMOTE into weighted Minkowski distance–based SMOTE and enhanced neighborhood selection but it is still affected by noisy features. To mitigate the generation of unwanted unsafe synthetic instance, Grina et al. [15] recommended SMOTE-BFT that uses an application of the Belief Function Theory for removing noisy samples effectively. Its dependence on belief thresholds hinders adaptability. Bao and Yang [16] reported the CP-SMOTE method as well as IO-SMOTE in relation to the enhancement of central reference points and minority sample separation. They show that these approaches curb noise but significantly fail in complex, high-dimensional data as a result of computational burden.

Apart from sampling, feature selection of imbalanced data is extremely important. Most traditional methods put the majority class on the lead. To solve this, Kamalov et al. [17] presented an F1-score filter using decision tree focusing on the relevance of the minority class. However, tree-based heuristics may ignore possible low-level inter-feature dependencies. Li et al. [18] presented multi-strategy grouped feature extraction, where redundancy is eliminated efficiently. For selection of dimension, Zhang et al. [19] took MOFS-BDE, a binary differential evolution [20] which is based on the novel mutation operators for balancing exploration and exploitation. Although effective, this method may struggle in high-noise environments.

DL in breast cancer detection of imbalanced data causes performance limitation. To address this limitation, Kumari et al. [20] introduced Smart GAN for improving classifier training on limited data. But GANs tend to become gradient, vanishing when discriminators drown generators. Ding et al. [21] proposed RGAN-EL, which integrates GANs with ensemble learning to address class imbalance; however, this approach introduces significant training overhead due to the ensemble component. In a subsequent work, Ding et al. [22] developed RVGAN-TL, combining GANs with transfer learning for tabular data classification. While this method improves performance, its ability to transfer effectively across different datasets remains limited. Convolutional Neural Networks (CNN) have also shown to be very feasible in deep neural networks (DNNs) for histopathological image classification of the domain. Overfitting and vanishing gradients are still there. However, Thapa et al. [23] proposed a deep CNN model trained by DL for bypassing these problems and Chattopadhyay et al. [24] introduced a Residual Attention Network Dense Residual Dual-Shuffle Attention (DRDA) on Low Data which is a lightweight version of DenseNet. Ultimately, while both models have been successful, they still require a lot of fine tuning and computational power. Saini and Susan [25] employed transfer learning with DNN, data augmentation, and regularization to address class imbalance but transfer learning not always aligns with target distributions. Bas et al. [26] later aimed at making the training method as good as possible using PSO (Particle Swarm Optimization) in order to improve Recurrent Neural Network (RNN) performance. Hu et al. [27] developed artificial derivatives as alternative activation functions which provide functionality while overcoming gradient issues. Nevertheless, they are sometimes difficult to interpret even in the originally built models and lack generalizability.

Recent work shows that data quality in resource-constrained sensing also benefits from freshness and attention mechanisms. Yang et al. [28] demonstrate how jointly optimizing Age of Information and information entropy improves data accuracy and energy efficiency in wireless sensing systems. Complementing this, Yang et al. [29] apply attention mechanisms in mobile crowd sensing—across task allocation, incentive design, privacy preservation, and data transmission—to dynamically prioritize high-value, low-noise inputs. These insights into temporal and contextual data weighting further motivate our framework's emphasis on robust oversampling and feature selection. Moreover, online learning has focused on addressing challenges related to incomplete and imbalanced data streams. You et al. [30] introduced the Online Learning from Incomplete and Imbalanced Data Streams algorithm, which utilizes empirical risk minimization to select informative features and employs a dynamic cost-sensitive strategy to handle real-time class imbalance. Their evaluation across 27 datasets under various stream patterns (trapezoidal, evolvable, and capricious) and imbalance conditions demonstrated superior adaptability, particularly under concept drift. Extending this, You et al. [31] proposed Online Learning for Data Streams with Incomplete Features and Labels (OLIFL), a framework designed to handle data streams with simultaneous feature and label incompleteness. OLIFL dynamically maintains a global informative matrix and estimates label confidence to minimize the influence of unlabeled samples, enabling accurate learning in volatile environments.

While significant progress has been made in addressing imbalance learning, feature selection, and vanishing gradients, critical challenges remain. Oversampling can still introduce noise, feature selection in high-dimensional imbalanced settings is not consistently optimal, and methods to prevent vanishing gradients can reduce

efficiency or scalability. In response to these limitations, the proposed framework leverages advanced oversampling techniques, ensemble-based feature selection, and gradient-stabilizing architectures to deliver high classification accuracy with improved efficiency and adaptability, particularly on real-world healthcare datasets.

3. Proposed Methodology

Figure 1 depicts the end-to-end Big Data classification pipeline, which consists of five sequential stages. First, the raw imbalanced dataset is rebalanced using FADA-SMOTE, a method that computes a fuzzy membership score for each minority instance and generates synthetic samples adaptively, placing more samples in sparse regions and fewer in dense regions to reduce noise and class overlap. Second, the balanced data undergo feature selection with XGB-BHPO, where XGB ranks features by importance and BHPO automatically tunes parameters, removing redundant or noisy attributes and lowering dimensionality. Third, before training the deep model, the COA rescales weight updates dynamically to prevent vanishing gradients and ensure stable convergence. Fourth, the processed features are fed into the GAN-ELM classifier, in which the GAN component creates realistic minority samples through adversarial training and the Extreme Learning Machine trains rapidly using random input weights with a closed-form solution for output weights. Finally, a Physics-Informed Policy Gradient (PIPG) module adds domain-specific penalty terms to the GAN's reward function to enforce real-world constraints and produce interpretable outputs. These stages—oversampling, feature selection, gradient stabilization, classification, and interpretability—work together in a unified framework to handle class imbalance, noisy inputs, optimization challenges, and transparency.

3.1. Fuzzy Adaptive SMOTE

FADA-SMOTE is a new technique designed to address the shortcomings of traditional SMOTE and its derivatives which use class imbalance in large imbalanced datasets to their advantage. By linearly interpolating between the minority instance examples, SMOTE creates new samples; however, this approach has several drawbacks including noise class overlap and computational inefficiency. Although they are more sophisticated and (very) computationally expensive, techniques like CP-SMOTE and IO-SMOTE can result in overfitting. They aim to improve the quality of synthetic samples.

To address these issues, FADA-SMOTE combines MapReduce-distributed processing, fuzzy logic, and dynamic control over sample generation, making it scalable and accurate. To determine how salient an instance is for oversampling, the process begins by computing a fuzzy membership score for each minority instance x_i. This weight is quantified by the fuzzy membership function as an importance:

$$\mu(x_i) = \frac{1}{1+e^{-\beta d(x_i,K)}}$$

Here, $d(x_i, K)$ is the Euclidean distance from x_i to its K-nearest majority class neighbors; β denotes the scaling parameter which adjusts sensitivity; $\mu(x_i)$ defines if the sample is vital for oversampling, with values closer to 1 indicating higher importance; and also, membership threshold θ is set to filter out low-significance instances, reducing the risk of introducing noise.

Once the high-significance instances are identified, FADA-SMOTE employs an adaptive sample generation strategy. It is determined using how many synthetic samples must be made for each selected instance x_i:

$$S(x_i) = \alpha \cdot \frac{N_{\text{maj}}}{N_{\text{min}}} \cdot \mu(x_i)$$

Here, N_{maj} and N_{min} refer to the number of majority and minority instances, respectively, and α is a weight factor. In sparse regions, the adaptive sample generation mechanism boosts synthetic generation while preventing oversampling in dense minority regions.

