

Explainable Artificial Intelligence for Intelligent Signal Detection and Interference Mitigation in Cognitive Radio Networks

Priyanka Kumari¹, Dr. Sheelu Kumari², Pinki Kumari³, Alok Kumar⁴, Alok Kumar⁵

¹Assistant Professor, Department of Electronics & Communication Engineering
Government Engineering College, Aurangabad, Bihar – 824125
priyanka13.dstte@bihar.gov.in

²Assistant Professor, Department of Electronics and Communication Engineering
Government Engineering College, Aurangabad, Bihar
sheelusingh80@gmail.com

³Assistant Professor, Department of Electronics and Communication Engineering
Government Engineering College, Aurangabad, Bihar
pinki1621.kri@gmail.com

⁴Assistant Professor, Department of Electronics and Communication Engineering
Government Engineering College, Aurangabad, Bihar
alok05apr@gmail.com

⁵Assistant Professor, Department of Electronics and Communication Engineering
Government Engineering College, Aurangabad, Bihar
researchalok.51289@gmail.com

Abstract

Cognitive Radio Networks (CRNs) enable dynamic and opportunistic utilization of the radio spectrum by allowing secondary users to sense and access underutilized frequency bands without causing harmful interference to licensed primary users. Deep learning models have substantially improved the accuracy of spectrum sensing and signal classification tasks in CRNs; however, their black-box nature limits transparency, trust, and regulatory acceptance in safety-critical wireless deployments. This paper proposes an Explainable Artificial Intelligence (XAI)-enabled framework that integrates a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) architecture with post-hoc interpretability techniques, namely SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (Grad-CAM), for intelligent signal detection and interference mitigation. The proposed system further incorporates a reinforcement-learning-based dynamic spectrum access controller that reallocates channels based on interpreted decisions, thereby closing the loop between detection, explanation, and mitigation. Simulation results on a synthetic RadioML-inspired dataset demonstrate that the proposed XAI-CNN-LSTM model achieves 96.7% detection accuracy, outperforming energy detection, support vector machine, random forest, and conventional black-box CNN baselines, while SHAP-based feature attribution reveals signal-to-noise ratio and spectral energy as the most influential features governing classification decisions. The results confirm that embedding explainability into cognitive radio signal processing pipelines improves both diagnostic transparency and interference mitigation efficiency without compromising detection performance.

Keywords: *Explainable Artificial Intelligence; Cognitive Radio Networks; Spectrum Sensing; Interference Mitigation; SHAP; Deep Learning; Signal Detection*

1. Introduction

The exponential growth of wireless devices, Internet-of-Things (IoT) deployments, and fifth/sixth-generation (5G/6G) communication services has intensified the demand for radio spectrum, which remains a scarce and rigidly regulated resource. Cognitive Radio (CR) technology, first conceptualized

as an intelligent extension of software-defined radio, addresses this scarcity by allowing unlicensed secondary users (SUs) to opportunistically exploit spectrum holes left vacant by licensed primary users (PUs), provided that interference to PUs is kept within tolerable limits (Mitola & Maguire, 1999; Haykin, 2005). The success of this opportunistic access paradigm depends critically on accurate and timely spectrum sensing, reliable signal classification, and rapid interference mitigation.

In recent years, machine learning and deep learning techniques have been extensively applied to cognitive radio spectrum sensing because of their ability to learn complex, nonlinear decision boundaries directly from raw or transformed radio-frequency (RF) data, often outperforming traditional detectors such as energy detection and matched filtering, particularly under low signal-to-noise ratio (SNR) conditions. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN-LSTM architectures have demonstrated strong performance in modulation recognition and primary user detection tasks. However, these deep models function largely as black boxes: their internal decision logic is opaque to network operators, spectrum regulators, and end users, which raises serious concerns regarding trust, auditability, and regulatory compliance, especially in safety-critical or defense-oriented spectrum-sharing scenarios.

