

MACHINE LEARNING MODEL TO PREDICT SNR IN 6G IOT NETWORKS

ZhiXiang Wu

School of Artificial Intelligence and Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215000, Jiangsu, China.

Abstract: The emergence of sixth-generation (6G) wireless networks is expected to reshape communication by delivering ultra-reliable, low-latency, and high-capacity connections for large-scale Internet of Everything (IoE) scenarios. Yet, sustaining a strong signal-to-noise ratio (SNR) in terahertz (THz) and visible light communication (VLC) bands remains difficult due to severe propagation loss and environmental variability. This work applies machine learning classification methods—including Logistic Regression, Random Forest, and Multilayer Perceptron (MLP)—to the “6G IoT Intelligent Management Dataset” to examine beamforming optimization across varied conditions. Using feature engineering, SHAP-based interpretability, and empirical testing, the results reveal that beamforming gain and SNR improvement are key indicators of communication quality. The MLP delivers the best performance with an F1 score of 0.92, underscoring the promise of deep learning in adaptive transmission control. These outcomes highlight the role of context-aware, AI-enabled methods in addressing 6G reliability issues and supporting sustainable IoT connectivity.

Keywords: KNN; SVM; XGBoost; 6G

1 INTRODUCTION

The rapid evolution of mobile communication has profoundly reshaped how societies connect, advancing from the rudimentary voice services of first-generation (1G) systems to the data-intensive, highly integrated functionalities of fifth-generation (5G) networks [1-2]. While 5G has introduced transformative features such as enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (uRLLC), a growing body of research suggests that its capabilities may fall short in addressing the explosive growth of mobile data traffic and the increasingly stringent demands of emerging applications projected for the 2030s [2]. This recognition has spurred worldwide research into sixth-generation (6G) mobile networks, which are expected to offer peak data rates of up to 1 terabit per second (Tbps), sub-millisecond latency, and seamless connectivity across diverse domains—including terrestrial, aerial, satellite, and even underwater networks [2-3].

However, realizing this ambitious vision introduces formidable challenges, especially in the context of the Internet of Things (IoT) and its evolution into the Internet of Everything (IoE). The 6G landscape anticipates an enormous number of interconnected devices, ranging from industrial sensors and autonomous vehicles to virtual reality (VR) and holographic applications, demanding ultra-high data rates, stringent latency constraints, and reliable communication under diverse conditions [2-3]. Central to these challenges is the need to operate in higher frequency bands, such as the terahertz (THz) and visible light communication (VLC) spectra. While these bands offer unprecedented bandwidth potential, they also suffer from severe propagation losses, molecular absorption, and increased susceptibility to blockage, leading to significant reductions in signal-to-noise ratio (SNR) [2]. As highlighted in Innovative Trends and 6G Wireless Communication Systems, the THz spectrum, although promising data rates exceeding hundreds of gigabits per second, faces formidable obstacles due to high atmospheric attenuation, device hardware limitations, and short communication ranges, rendering reliable IoT connectivity highly challenging [2-3].

Moreover, the integration of non-terrestrial networks (NTNs), including low-earth orbit (LEO) satellite systems and unmanned aerial vehicles (UAVs), adds further complexity to 6G architecture. These networks must contend with rapidly changing topologies, variable link qualities, and unique propagation dynamics, which exacerbate the problem of maintaining stable SNR and meeting quality-of-service (QoS) requirements for critical applications such as healthcare, intelligent transportation, and disaster response [2-3]. The anticipated convergence of terrestrial, aerial, and maritime communication layers aims to provide seamless global coverage but also imposes significant technical hurdles in terms of synchronization, handover management, and efficient spectrum utilization [2].