A method of fuzzy-weighted interpolation is employed to generate artificial samples. For every x_i, a nearby minority instance x_j is randomly selected and a new synthetic sample x_{new} is created using the process outlined below:

$$x_{\text{new}} = x_i + \lambda(x_j - x_i)$$

Here, x_j refers to an arbitrarily elected fuzzy nearest minority neighbor and λ denotes a dynamic interpolation coefficient:

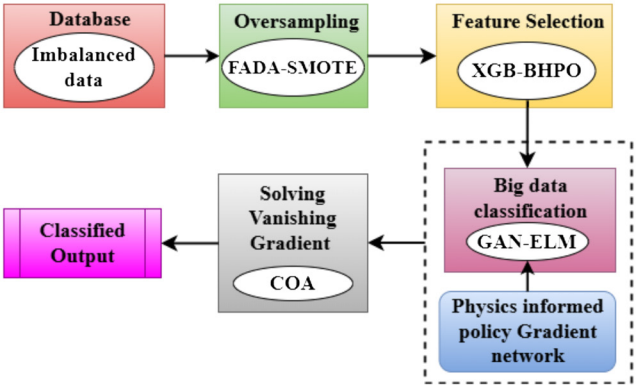
$$\lambda = \mu(x_i) \cdot \delta$$

Here, δ is a value chosen at random from a uniform distribution in the interval [0, 1].

This minimizes class overlap by ensuring synthetic samples that stay inside clearly defined decision boundaries.

Also, FADA-SMOTE is integrated as a distributed execution method into the MapReduce framework so that it could scale with large datasets. Each minority instance determines the fuzzy memberships and counts of the corresponding necessary synthetic samples after the dataset has been partitioned during the Mapper phase. To create a balanced dataset, several synthetic samples from each partition are combined during the Reducer phase. Even when processing high-dimensional large-scale data, the FADA-SMOTE–parallelized architecture retains its effectiveness and performance.

Figure 1
Block diagram of the Proposed GAN-ELM via Physics-Informed Policy Gradient Network



3.2. Feature selection using ensemble learning

A hybrid feature selection approach based on XGB and BHPO is proposed to enhance classification efficiency and mitigate overfitting in high-dimensional, imbalanced data. The study introduces a novel hybrid feature selection approach that combines XGB and BHPO to address the problem of overfitting in high-dimensional imbalanced data and improve classification efficiency. The three conventional filter wrapper and embedded approaches all have drawbacks such as missing feature interactions, being model dependent, and being computationally inefficient. The suggested method makes use of XGB ensemble learning which is reliably stable and BHPO due to its capacity for probabilistic optimization in order to address these issues.

3.2.1. XGB for feature selection

XGB is a regularized gradient-boosting framework that iteratively assesses and prunes features according to three key metrics: Cover (sample coverage), Weight (split frequency), and Gain (loss reduction). In subsequent iterations, low-importance features are eliminated minimizing dimensionality and redundancy while maintaining predictive relevance.

3.2.2. Bayesian hyperparameter optimization for XGB

BHPO will be used to fine tune critical XGB hyperparameters (learning rate depth and estimators) on the fly using Gaussian process–based surrogate modeling. It continuously explores its surroundings to find solutions that lead to model generalization and better fit test data taking advantage of similarly promising points in order to find the ideal configuration in the bounded space. The classification accuracy gain and scalability in mass datasets using the XGB-BHPO framework were made possible by a condensed and interpretable feature subset.

3.3. Classification using GAN-ELM

ELMs and GANs are integrated into the Dynamic GAN-ELM framework to address classification problems in Big Data settings particularly for their use in high-dimensional noisy and imbalanced dataset classification problems. This hybrid architecture as shown in Figure 2 includes a discriminator (B) and a generator (C) as multichannel signal inputs into the GAN assembly. Based on inverse dynamics, the generator creates positive samples and the discriminator evaluates the quality of both generated and real samples. A crucial enhancement to this architecture superstructure is the use of a Physics-Informed Policy which ensures the legal expansion of practically possible synthetic information within constraints. The multilayer ELM classifier will use these improved samples following GAN refinement of the data. ELMs are a type of single-pass neural network in which the back propagation iteration is eliminated by randomly initializing the input weights and using Moore–Penrose pseudo-inversion to calculate the output weights in a closed-form solution. With minimal loss of accuracy, this saves a significant amount of time during the training session. Together, GAN-based augmentation and ELM rapid learning enabled classification, making it possible to handle noise redundancy and imbalanced data.

3.3.1. Generative Adversarial Network

GAN comprises of generator (C) and discriminator (B) networks which are trained concurrently in an adversarial environment. The work of the generator is to generate the samples from the data which resemble the real data. On the other hand, the discriminator finds the difference between the real and synthetic samples. The process of generating data samples is explained in Figure 3 [32]. This training process can be defined by a minimax optimization problem:

$$\min_{C}\max_{B} Y(B,C) = A_{t\sim r_{\text{data}}(t)}[\log B(t)] + A_{s\sim r_s(s)}[\log(1-B(C(s)))] \quad (1)$$

Figure 2

The architectural diagram of Dynamic GAN-ELM

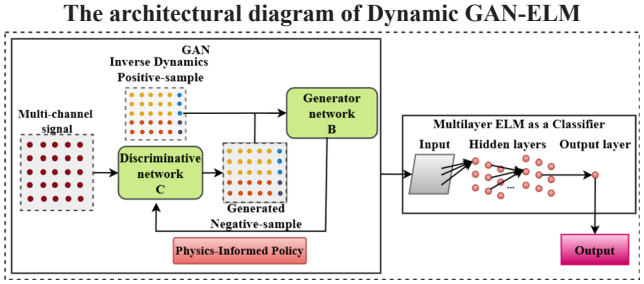
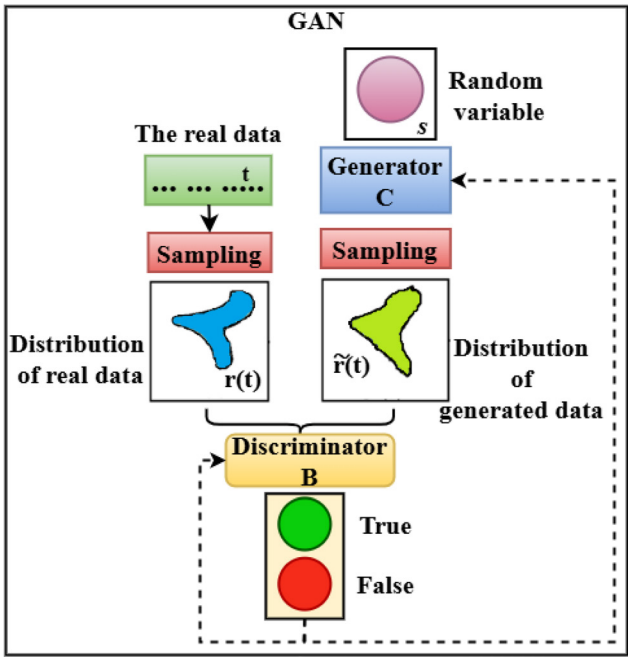


Figure 3

The complete process of generating data of GAN



Here, the distribution of the training set is denoted as $r_{\text{data}}(t)$, while r_s denotes distribution of noise. At equilibrium, the generator's distribution approximates the real data distribution $r_c = r_{\text{data}}$ making the discriminator output $B_c^*(t) = 1/2$, indicating indistinguishability between real and fake data [32].

3.3.2. Integration with ELM for feature generation

The data generation is based on an ELM model. ELM uses a self-adaptive multi-layer ELM model for relatively accurate classifiers. The ELM is a machine learning model based on a multilayer perceptron (MLP) architecture. The ELM does not rely on MLP as W and b are initialized randomly using respective uniform random distributions. Also, using a linear combination yields the prediction.