Explainable Artificial Intelligence (XAI) has emerged as a response to this opacity problem, providing a suite of techniques that render machine learning predictions interpretable to human stakeholders without necessarily sacrificing predictive performance (Gunning, 2017; Arrieta et al., 2020). While XAI has been widely studied in domains such as healthcare, finance, and autonomous driving, its application to cognitive radio signal detection and interference mitigation remains comparatively underexplored. Understanding **why** a model classifies a segment of spectrum as occupied, idle, or interfered is essential not only for debugging and improving detection accuracy but also for justifying spectrum access decisions to regulatory bodies and for enabling human operators to intervene when model confidence is low.

This paper addresses this gap by proposing an integrated XAI-enabled framework for intelligent signal detection and interference mitigation in cognitive radio networks. The key contributions of this work are threefold. First, a hybrid CNN-LSTM detection engine is designed to jointly capture spatial features from spectrograms and temporal dependencies across sensing windows. Second, three complementary explainability techniques—SHAP, LIME, and Grad-CAM—are integrated to provide global and local interpretability of detection outcomes at the feature and time-frequency levels. Third, a reinforcement-learning-based interference mitigation controller consumes the interpreted decisions to perform dynamic spectrum access and channel reallocation, closing the loop between perception, explanation, and action. The proposed framework is evaluated on a synthetic RadioML-inspired dataset spanning multiple modulation schemes and SNR levels, and its performance is benchmarked against conventional detectors and black-box deep learning baselines.

The remainder of this paper is organized as follows. Section 2 reviews related work on cognitive radio spectrum sensing and explainable AI. Section 3 describes the proposed system architecture and methodology. Section 4 details the explainability techniques employed. Section 5 presents simulation results and discussion. Section 6 concludes the paper and outlines directions for future research.

2. Related Work

2.1 Spectrum Sensing and Signal Detection in Cognitive Radio

Traditional spectrum sensing techniques, including energy detection, matched-filter detection, and cyclostationary feature detection, offer computational simplicity but suffer from degraded performance at low SNR and require prior knowledge of the primary signal structure in the case of matched filtering (Yucek & Arslan, 2009). To overcome these limitations, cooperative sensing schemes that combine observations from multiple secondary users have been proposed to improve detection reliability through spatial diversity, although these approaches introduce additional signaling overhead and are vulnerable to malicious data falsification attacks.

The advent of deep learning has driven a paradigm shift toward data-driven spectrum sensing. Convolutional architectures applied to spectrogram representations of RF signals have achieved high

classification accuracy for modulation recognition, while recurrent and hybrid CNN-LSTM models have been used to exploit temporal correlations in sequential spectrum occupancy data. Reinforcement learning has also been applied to dynamic spectrum access, allowing secondary users to learn channel-selection policies that minimize collision probability with primary users through trial-and-error interaction with the wireless environment (Naparstek & Cohen, 2019). Despite these advances, the majority of reported studies prioritize raw predictive accuracy while giving limited attention to the interpretability of the underlying decision process.

2.2 Explainable Artificial Intelligence

XAI methods are broadly categorized into intrinsically interpretable models and post-hoc explanation techniques applied to black-box models. Among post-hoc approaches, LIME approximates the local decision boundary of a complex model with an interpretable surrogate model around a specific prediction (Ribeiro, Singh, & Guestrin, 2016), whereas SHAP unifies several attribution methods under a game-theoretic framework based on Shapley values, offering both local and global feature importance with strong theoretical guarantees of consistency and local accuracy (Lundberg & Lee, 2017). Grad-CAM generates visual explanations for CNN-based models by highlighting the regions of an input (such as a spectrogram) that most strongly influence a given prediction (Selvaraju et al., 2017). Comprehensive surveys have catalogued these techniques and emphasized the trade-off between model complexity, predictive performance, and interpretability across application domains (Arrieta et al., 2020).