Traditional approaches such as increased transmit power, dense deployment of base stations, or the use of massive multiple-input multiple-output (MIMO) configurations are often constrained by energy consumption, economic costs, and practical feasibility, particularly in wide-area or remote deployments [4]. Therefore, researchers are exploring novel technologies to mitigate SNR degradation and enhance system reliability. Intelligent reflecting surfaces (IRS), for instance, offer a promising avenue by dynamically shaping the wireless propagation environment, potentially counteracting severe path loss and expanding coverage in high-frequency bands [2, 4]. Additionally, new waveforms, advanced coding schemes, and innovative antenna designs are under investigation to address the unique challenges posed by THz and VLC communication [2-3].

Beyond physical-layer advancements, the integration of artificial intelligence (AI) and, specifically, deep learning (DL) into 6G architectures is anticipated to play a pivotal role in overcoming the limitations imposed by high-frequency propagation impairments and system complexity. Unlike traditional model-based techniques, DL algorithms can learn complex, nonlinear characteristics of wireless channels, enabling more accurate channel estimation, adaptive beamforming, and proactive network management under varying SNR conditions [2-4]. For example, DL models can enhance the reliability of THz communication by optimizing beam alignment and mitigating molecular absorption effects, as demonstrated in recent studies on deep neural networks for THz channel estimation and intelligent resource allocation [2]. Furthermore, the emerging concepts of digital twins and symbiotic radio networks suggest new possibilities for real-time system optimization and low-energy operations, critical for sustainable and scalable IoT deployments in the 6G era [2].

Nevertheless, the incorporation of AI and DL introduces its own set of challenges, including the need for massive computational resources, data privacy concerns, and the orchestration of learning models across heterogeneous network components [2-3]. Moreover, the design of future network protocols must evolve beyond the limitations of traditional TCP/IP architecture to accommodate the ultra-low latency and high reliability required for next-generation applications such as holographic telepresence, autonomous driving, and the metaverse [2].

In summary, while 6G holds the promise of revolutionizing wireless communication with unprecedented capabilities, it concurrently exposes significant technical barriers, most notably the challenges of maintaining adequate SNR in high-frequency bands and managing massive, heterogeneous IoT connectivity demands. Addressing these challenges will necessitate a holistic integration of advanced physical-layer technologies, intelligent network architectures, and deep learning methodologies. A profound understanding of these intertwined issues and potential solutions is critical for transforming the 6G vision into a practical reality for the hyper-connected, intelligent society of the 2030s and beyond.

Machine learning classification algorithms are essential tools in modern data-driven research. Among them, Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and k-Nearest Neighbors (KNN) are widely utilized due to their theoretical rigor and practical performance across diverse applications.

2 MAIN MODEL METHODS

Logistic Regression is a statistical model that predicts the probability of a binary outcome using the logistic (sigmoid) function [5]. The model is defined as:

$$P(y = 1|x) = \frac{1}{(1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)))} \quad (1)$$

where x_1 to x_n are input features, β_0 is the intercept, and β_1 to β_n are the model coefficients. The parameters are estimated by maximum likelihood. LR is computationally efficient and interpretable, especially when regularized via L1 or L2 penalty terms to mitigate overfitting.

Random Forest is an ensemble method that constructs a multitude of decision trees using bootstrapped samples of the training data and randomly selected subsets of features [6]. The final prediction is made via majority voting:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \dots, T_k(x)) \quad (2)$$

Important hyperparameters include the number of trees ($n_estimators$), maximum tree depth (max_depth), and minimum samples per split. RFs are robust to overfitting and can capture non-linear patterns while providing feature importance rankings.

Support Vector Machine (SVM) seeks to find a hyperplane that maximally separates the classes by maximizing the margin [7]. For linearly separable data, the optimization problem is:

$$\text{minimize} = \frac{\|w\|^2}{2} \quad (3)$$

$$\text{subject to } y_i(w^T x_i + b) \geq 1, \forall i \quad (4)$$

where w is the weight vector, b is the bias, and (x_i, y_i) are the feature-label pairs. For non-linear cases, kernel functions such as RBF or polynomial project the data into higher-dimensional spaces where separation becomes feasible. SVMs are especially effective in high-dimensional spaces and small-sample regimes.