ELM-based data generation: In order to address the issue of under-representation in our minority class, we will use new ELM models dynamically with GANs to generate samples. The ELM samples will be derived using randomly segmented portions of the data fit with $\text{Set}_a\ \{(\mathbb{i}_{\mathbb{l}}, \mathbb{w}_{\mathbb{l}})\}_{\mathbb{l}=1}^{M_a}$, providing fragments W_a^* [33]. Using the Dynamic ELM model will use the samples to assess class imbalance in the generated data.

$$LP_a = M_{\mathbb{j}}^{\text{least}} / M_{\mathbb{j}}^{\text{most}} \quad (2)$$

$$\text{num}_{\text{class+}} = \lfloor M_{\mathbb{j}}^{\text{most}} / M_{\mathbb{j}}^{\text{class+}} \rfloor \quad (3)$$

Generated features W_C are obtained through

$$W_C = \mathcal{D}_C \gamma_C \quad (4)$$

where $\mathcal{D}_C$ is the output value from the hidden layer and γ_C the output weight matrix. The dynamic ELM can adaptively increase the number of hidden nodes in the generator, while simultaneously determining the structure of the output layer. Hidden layer I can be expressed as follows:

$$\mathcal{D}_C, I+1 = [\mathcal{D}_C, I_{\mathbb{h}}] \quad (5)$$

where h is a random variable representing a new hidden node output. Optimization is aimed at minimizing expected cross-entropy:

$$\min_{C}\min_{B}\mathcal{Y}(B,C)=\mathcal{A}_{W_{a\sim P_{data}(W_a)}}[\log B(W_a)]+\mathcal{A}_{T\sim P_T(T)}[\log(1-B(C(T)))] \tag{6}$$

Self-adaptive multilayer ELM modeling: The self-adaptive multilayer ELM automatically tunes the number of hidden layers and neurons to optimize model performance while balancing time complexity and generalization. Too few nodes result in underfitting, while too many increase computational overhead. The selection is guided by the data distribution and imbalance ratio, enabling dynamic reconfiguration during training [32]. The hidden layer size is computed based on sample complexity and feature richness:

$$S=\text{function}(n_{\text{samples}}, n_{\text{features}}, \mathbb{LP}) \tag{7}$$

where $\mathbb{LP}$ is the imbalance ratio. This configuration allows for adaptive learning that supports the evolving complexity of input data.

3.4. Training of GAN-ELM through Physics-Informed Policy Gradient

By using a PIPG within the GAN-ELM training framework, we introduce a domain-aware reinforcement mechanism to assist with the generative process. Unlike traditional adversarial learning which minimizes statistical differences only, PIPG embeds domain-based constraints as inductive biases in the reward function; therefore, generated outputs are required not only to fool the discriminator but to be consistent with the physics equations. This is important for any generalization especially when data are scarce [34]. Beyond enforcing physical consistency, PIPG also yields a constraint compliance profile for each generated sample (and thus each classification) by reporting the residual penalties associated with each rule. In medical imaging, for example, these rules can encode tissue stiffness ranges or contrast uptake kinetics; the model then explains its decisions in terms of which rules were most strongly satisfied or violated. This traceability gives clinicians actionable insights into the physical properties driving each prediction, enhances trust by linking outputs to established domain knowledge, and supports regulatory compliance by providing clear, rule-based justifications.

The reward function $\mathcal{M}(\varsigma)$ is composed of two parts: the structural incentive $\mathcal{P}_{C_\varsigma}$ and the physical representation action $\mathcal{S}^{C(\varsigma)}_{B(\varphi)}$, expressed as follows:

$$\mathcal{M}(\varsigma)=\mathbb{A}[\mathcal{P}_{C_\varsigma}(C_\varsigma(s_{0:\mathcal{O}}))]\cdot\mathcal{S}^{C(\varsigma)}_{B(\varphi)}\ (C_\varsigma(s_{0:\mathcal{O}})),[\mathcal{B},\Theta]_{0:\mathcal{O}}) \tag{8}$$

where $s_{0:\mathcal{O}}$ denotes the multichannel signal sequence. C_ς will produce unbiased samples $([\hat{\mathcal{B}},\hat{\Theta}]_{0:\mathcal{O}})$ consistent with physical dynamics. The structural reward is defined by

$$\mathcal{P}_C([\hat{\mathcal{B}},\hat{\Theta}]_{0:\mathcal{O}})=\exp^{\mathcal{RI}^2}([\hat{\mathcal{B}},\hat{\Theta}]_{0:\mathcal{O}}) \tag{9}$$

where $\mathcal{RI}(.)$ is a residual informing the Lagrange motion equation

$$\mathcal{RI}([\hat{\mathcal{B}},\hat{\Theta}]_{0:\mathcal{O}})=\tfrac{1}{\mathcal{O}}\textstyle\sum_{\circledcirc=1}^{\mathcal{O}}(\mathbb{J}(\hat{\Theta}_\circledcirc)\ddot{\hat{\Theta}}_\circledcirc+\hat{\mathbb{g}}(\Theta_\circledcirc,\dot{\hat{\Theta}}_\circledcirc)+\hat{\mathbb{C}}(\Theta_\circledcirc)-\textstyle\sum_{\Bbbk}\mathcal{K}\mathcal{B}_\circledcirc^{\Bbbk})^2 \tag{10}$$

This enforces alignment to system dynamics with minimal data. We sequentially obtain the policy gradient as follows:

$$\delta_\varsigma\mathcal{M}(\varsigma)\simeq\tfrac{1}{\mathcal{O}}\textstyle\sum_{\circledcirc=1}^{\mathcal{O}}\sum_{\mathbb{v}_\mathcal{O}\in[\hat{\mathcal{B}},\hat{\Theta}]_\mathcal{O}}\delta_\varsigma\mathcal{P}_{C_\varsigma}(\mathbb{v}_\circledcirc|[\mathcal{B},\Theta]_\mathcal{O})\cdot\mathcal{S}^{C(\varsigma)}_{B(\varphi)}(\mathbb{v}_\circledcirc,[\mathcal{B},\Theta]_{0:\mathcal{O}}) \tag{11}$$

The policy is then updated using the following:

$$\varsigma\leftarrow\varsigma+\beta\delta_\varsigma C(\varsigma) \tag{12}$$

where the learning rate is $\beta\in P$ [34].

3.5. COA for solving the vanishing gradient problem

The COA minimizes disappearing gradients and allows for effective multi-objective optimization with little parameter tuning. The advantages of the COA include having no control parameters needed, strong global search, fast convergence, and good performance on high-dimensional tasks.

Each coati represents a solution in the population and it is created as follows:

$$U_n: y_{n,m}=lb_m+p(ub_m-ub_m), n=q,w,\ldots,M, \quad m=1,2,\ldots,j \tag{13}$$

Here, $p\in[0,1]$ denotes an arbitrary number and lb_m and ub_m refer to the lower and upper bounds of the m-th variable, respectively. The COA employs two distinct behaviors: exploration and exploitation [35].