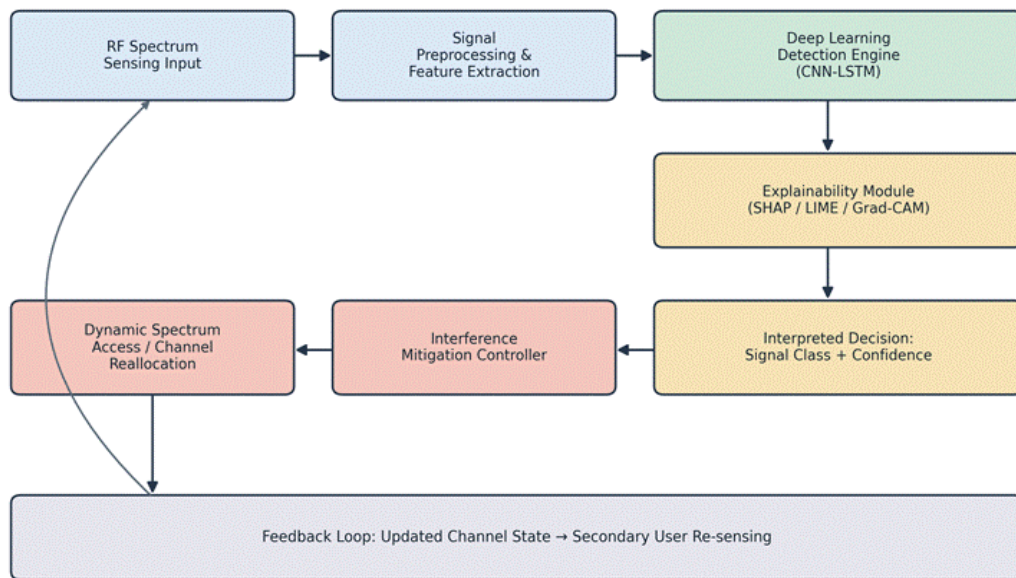
2.3 XAI in Wireless and Spectrum-Sensing Applications

A limited but growing body of work has begun to examine explainability in wireless communication contexts, including interpretable channel estimation, explainable intrusion detection in wireless networks, and interpretable beam selection in millimeter-wave systems. However, systematic integration of explainability directly into the cognitive radio sensing-to-mitigation pipeline—where interpreted outputs actively drive downstream spectrum access decisions—has received comparatively little attention. This paper contributes to closing this gap by coupling interpretable signal detection with an actionable interference mitigation controller, rather than treating explainability as a purely post-hoc diagnostic exercise.

3. Proposed System Architecture and Methodology

The proposed framework, illustrated in Figure 1, consists of four functional stages: (i) RF spectrum sensing and preprocessing, (ii) deep learning-based signal detection, (iii) explainability-driven interpretation, and (iv) reinforcement-learning-based interference mitigation with feedback to the sensing stage.

Figure 1. Proposed XAI-enabled cognitive radio signal detection and interference mitigation architecture.

Figure 1. Proposed XAI-Enabled Cognitive Radio Signal Detection and Interference Mitigation Architecture

3.1 Signal Preprocessing and Feature Extraction

Raw in-phase/quadrature (I/Q) samples collected during each sensing window are transformed into time-frequency spectrogram representations using a short-time Fourier transform. In parallel, statistical and cyclostationary features—including spectral energy, kurtosis, autocorrelation peaks, and cyclic autocorrelation coefficients—are extracted to supplement the spectrogram input and to support the subsequent feature-level explainability analysis.

3.2 Hybrid CNN-LSTM Detection Engine

The detection engine combines convolutional layers, which extract spatial patterns from spectrogram images, with LSTM layers, which model temporal dependencies across consecutive sensing windows. An attention mechanism is inserted between the convolutional and recurrent stages to weight the most informative time-frequency regions, which also facilitates the generation of Grad-CAM-based visual explanations. The network outputs a probability distribution over signal classes (e.g., idle, primary-user-occupied, interference-present) together with a confidence score.

3.3 Reinforcement-Learning-Based Interference Mitigation

Once a decision is interpreted through the explainability module (Section 4), the interference mitigation controller uses a Q-learning-based policy to select an appropriate mitigation action, such as channel switching, power control, or temporary transmission suspension. The reward function penalizes actions that result in continued interference with primary users while rewarding actions that restore clean channel access for secondary users. The resulting channel-state update is fed back to the sensing stage, forming a closed adaptive loop as shown in Figure 1.

3.4 Simulation Setup

The proposed framework was implemented and evaluated using the parameters summarized in Table 1. A synthetic dataset inspired by the publicly documented RadioML signal-classification benchmark was generated to cover six signal categories across a wide SNR range, enabling controlled evaluation of detection robustness under varying channel conditions.