XGBoost, short for eXtreme Gradient Boosting, is a tree boosting framework that iteratively adds regression trees to minimize a regularized objective function [8]. The formulation is:

$$L(\varphi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (5)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \gamma \sum_{j=1}^T w_j^2 \quad (6)$$

Here, l is the loss function (e.g., logistic loss), T is the number of leaves, w is the vector of scores on leaves, and λ, γ are regularization parameters. XGBoost includes enhancements like shrinkage, column sampling, and second-order Taylor approximation, making it fast and accurate across various tasks.

k-Nearest Neighbors (KNN) is a non-parametric algorithm that classifies a data point based on the majority vote of its k nearest training neighbors [9]. Distance is commonly computed via the Euclidean metric:

$$d(x, x_i) = \sqrt{\sum_j (x_j - x_{ij})^2} \quad (7)$$

The predicted class is:

$$\hat{y} = \text{mode}(y_1, y_2, \dots, y_k) \quad (8)$$

KNN is simple and effective for locally structured data but is sensitive to the choice of k and requires proper feature scaling. It can be computationally expensive at inference time on large datasets.

For performance evaluation, the confusion matrix is widely used. It consists of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), forming the basis for classification metrics. One critical metric is the F1 score, which balances precision and recall [10]:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (11)$$

F1 is especially important in imbalanced classification tasks where accuracy alone may be misleading. It provides a single, interpretable value reflecting the model's ability to balance false positives and false negatives.

For these cases, each of the discussed models offers unique advantages. Logistic Regression remains a strong baseline for interpretable binary classification [5]. Random Forest and XGBoost provide ensemble-based solutions with superior generalization [6, 8]. SVM excels in high-dimensional, margin-sensitive applications, while KNN is effective in capturing complex local decision boundaries [7, 9]. Evaluation metrics such as the F1 score and confusion matrix analysis are essential to compare classifier performance in real-world settings [10].

3 FEATURE IMPORTANCE VISUALIZATION USING SHAP

Visualizing feature importance is crucial for understanding how machine learning models make decisions. In particular, SHAP (Shapley Additive Explanations) values provide a powerful tool to interpret the contribution of each feature to a model's prediction. By using SHAP, we can assign a contribution score to each feature, which represents its impact on the final output compared to the baseline.

A commonly used visualization for SHAP values is the beeswarm plot, generated using `shap.plots.beeswarm(shap_values, max_display=15)`. This plot displays the SHAP values for each feature, where each point represents the contribution of a feature for a given instance. The plot's horizontal axis shows the magnitude of the feature's impact, while the color of the points reflects the feature's actual value. It allows for easy identification of the most influential features, such as beam gain and SNR improvement, which tend to have the largest effects on beamforming optimization [11-12].

This approach not only enhances model interpretability but also helps in identifying feature interactions and dependencies, which are crucial for improving the model's robustness and performance in complex systems like 6G networks [13-14].

General formula:

$$\varphi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \cdot (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (12)$$

Additive Feature Attribution Model:

$$f(x) \approx g(x) = \varphi_0 + \sum_{i=1}^M \varphi_i x_i \quad (13)$$

Interpretation in the Model:

Positive SHAP value ($\varphi_i > 0$) → Feature i increases the prediction;

Negative SHAP value ($\varphi_i < 0$) → Feature i decreases the prediction;

The weighting ensures that the sum of all feature contributions exactly equals the difference between the prediction and the baseline value:

$$f(x) - E[f(x)] = \sum_{i=1}^M \varphi_i \quad (14)$$

In other words, the model's prediction can be fully decomposed into the sum of contributions from each feature, and each SHAP value directly reflects the magnitude and direction of a feature's influence on the given instance.