3.5.1. Exploration phase

Half of the population waits below (local search), and the other half climbs the tree (broad search). That for the climbing team is as follows:

$$U_n^{Q1}: y_{n,m}^{Q1}=y_{n,m}+p\cdot(\text{Iguana}_m+L\cdot y_{n,m}), \quad \text{for } n=1,2,..,\left\lfloor\tfrac{M}{2}\right\rfloor \text{ and } m=1,2,\ldots,j \tag{14}$$

where Iguana_m is the best-known position and $L\in\{1,2\}$ is a random integer. That for the remaining coati is as follows:

$$\text{Iguana}^T: \text{Iguana}_m^T=lb_m+p\cdot(ub_m-lb_m), m=1,2,\ldots,j, \tag{15}$$

$$U_n^{Q1}: y_{n,m}^{Q1}=\begin{cases}y_{n,m}+p\cdot\left(\text{Iguana}_m^T-L\cdot y_{n,m}\right), V_{\text{Iguana}^{T<V_n}}\\ y_{n,m}+p\cdot\left(y_{n,m}-\text{Iguana}_m^T\right), \text{ else}\end{cases}, \quad \text{for } n=\left\lfloor\tfrac{M}{2}\right\rfloor+1,\left\lfloor\tfrac{M}{2}\right\rfloor+2,\ldots,M \text{ and } m=1,2,\ldots,j \tag{16}$$

Updated positions are retained only if they improve the objective function:

$$U_n=\begin{cases}U_n^{Q1}, V_n^{Q1}<V_n\\ U_n, \text{ else}\end{cases} \tag{17}$$

Here, the new position calculated for n^{th} is U_n^{Q1}; U_n^{Q1} is the m^{th} dimension; the objective value function is represented as V_n^{Q1}; the random real number in the interval [0,1] is indicated as p; Iguana denotes the iguana's location within the search space, representing the position of its finest member; Iguana_m is its m^{th} dimension; Lis an integer which is directly selected from the set {1,2}; the position of the iguana on the ground generated randomly is represented as Iguana^T; and its dimension is Iguana_m^T; the value of this objective function is V_{Iguana^T}, and the floor function is signified as $\lfloor\cdot\rfloor$ [35].

3.5.2. Exploitation phase

This stage focuses on local refinement while simulating coati behavior in avoiding predators. This defines a local search space.

$$lb_m^{\text{local}} = \frac{lb_m}{h}, ub_m^{\text{local}} = \frac{lb_m}{h}, h = 1, 2, \ldots . N \tag{18}$$

Each coati searches near its current position:

$$U_n^{Q2} : y_{n,m}^{Q2} + (1 - 2p).\left(lb_m^{\text{local}} + p.\left(ub_m^{\text{local}} - lb_m^{\text{local}}\right)\right), \quad n = 1, 2, \ldots, M,\ m = 1, 2, \ldots, j \tag{19}$$

The new position is retained only if it yields better objective values:

$$U_n = \begin{cases} U_n^{Q2}, V_n^{Q2} < V_n \\ U_n, \text{ else} \end{cases} \tag{20}$$

4. Experimental Setup

A PIPG network combined with a GAN-ELM was used to classify unbalanced Big Data using images from breast cancer histopathology. Performance analysis was expanded to include ImageNet-LT and the CelebA Facial Attribute Dataset. Every model was put into practice in Python and compared with other approaches. A strong evaluation scenario is provided by the dataset class imbalance. A summary of the main features of the datasets used is provided in Table 1.

4.1. Hyperparameter settings

To maximize performance and guarantee stability during training, the suggested hyperparameters were carefully chosen. Extensive testing and validation on the dataset of Breast Cancer Histopathology images led to the determination of these values. Table 2 provides specifics about the selected hyperparameters and their values.

Table 2
Hyperparameters of the XGB learning model

Method	Parameter	Optimal value
XGB	No. of estimators	300
	Learning rate	0.01
	max depth	15
	reg_lambda	2
	objective function	Binary-logistic
	gamma	1
	booster	gbtree
	reg_alpha	1
GAN-ELM	Generator time	1.1
	Bot filter quantile	0.0001
	Top filter quantile	0.99
	Maximum depth	2
	Maximum bin	100
	Learning rate	0.001
	Random state	Yes
	Estimators	100
	Batch size	500
	Patience	25
	Amplification factor	0.7
	Population size	30
	Number of generators	50
	Lower bounds	(1,0,0)
	Upper bounds	(300,9,1)
COA	Parameter free	--

5. Results and Discussion

Key metrics like accuracy, precision, recall, specificity, F1-score, MCC, AUC, and computational time are used to assess the suggested GAN-ELM with a policy gradient network. Performance is evaluated in comparison to the DCNN, RGAN-EL, RVGAN-TL, and Smart GAN models. ImageNet-LT and CelebA are used for additional comparisons after the initial assessment of Breast Cancer Histopathology images. Findings demonstrate the model performance on unbalanced Big Data.

Figure 4 presents the 2D distribution of samples generated by No-Sampling, SMOTE, IO-SMOTE, CP-SMOTE, and the Proposed FADA-SMOTE on the Breast Cancer Histopathology dataset. Majority, minority, and synthetic samples are marked as blue circles, red pluses, and black snowflakes, respectively. While SMOTE-based methods reduce imbalance, they often generate overlapping or excessive synthetic points. FADA-SMOTE outperforms by adaptively generating fewer, high-quality samples away from decision boundaries. As shown in Table 3, FADA-SMOTE achieves the highest classification accuracy (99.81% training, 99.37% testing), improving generalization and minimizing error propagation.

Figure 5 compares error convergence across four oversampling techniques—No sampling, SMOTE, IO-SMOTE, and CP-SMOTE—on the Breast Cancer Histopathology images dataset. All methods start with high error due to imbalance, but FADA-SMOTE achieves the lowest final error with smoother and faster convergence. While SMOTE and IO-SMOTE offer moderate improvements, they suffer from instability and sensitivity to class boundaries. CP-SMOTE performs better, but FADA-SMOTE stands out by generating noise-resistant, well-separated synthetic samples, leading to superior model learning and error reduction.

After feature selection, classifiers such as SVM, LGBM, ETC, FKNN, and the suggested XGB-BHPO are compared on the Breast Cancer Histopathology images dataset in Table 4. Metrics like accuracy, F1-score, recall, specificity, MCC, and AUC distinguish XGB-BHPO from the competition. Its combination of BHPO and XGB allows for

Table 1
Research gaps

Dataset	Number of samples	Classes	Size	Number of attributes	Data availability
Breast Cancer Histopathology images	277,524	2	50 × 50	4	Publicly available
CelebA Facial Attribute Dataset	202,599	10,177	178 × 218	40	Publicly available
ImageNet-LT (Long-Tailed ImageNet)	115,846	1,000	Variable	N/A	Publicly available

Figure 4
Distribution of two-dimensional discrete point models using oversampling techniques for the Breast Cancer Histopathology images dataset