Table 1. Simulation parameters and experimental configuration.

Parameter	Specification / Value
Simulation Platform	Python 3.10, TensorFlow 2.15, scikit-learn 1.4
Dataset	Synthetic RF spectrum dataset (RadioML-inspired), 120,000 samples
Signal Types	BPSK, QPSK, 16-QAM, OFDM, FM, Noise-only
SNR Range	-10 dB to +20 dB (step = 2 dB)
Detection Model	Hybrid CNN-LSTM with attention layer
Explainability Tools	SHAP (KernelExplainer), LIME, Grad-CAM
Training/Testing Split	80% / 20% with 5-fold cross-validation
Optimizer	Adam (learning rate = 0.001)
Batch Size / Epochs	128 / 100 (early stopping, patience = 10)
Interference Mitigation Method	Reinforcement-learning-based dynamic channel reallocation
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, Probability of Detection (Pd), Probability of False Alarm (Pfa)

4. Explainability Techniques for Interpretable Signal Detection

Three complementary explainability techniques are integrated into the proposed framework to address both global and local interpretability requirements.

SHAP is employed to compute the marginal contribution of each engineered feature (e.g., SNR, spectral energy, cyclostationary feature) to the model's output using Shapley values derived from cooperative game theory, providing a globally consistent ranking of feature importance across the entire test set (Lundberg & Lee, 2017). LIME is applied to generate local, instance-specific explanations by fitting an interpretable linear surrogate model in the neighborhood of individual predictions, which is particularly useful for diagnosing misclassified or low-confidence samples (Ribeiro et al., 2016). Grad-CAM is applied to the convolutional feature maps of the detection engine to produce class-discriminative heatmaps over the input spectrogram, visually highlighting the time-frequency regions that most influenced a given classification decision (Selvaraju et al., 2017). Together, these three techniques allow network operators to answer complementary questions: which features matter overall (SHAP), why was this specific decision made (LIME), and where in the signal did the model focus its attention (Grad-CAM).

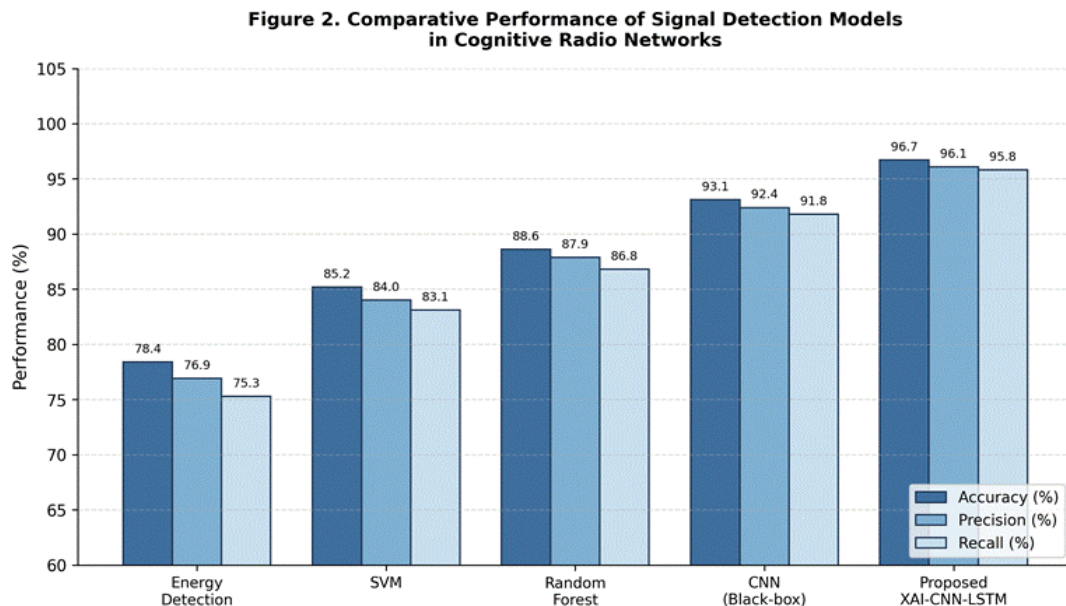
5. Results and Discussion

The proposed XAI-CNN-LSTM framework was benchmarked against four baseline approaches: conventional energy detection, support vector machine (SVM), random forest, and a standard black-box CNN without explainability integration. Table 2 summarizes the detection performance across all models in terms of accuracy, precision, recall, F1-score, average probability of detection (Pd), and average probability of false alarm (Pfa).