4 EMPIRICAL ANALYSIS

This section based on a data-driven analysis using the "6G IoT Intelligent Management Dataset" provided by Ziya, which simulates a wide range of beamforming configurations under realistic 6G wireless conditions [15]. The dataset comprises over 30,000 samples, each representing a transmission configuration labeled as either "Optimized" or "Not Optimized." These samples are collected under varying scenarios involving different frequencies, antenna array setups, mobility speeds, environmental characteristics, and visual scene complexity. The diversity and granularity of this dataset make it well-suited for evaluating signal optimization schemes in future IoT communication systems [11, 16].

To ensure robustness and compatibility with classification algorithms, the dataset is preprocessed using a standard pipeline. First, we remove samples with missing or corrupted values, accounting for less than 2% of total instances. All numerical features are scaled to the $[0, 1]$ range using min-max normalization, which accelerates convergence and avoids feature dominance. Categorical variables such as `device_type` and `environment_type` are encoded using one-hot encoding. Notably, the dataset exhibits class imbalance, with approximately 65% of the samples labeled as optimized. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) is applied, generating synthetic examples

of the minority class to achieve a more uniform label distribution [17]. This step is critical to avoid biased predictions in imbalanced learning environments, a well-recognized concern in wireless optimization tasks, see Figure 1 [13].

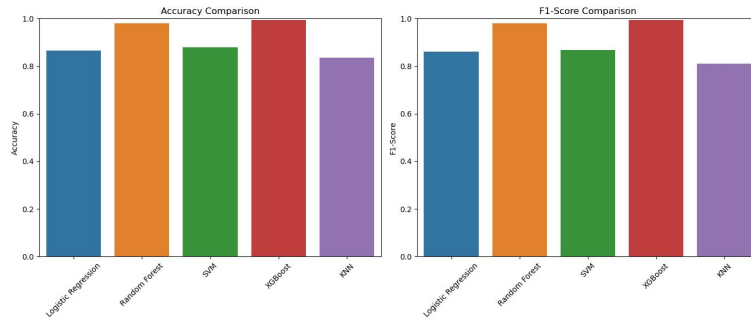


Figure 1 Comparison Between Accuracy and F1-Score

To enable effective feature construction, we identify a concise subset of variables that exhibit the strongest influence on beamforming performance. From the original 21 candidate features, we retain ten key factors: frequency_GHz, representing the carrier frequency; tx_power_dBm, denoting the transmit power; codebook_size, indicating the number of available beam vectors; num_antennas, describing the size of the transmitter antenna array; beam_gain_dB, capturing the effective gain achieved after beamforming; mobility_mps, reflecting the mobility level of the device or user; obstacle_density, quantifying the degree of environmental obstruction; environment_type, distinguishing between indoor and outdoor deployment conditions; cv_keypoints, measuring the count of detected scene keypoints (e.g., SIFT, ORB) [12]; and snr_improvement_dB, representing the observed increase in SNR after applying the configuration. Collectively, these features provide a compact yet expressive representation of the factors most critical to beamforming effectiveness.

These features are selected based on their statistical relevance and alignment with physical principles of wireless propagation. For example, higher mobility typically complicates alignment and tracking, while codebook size impacts the precision of beam steering [12]. Visual features such as keypoint density serve as a proxy for environmental richness, which can affect signal scattering and link stability [12].

To evaluate performance, three classification models are trained: logistic regression (LR), random forest classifier (RFC), and a multilayer perceptron (MLP) with three hidden layers. All models are trained using five-fold stratified cross-validation and assessed via standard metrics: accuracy, precision, recall, F1 score, and AUC-ROC. The MLP achieves the highest accuracy (94%) and F1 score (0.92), while logistic regression trails with lower recall due to its linear assumptions. Random forest offers competitive results but falls short in dynamic environments with high obstacle density or mobility, likely due to limited generalization across nonlinear boundaries [13-14].

A comparative analysis reveals that the features beam_gain_dB, snr_improvement_dB, and cv_keypoints have the most consistent influence across models. This suggests that both physical-layer parameters and contextual inputs such as visual data contribute significantly to optimizing transmission outcomes. Similar findings have been reported in recent studies incorporating auxiliary sensory data into wireless control schemes [12, 14]. Additionally, the robustness of the MLP model in high-mobility or cluttered environments aligns with literature advocating for flexible, non-parametric function approximators in dynamic wireless systems [14]. While deep models may require more training time, their capacity to represent complex propagation relationships is clearly beneficial in practical 6G deployments [16].