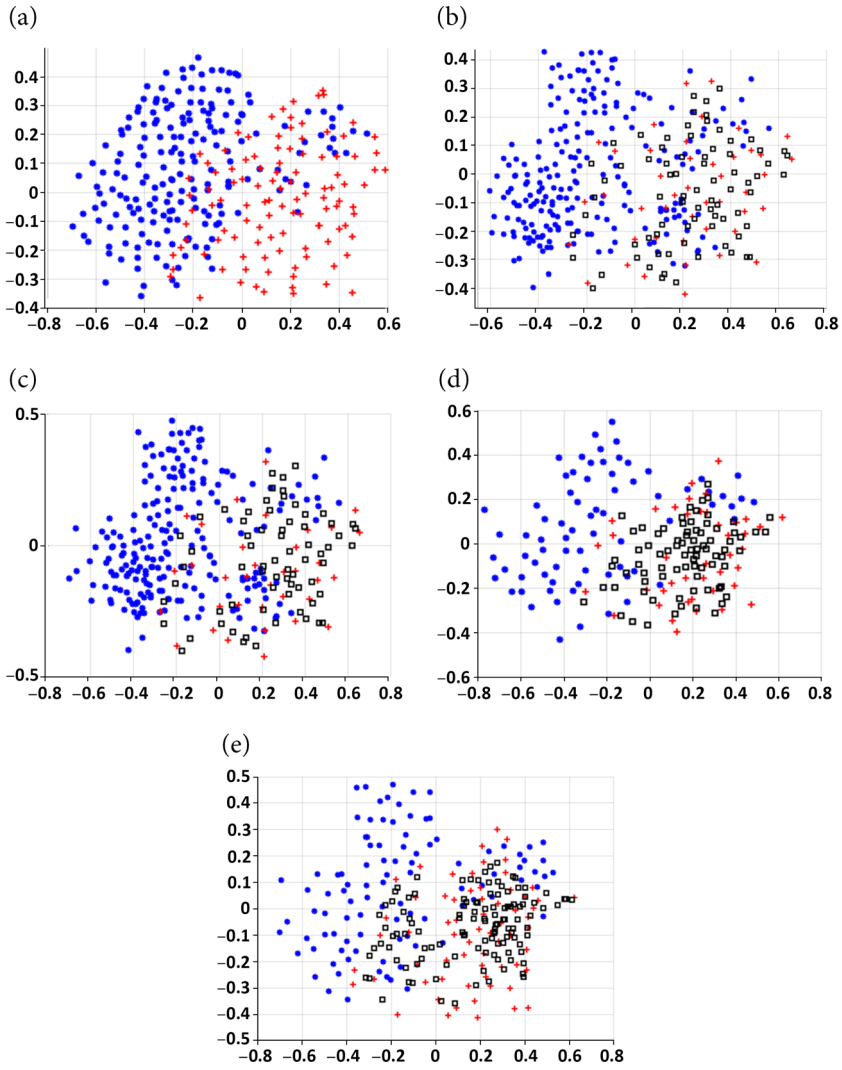


Table 3
Classification accuracy results for four different oversampling techniques in ELM with Breast Cancer Histopathology images dataset

Methods	Models	Accuracy Values (%)
No sampling	Training	99.36%
	Test	98.64%
SMOTE	Training	99.40%
	Test	98.80%
IO-SMOTE	Training	99.50%
	Test	99.00%
CP-SMOTE	Training	99.60%
	Test	99.20%
FADA-SMOTE	Training	99.81%
	Test	99.37%

Figure 5
Comparison of oversampling approaches for error function with Breast Cancer Histopathology images dataset

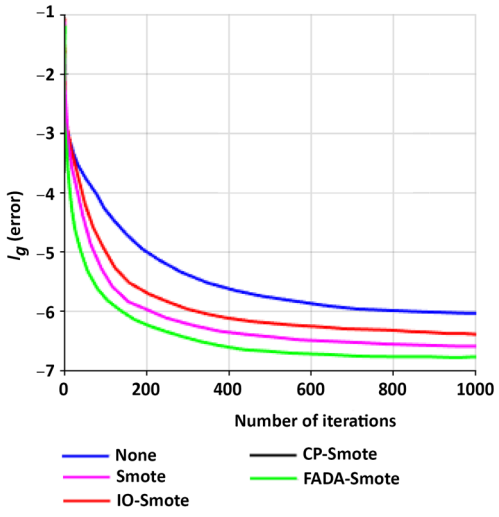


Table 4
Performance analysis for feature selection with Breast Cancer Histopathology images

Classifiers	SVM	LGBM	ETC	FKNN	Proposed XGB-BHPO
Accuracy	90.17	92.84	94.73	96.38	99.89
F1-score	89.57	90.83	92.69	94.82	98.01
Recall	89.46	91.65	94.48	95.38	97.75
Specificity	91.81	91.65	93.45	96.23	97.74
MCC	0.8493	0.8675	0.8814	0.8951	0.9184
AUC	0.9651	0.9704	0.9786	0.9729	0.9961

Figure 6
Comparative result for the feature selection model with the Breast Cancer Histopathology images dataset

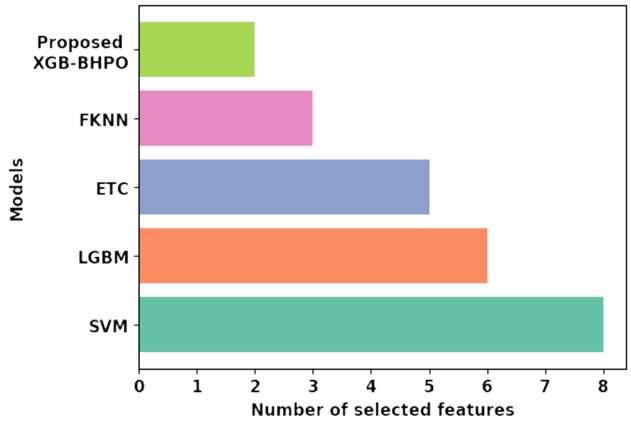


Table 5
Comparison of baseline models for feature selection approach

Reference	Method	Accuracy	Recall	Specificity
[36]	LightGBM	86.59	81.7	90.66
[36]	CatBoost	87.68	83.3	91.33
[37]	IG+PCA	98.24	-	-
[37]	CFS-BA	99.81	-	-
Ours	Proposed XGB-BHPO	**99.89**	**98.01**	**97.75**

efficient feature removal and ideal tuning. This improves robustness and discrimination by class particularly for unbalanced data. For medical diagnosis tasks, its high recall, specificity, MCC, and AUC attest to its dependability.

The feature selection effectiveness of SVM, FKNN, ETC, LGBM, and XGB-BHPO on the Breast Cancer Histopathology images dataset is contrasted in Figure 6. The fact that XGB-BHPO chooses the fewest features suggests that it is better at keeping only the most important characteristics. More features are retained by LGBM and SVM whereas FKNN and ETC exhibit a moderate reduction. For high-dimensional medical data, the results demonstrate how well XGB-BHPO reduces dimensionality while preserving classification performance and scalability.

Feature selection models applied to images of breast cancer histopathology are technically compared in Table 5. The suggested XGB-BHPO achieves the best accuracy recall and specificity surpassing baseline techniques. In contrast to CFS-BA and IG+PCA which do not provide comprehensive metric reporting, XGB-BHPO exhibits balanced and excellent performance. It provides better feature selection and classification reliability than LightGBM and CatBoost demonstrating its resilience for medical image analysis.

Figure 7 represents the classification outcome of the proposed model regarding accuracy, precision, recall, and MCC of different ML models or approaches. Five models are compared such as Proposed GAN-ELM, Smart-GAN, RGAN-EL, RVGAN-TL, and DCNN. Among the data balancing techniques, the proposed approach achieves the highest accuracy, approaching 100%. Smart-GAN slightly lowers accuracy than Proposed but is still very high. RGAN-EL has lower accuracy than Proposed and Smart-GAN but is still above mid-range. RVGAN-TL has accuracy around the mid-range and DCNN has very low accuracy among all listed models. The proposed approach shows the promising results in accuracy, recall, and MCC.

A comparison of time complexity across models for unbalanced Big Data is shown in Figure 8. Large-scale data processing benefits greatly from the suggested model's exceptional computational efficiency as input size grows. On the other hand, because they generate a lot of synthetic data, models like DCNN, Smart-GAN, RGAN-EL, and RVGAN-EL exhibit a much higher time complexity. The suggested method is perfect for real-time and resource-constrained situations since it strikes a balance between speed and performance.

Model space complexity for imbalanced Big Data classification is compared in Figure 9. For large datasets with limited storage, the suggested model is the most space-efficient option. On the other hand, because of their deep architectures and artificial data generation, RGAN-EL, RVGAN-TL, and DCNN demand more memory. The suggested method has little storage overhead and provides a scalable solution.

Table 6 provides a statistical comparison of classification methods based upon standard deviation, 95% confidence intervals (CI), and p-values. The proposed GAN-ELM has the highest classification accuracy (98.43%) and has the lowest variance (±0.15). This suggests that GAN-ELM has stable performance. The 95% CI (98.28–98.58) is both narrow and higher than other classification methods. All p-values are <0.05 suggesting statistically significant improvements (with respect to classification accuracy) over existing methods. The results tabulated here suggest that GAN-ELM is robust, reliable, or both.

The computational time of several optimization algorithms for solving the vanishing gradient problem is contrasted in Table 7. The suggested COA demonstrates superior efficiency in sustaining gradient flow by achieving the fastest convergence. With slower convergence, WOA, ACO, and ABC are less appropriate for deep networks that are susceptible to vanishing gradients whereas PSO and GWO likewise demonstrate strong performance.

Table 8 offers a complete modular ablation study explicitly examining the individual contributions of each key component of the proposed GAN-ELM model. The findings emphasize the sequential increases in classification performance gained by stacking sophisticated modules like FADA-SMOTE for generating synthetic samples, XGB-BHPO for selecting optimal features, COA for resolving gradient vanishing in deep layers, and PIPG for implementing domain-specific constraints via policy-guided learning. Every module makes the model more robust, improves generalizability, and increases accuracy in dealing with class imbalance. The combined configuration that incorporates all modules has the maximum performance, illustrating the synergistic contribution of the entire architecture.

Figure 7
Classification outcome of the proposed model with the Breast Cancer Histopathology images dataset

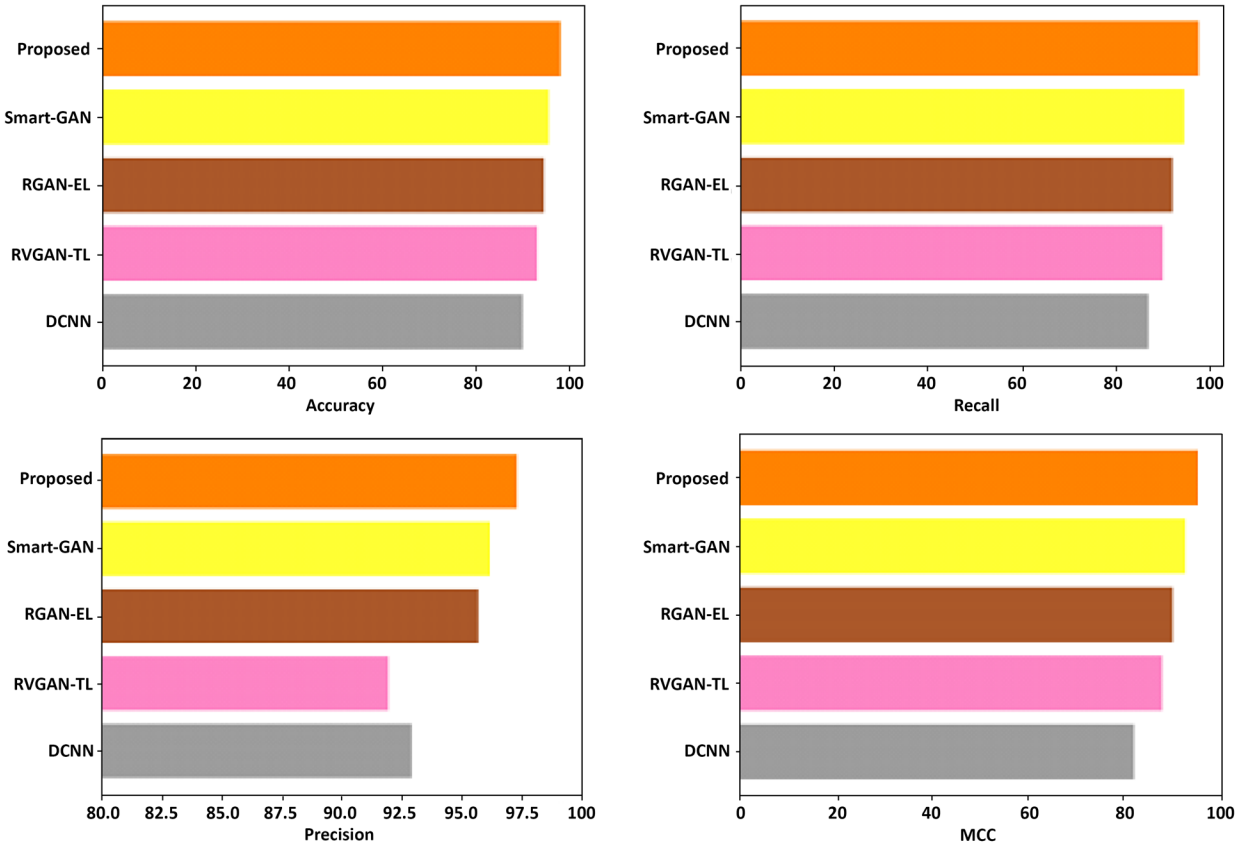


Figure 8
Comparison of time complexity analysis

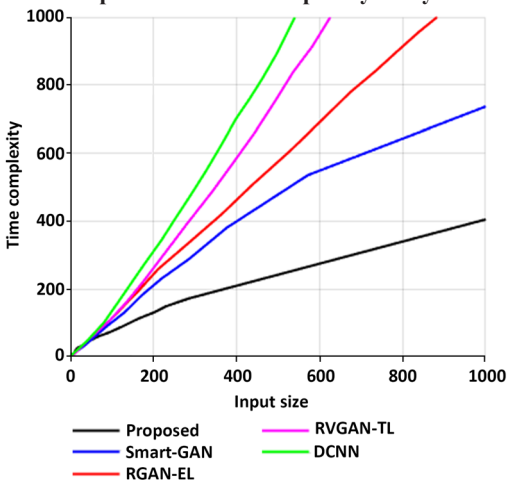
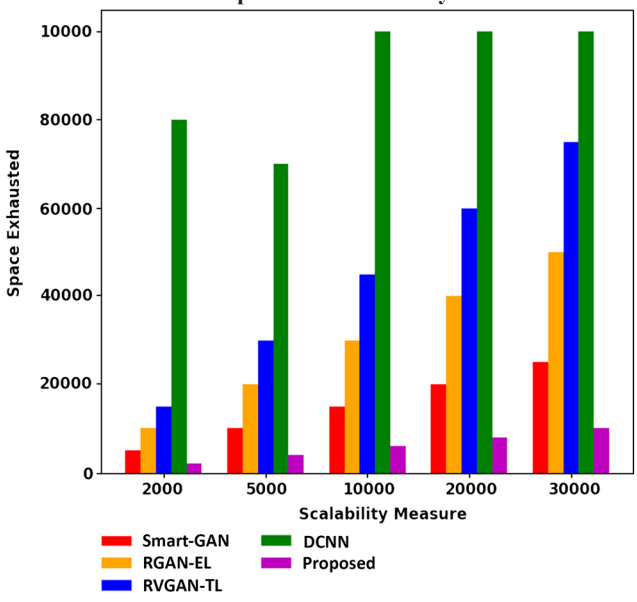


Figure 9
Comparison of scalability



In Table 9, the proposed GAN-ELM model is compared with other peers using three unbalanced datasets: ImageNet-LT, CelebA, and Breast Cancer Histopathology images. With regard to medical imaging data, the classification performance of the model is outstanding with high precision, recall, and F1-score, indicating successful identification of both the majority and minority classes. The model demonstrates good generalization ability on the CelebA dataset; however, ImageNet-LT performers are slightly lower, reflecting the challenges of long-tailed distributions. In this sense, GAN-ELM proves to be highly adaptable and efficient in tackling various issues of imbalanced image classification.