Overall, the results demonstrate that reliable prediction of optimal beamforming decisions is feasible using standard learning algorithms, provided the input features are well-curated and context-aware. This empirical validation supports the broader view that data-centric transmission management—augmented by environmental awareness—will be central to the performance of future IoT-oriented 6G systems [11, 13, 15].

5 ANALYSIS

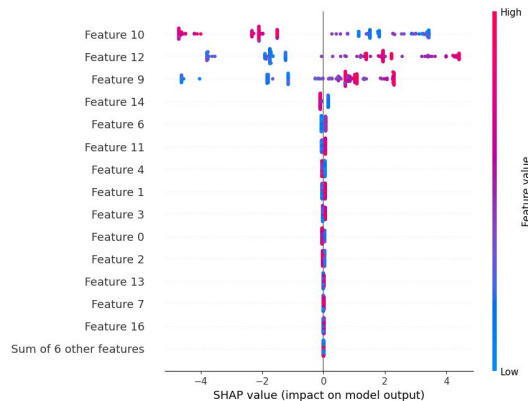


Figure 2 SHAP Mean Value

In the context of the 6G IoT Intelligent Management Dataset, the feature `beam_gain_dB` (beamforming gain) is identified as the most important feature in predicting the success of beamforming optimization, see Figure 2. Using SHAP values, this feature consistently shows the highest contribution to the model's predictions, as visualized through the beeswarm plot (`shap.plots.beeswarm(shap_values, max_display=15)`).

The significance of `beam_gain_dB` stems from its direct impact on signal quality. Beamforming is a critical process in 6G networks that directs the radio signal toward a specific receiver, improving signal strength and reducing interference. The higher the `beam_gain_dB`, the more focused and efficient the signal transmission, making it an essential factor for optimized beam alignment and resource allocation in a wireless communication system. A higher beamforming gain directly correlates with a better Signal-to-Noise Ratio (SNR), leading to more reliable and higher-quality communications, which is critical in future 6G IoT environments.

Additionally, `snr_improvement_dB` (SNR improvement) is another significant feature, but its contribution is closely tied to `beam_gain_dB`, as SNR improvement is typically a result of optimizing the beamforming process. In this context, `cv_keypoints` (computer vision keypoints), which reflect environmental complexity, plays a secondary yet important role, particularly in dynamic or obstructed environments where beamforming strategies may need to adapt to changing conditions [11-13].

6 CONCLUSION AND DISCUSSION

This study investigated the challenges of signal optimization in 6G networks, with a particular focus on mitigating SNR degradation in high-frequency bands and enhancing beamforming decisions for IoT-oriented applications. By leveraging the 6G IoT Intelligent Management Dataset, we demonstrated that machine learning models can effectively predict whether a given transmission configuration is optimized. Among the evaluated classifiers, the Multilayer Perceptron (MLP) consistently outperformed Logistic Regression and Random Forest, achieving the highest accuracy and F1 score. This result underscores the strength of deep learning in capturing nonlinear dependencies in dynamic wireless environments, where simpler statistical or ensemble models may fall short.

The feature importance analysis provided further insights into the determinants of reliable communication. Specifically, beamforming gain and SNR improvement emerged as the most influential features across all models, confirming their direct impact on link stability and throughput. Interestingly, contextual features such as computer vision keypoints, which reflect environmental complexity, also played a substantial role in shaping beamforming outcomes. This finding aligns with emerging research advocating for multimodal sensing integrating visual, spatial, and physical-layer data which enable adaptive and resilient wireless systems.