Table 6
Statistical evaluation of classification performance

Methods	Accuracy (%)	Std Dev (±)	95% CI	*p*-value (vs Proposed GAN-ELM)
nCOVnet [37]	88.00	±0.46	[87.52–88.48]	<0.001
Mini-COVIDNet [38]	83.20	±0.33	[82.85–83.55]	<0.001
Deep Neural Network [39]	95.81	±0.28	[95.50–96.12]	0.012
T-ResNet50+STGAN+TTA [40]	98.23	±0.22	[98.01–98.45]	0.041
Proposed GAN-ELM (Ours)	98.43	±0.15	[98.28–98.58]	—

Table 7
Comparison of computational time for different optimization algorithms

Methods	Computational time(s)
ABC	167.3
ACO	146.2
WOA	118.8
GWO	109.6
PSO	97.05
Proposed COA	76.41

Table 8
Ablation study for the proposed model

Method	Accuracy (%)
CWGAN	91.80
IT-GAN	94.20
STGAN	96.60
GAN-ELM (Baseline only)	96.90
GAN-ELM + FADA-SMOTE	97.40
GAN-ELM + FADA-SMOTE + XGB-BHPO	97.85
GAN-ELM + FADA-SMOTE + XGB-BHPO + COA	98.10
GAN-ELM + FADA-SMOTE + XGB-BHPO + PIPG	98.15
Proposed GAN-ELM (all modules)	98.43

Table 9
Comparison of performance metrics with different datasets using the proposed GAN-ELM

Dataset	Accuracy	Precision	Recall	F1-Score
Breast Cancer Histopathology images (ours)	98.43	98.01	98.10	97
CelebA Facial Attribute Dataset	97.36	97.21	97.6	96.42
ImageNet-LT (Long-Tailed ImageNet)	96.28	96.86	96.1	96

6. Conclusion

This research introduces a scalable classification strategy that addresses the challenges of imbalanced Big Data, specifically in the medical imaging field. The defined architecture incorporates several sophisticated elements which include FADA-SMOTE for oversampling, XGB-BHPO for feature selection, COA to mitigate the vanishing gradient problem, and a hybrid GAN-ELM classifier for enhanced learning. Moreover, the addition of a Physics-Informed Policy Gradient Network greatly improves the explicability of the system while ensuring adherence to domain-specific constraints. Testing on the Breast Cancer Histopathology images datasets verifies that FADA-SMOTE not only increases the informativeness of synthetic samples but also reduces redundancy compared with other oversampling techniques. XGB-BHPO achieved best results among all classical and ensemble feature selection methods by being the most selective and discriminative. The COA-optimized GAN-ELM classification module also provided better accuracy, recall, and robustness than the baseline classifiers which demonstrates the effectiveness of the proposed system. The results obtained validated the framework's efficacy using the Breast Cancer Histopathology images datasets where the system achieved 98.43% accuracy, the highest reported to date, which is a considerable improvement over conventional methods. This was due to an integration of advanced oversampling techniques, optimized feature selection, and robust classification methods. Further research will focus on the responsive adjustment to data shifts, decentralized data through federated learning, and integration of multi-modal medical datasets for enhanced classification generalizability.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The Breast Cancer Histopathology image data that support the findings of this study are openly available at https://www.kaggle.com/code/paultimothymooney/predict-idc-in-breast-cancer-histology-images/notebook. The CelebA data that support the findings of this study are openly available at https://www.kaggle.com/datasets/jessicali9530/celeba-dataset. The ImageNet-LT data that support the findings of this study are openly available at https://www.tensorflow.org/datasets/catalog/imagenet_lt.

Author Contribution Statement

Rithani Mohan: Conceptualization, Software, Validation, Resources, Data curation, Writing – original draft. **Prasanna Kumar Rangarajan:** Methodology, Formal analysis, Investigation, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Elkano, M., Galar, M., Sanz, J., & Bustince, H. (2018). CHI-PG: A fast prototype generation algorithm for Big Data classification problems. *Neurocomputing*, *287*, 22–33. https://doi.org/10.1016/j.neucom.2018.01.056

[2] Rithani, M., Kumar, R. P., & Doss, S. (2023). A review on big data based on deep neural network approaches. *Artificial Intelligence Review*, *56*(12), 14765–14801. https://doi.org/10.1007/s10462-023-10512-5

[3] Mohan, G. B., & Kumar, R. P. (2022). A comprehensive survey on topic modeling in text summarization. In *Micro-Electronics and Telecommunication Engineering: Proceedings of 5th ICMETE 2021*, 231–240. https://doi.org/10.1007/978-981-16-8721-1_22

[4] Bharathi Mohan, G., & Prasanna Kumar, R. (2023). Survey of text document summarization based on ensemble topic vector clustering model. In *IoT Based Control Networks and Intelligent Systems: Proceedings of 3rd ICICNIS 2022*, 831–847. https://doi.org/10.1007/978-981-19-5845-8_60

[5] Hassib, E. M., El-Desouky, A. I., Labib, L. M., & El-Kenawy, E. S. M. (2020). WOA+ BRNN: An imbalanced big data classification framework using Whale optimization and deep neural network. *Soft Computing*, *24*(8), 5573–5592. https://doi.org/10.1007/s00500-019-03901-y

[6] Prasanna Kumar, R., Rithani, M., Bharathi Mohan, G., & Venkatakrishnan, R. (2024). Empirical evaluation of large language models in resume classification. In *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies*, 1–4. https://doi.org/10.1109/ICAECT60202.2024.10469472

[7] Elmoghazy, Y., Abuelgasim, E. M. O., Osman, S. A., Afaneh, Y. R. H., Eissa, O. M. A., & Safaei, B. (2023). Effective mechanical properties evaluation of unidirectional and bidirectional composites using virtual domain approach at microscale. *Archives of Advanced Engineering Science*, *1*(1), 27–37. https://doi.org/10.47852/bonviewAAES32021723

[8] Asniar, Maulidevi, N. U., & Surendro, K. (2022). SMOTE-LOF for noise identification in imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences*, *34*(6), 3413–3423. https://doi.org/10.1016/j.jksuci.2021.01.014

[9] Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective class-imbalance learning based on SMOTE and convolutional neural networks. *Applied Sciences*, *13*(6), 4006. https://doi.org/10.3390/app13064006

[10] Li, S., Peng, Y., Shen, Y., Zhao, S., Shao, H., Bin, G., ..., & Fan, C. (2024). Rolling bearing fault diagnosis under data imbalance and variable speed based on adaptive clustering weighted oversampling. *Reliability Engineering & System Safety*, *244*, 109938. https://doi.org/10.1016/j.ress.2024.109938

[11] Akbar, S., Ali, H., Ahmad, A., Sarker, M. R., Saeed, A., Salwana, E., ..., & Ali, F. (2023). Prediction of amyloid proteins using embedded evolutionary & ensemble feature selection based descriptors with eXtreme gradient boosting model. *IEEE Access*, *11*, 39024–39036. https://doi.org/10.1109/ACCESS.2023.3268523

[12] Vairetti, C., Assadi, J. L., & Maldonado, S. (2024). Efficient hybrid oversampling and intelligent undersampling for imbalanced big data classification. *Expert Systems with Applications*, *246*, 123149. https://doi.org/10.1016/j.eswa.2024.123149

[13] Halim, A. M., Dwifebri, M., & Nhita, F. (2023). Handling imbalanced data sets using SMOTE and ADASYN to improve classification performance of Ecoli data sets. *Building of Informatics, Technology and Science (BITS)*, *5*(1), 246–253. https://doi.org/10.47065/bits.v5i1.3647

[14] Maldonado, S., Vairetti, C., Fernandez, A., & Herrera, F. (2022). FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification. *Pattern Recognition*, *124*, 108511. https://doi.org/10.1016/j.patcog.2021.108511

[15] Grina, F., Elouedi, Z., & Lefevre, E. (2020). A preprocessing approach for class-imbalanced data using SMOTE and belief function theory. In *Intelligent Data Engineering and Automated Learning – IDEAL 2020: 21st International Conference*, 3–11. https://doi.org/10.1007/978-3-030-62365-4_1

[16] Bao, Y., & Yang, S. (2023). Two novel SMOTE methods for solving imbalanced classification problems. *IEEE Access*, *11*, 5816–5823. https://doi.org/10.1109/ACCESS.2023.3236794

[17] Kamalov, F., Thabtah, F., & Leung, H. H. (2023). Feature selection in imbalanced data. *Annals of Data Science*, *10*(6), 1527–1541. https://doi.org/10.1007/s40745-021-00366-5

[18] Li, M., Wang, H., Yang, L., Liang, Y., Shang, Z., & Wan, H. (2020). Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction. *Expert Systems with Applications*, *150*, 113277. https://doi.org/10.1016/j.eswa.2020.113277

[19] Zhang, Y., Gong, D.-W., Gao, X.-Z., Tian, T., & Sun, X.-Y. (2020). Binary differential evolution with self-learning for multi-objective feature selection. *Information Sciences*, *507*, 67–85. https://doi.org/10.1016/j.ins.2019.08.040

[20] Kumari, D., Vyshnavi, S. K., Dhar, R., Rajita, B. S. A. S., Panda, S., & Christopher, J. (2024). Smart GAN: A smart generative adversarial network for limited imbalanced dataset. *The Journal of Supercomputing*, *80*(14), 20640–20681. https://doi.org/10.1007/s11227-024-06198-3

[21] Ding, H., Sun, Y., Wang, Z., Huang, N., Shen, Z., & Cui, X. (2023). RGAN-EL: A GAN and ensemble learning-based hybrid approach for imbalanced data classification. *Information Processing & Management*, *60*(2), 103235. https://doi.org/10.1016/j.ipm.2022.103235

[22] Ding, H., Sun, Y., Huang, N., Shen, Z., Wang, Z., Iftekhar, A., & Cui, X. (2023). RVGAN-TL: A generative adversarial networks and transfer learning-based hybrid approach for imbalanced data classification. *Information Sciences*, *629*, 184–203. https://doi.org/10.1016/j.ins.2023.01.147

[23] Thapa, A., Alsadoon, A., Prasad, P. W. C., Bajaj, S., Alsadoon, O. H., Rashid, T. A., ..., & Jerew, O. D. (2022). Deep learning for breast cancer classification: Enhanced tangent function. *Computational Intelligence*, *38*(2), 506–529. https://doi.org/10.1111/coin.12476

[24] Chattopadhyay, S., Dey, A., Singh, P. K., & Sarkar, R. (2022). DRDA-Net: Dense residual dual-shuffle attention network for breast cancer classification using histopathological images. *Computers in Biology and Medicine*, *145*, 105437. https://doi.org/10.1016/j.compbiomed.2022.105437

[25] Saini, M., & Susan, S. (2023). Vggin-Net: Deep transfer network for imbalanced breast cancer dataset. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *20*(1), 752–762. https://doi.org/10.1109/TCBB.2022.3163277

[26] Bas, E., Egrioglu, E., & Kolemen, E. (2022). Training simple recurrent deep artificial neural network for forecasting using par-

ticle swarm optimization. *Granular Computing*, *7*(2), 411–420. https://doi.org/10.1007/s41066-021-00274-2

[27] Hu, Z., Zhang, J., & Ge, Y. (2021). Handling vanishing gradient problem using artificial derivative. *IEEE Access*, *9*, 22371-22377. https://doi.org/10.1109/ACCESS.2021.3054915

[28] Yang, Y., Du, H., Xiong, Z., Xu, R., Niyato, D., & Han, Z. (2025). Exploring impacts of age of information on data accuracy for wireless sensing systems: An information entropy perspective. *IEEE Transactions on Mobile Computing*, *24*(6), 4907–4924. https://doi.org/10.1109/TMC.2025.3527587

[29] Yang, Y., Du, H., Xiong, Z., Niyato, D., Jamalipour, A., & Han, Z. (2024). Enhancing wireless networks with attention mechanisms: Insights from mobile crowdsensing. *arXiv Preprint:2407.15483*.

[30] You, D., Xiao, J., Wang, Y., Yan, H., Wu, D., Chen, Z., ..., & Wu, X. (2023). Online learning from incomplete and imbalanced data streams. *IEEE Transactions on Knowledge and Data Engineering*, *35*(10), 10650–10665. https://doi.org/10.1109/TKDE.2023.3250472

[31] You, D., Yan, H., Xiao, J., Chen, Z., Wu, D., Shen, L., & Wu, X. (2024). Online learning for data streams with incomplete features and labels. *IEEE Transactions on Knowledge and Data Engineering*, *36*(9), 4820–4834. https://doi.org/10.1109/TKDE.2024.3374357

[32] Zhang, L., Yang, H., & Jiang, Z. (2018). Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. *BioMedical Engineering OnLine*, *17*(1), 181. https://doi.org/10.1186/s12938-018-0604-3

[33] Ma, W., Zhang, Y., Guo, J., & Li, K. (2021). Unbalanced abnormal traffic detection based on improved Res-BIGRU and integrated dynamic ELM optimization. *Computer Communications*, *179*, 112–130. https://doi.org/10.1016/j.comcom.2021.08.005

[34] Shi, Y., Ma, S., Zhao, Y., & Zhang, Z. (2023). A physics-informed low-shot learning for sEMG-based estimation of muscle force and joint kinematics. *arXiv Preprint:2307.05361*. https://doi.org/10.48550/arXiv.2307.05361

[35] Dehghani, M., Montazeri, Z., Trojovská, E., & Trojovský, P. (2023). Coati Optimization Algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems. *Knowledge-Based Systems*, *259*, 110011. https://doi.org/10.1016/j.knosys.2022.110011

[36] Albaqami, H., Hassan, G. M., Subasi, A., & Datta, A. (2021). Automatic detection of abnormal EEG signals using wavelet feature extraction and gradient boosting decision tree. *Biomedical Signal Processing and Control*, *70*, 102957. https://doi.org/10.1016/j.bspc.2021.102957

[37] Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., & Singh, V. (2020). Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. *Chaos, Solitons & Fractals*, *138*, 109944. https://doi.org/10.1016/j.chaos.2020.109944

[38] Awasthi, N., Dayal, A., Cenkeramaddi, L. R., & Yalavarthy, P. K. (2021). Mini-COVIDNet: Efficient lightweight deep neural network for ultrasound based point-of-care detection of COVID-19. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, *68*(6), 2023-2037. https://doi.org/10.1109/TUFFC.2021.3068190

[39] Aminu, M., Ahmad, N. A., & Noor, M. H. M. (2021). Covid-19 detection via deep neural network and occlusion sensitivity maps. *Alexandria Engineering Journal*, *60*(5), 4829–4855. https://doi.org/10.1016/j.aej.2021.03.052

[40] Su, Q., Hamed, H. N. A., Isa, M. A., Hao, X., & Dai, X. (2024). A GAN-based data augmentation method for imbalanced multi-class skin lesion classification. *IEEE Access*, *12*, 16498–16513. https://doi.org/10.1109/ACCESS.2024.3360215

How to Cite: Mohan, R., & Rangarajan, P. K. (2025). An Extreme Gradient Boosting Feature Selection–Based GAN-ELM for Classification of Imbalanced Big Data. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE52025973