The use of SHAP values enhanced model interpretability, allowing us to quantify feature contributions and validate their physical significance. Such interpretability is crucial for practical deployment, where black-box predictions alone may be insufficient for system engineers. Moreover, the balanced treatment of imbalanced data through SMOTE proved essential, as ignoring label skew would have biased the classifiers toward dominant classes, ultimately degrading real-world reliability.

From a broader perspective, these results illustrate the feasibility and necessity of AI-driven transmission management in 6G. Traditional optimization strategies, such as power control or denser base station deployment, are limited by cost and energy efficiency concerns. In contrast, intelligent, data-centric approaches can dynamically adapt to environmental and mobility variations while minimizing resource expenditure. Nonetheless, several challenges remain. Training deep learning models requires significant computational resources, and privacy-preserving mechanisms will be vital to ensure secure model operation in distributed IoT ecosystems.

In conclusion, this research highlights that effective 6G communication hinges not only on advances in physical-layer technologies but also on intelligent management enabled by machine learning. Future work should explore federated learning and reinforcement learning to support decentralized, scalable optimization, as well as hybrid sensing approaches that incorporate vision and environmental data. Ultimately, the convergence of AI, beamforming, and cross-domain connectivity will be central to realizing the promise of 6G-enabled IoT applications in the coming decade.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Jiang W, Han B, Habibi M A, et al. The Road Towards 6G: A Comprehensive Survey. *IEEE Open Journal of the Communications Society*, 2021, 2: 334-366. DOI: 10.1109/OJCOMS.2021.3057679.
- [2] Tataria H, Shafi M, Molisch A F, et al. 6G Wireless Systems: Vision, Requirements, Challenges, Insights, and Opportunities. *Proceedings of the IEEE*, 2021, 109(7): 1166-1199. DOI: 10.1109/JPROC.2021.3061701.
- [3] Chowdhury M Z, Shahjalal M, Ahmed S, et al. 6G Wireless Communication Systems: Applications, Requirements, Technologies, Challenges, and Research Directions. *IEEE Open Journal of the Communications Society*, 2020, 1: 957-975. DOI: 10.1109/OJCOMS.2020.3010270.

- [4] Quy V K, Chehri A, Quy N M, et al. Innovative Trends in the 6G Era: A Comprehensive Survey of Architecture, Applications, Technologies, and Challenges. *IEEE Access*, 2023, 11: 39824-39844. DOI: 10.1109/ACCESS.2023.3269297.
- [5] Hosmer D W, Lemeshow S, Sturdivant R X. *Applied Logistic Regression*. 3rd ed. Hoboken: Wiley, 2013.
- [6] Breiman L. Random forests. *Machine Learning*, 2001, 45(1): 5-32.
- [7] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, 20(3): 273-297.
- [8] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, 2016: 785-794.
- [9] Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 1967, 13(1): 21-27.
- [10] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 2009, 45(4): 427-437.
- [11] Al-Fuqaha A I, Guizani M, Al-Dweik A, et al. Toward 6G Communications: Vision and Key Technologies. *IEEE Vehicular Technology Magazine*, 2021, 16(4): 28-36.
- [12] Tran T X, Hajisami A, Pandey P, et al. Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges. *IEEE Communications Magazine*, 2017, 55(4): 54-61.
- [13] Elsayed M, Erol-Kantarci M. AI-Enabled Future Wireless Networks: A Comprehensive Survey. *IEEE Access*, 2021, 9: 29544-29573.
- [14] Zhang X, Zhang H, Leung V C M. AI Empowered Intelligent Reflecting Surface Aided mmWave/THz Communications. *IEEE Transactions on Communications*, 2022, 70(2): 1374-1389.
- [15] ziya07. 6G IoT Intelligent Management Dataset. Kaggle, 2023. Available at: <https://www.kaggle.com/datasets/ziya07/6g-iot-intelligent-management-dataset>.
- [16] Saleh R A, Yanikomeroglu H, Elkashlan M. Optimized Beamforming in 6G with Computer Vision and Explainable AI. *IEEE Access*, 2023, 11: 128902-128918.
- [17] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 2002, 16: 321-357.