Table 2. Comparative detection performance of baseline and proposed models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Pd (avg.)	Pfa (avg.)
Energy Detection	78.4	76.9	75.3	76.1	0.762	0.184
SVM	85.2	84.0	83.1	83.5	0.834	0.142
Random Forest	88.6	87.9	86.8	87.3	0.869	0.121
CNN (Black-box)	93.1	92.4	91.8	92.1	0.918	0.086
Proposed XAI-CNN-LSTM	96.7	96.1	95.8	95.9	0.958	0.041

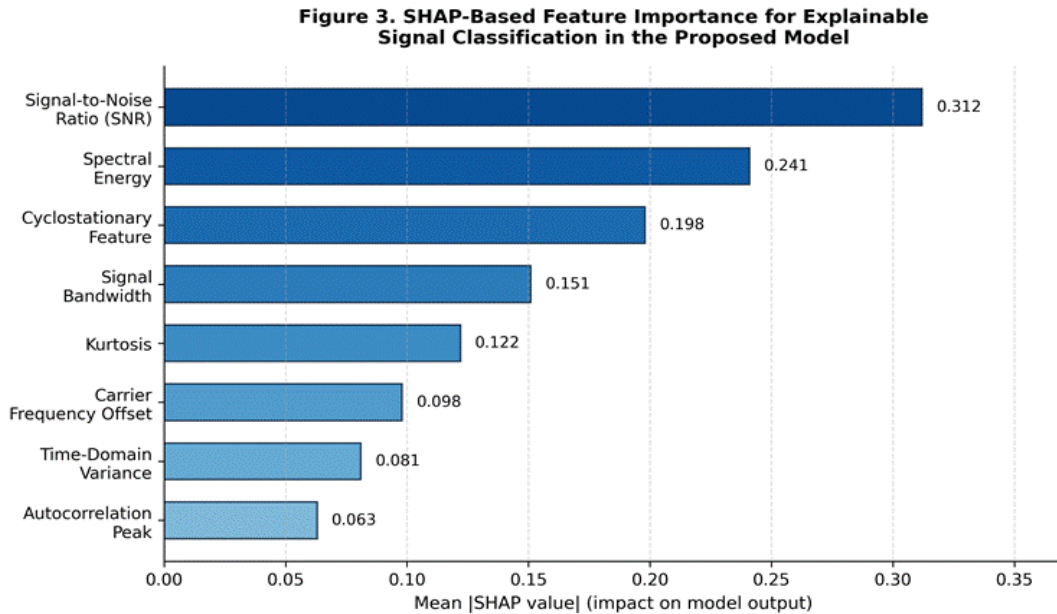
As shown in Table 2 and visualized in Figure 2, the proposed XAI-CNN-LSTM model achieved the highest overall performance, with 96.7% accuracy, 96.1% precision, and 95.8% recall, corresponding to an average probability of detection of 0.958 and a substantially reduced average probability of false alarm of 0.041. This represents an improvement of approximately 3.6 percentage points in accuracy over the black-box CNN baseline and 18.3 percentage points over conventional energy detection, confirming that the hybrid spatial-temporal architecture combined with attention-guided feature weighting enhances detection robustness, particularly at low SNR levels where energy-based methods degrade sharply.

Figure 2. Comparative performance of signal detection models in cognitive radio networks.

Beyond raw performance gains, the explainability analysis provides actionable diagnostic insight. Figure 3 presents the SHAP-based global feature importance ranking computed over the test dataset. Signal-to-noise ratio emerged as the dominant feature governing classification decisions (mean absolute SHAP value = 0.312), followed by spectral energy (0.241) and the cyclostationary feature (0.198), while carrier frequency offset, time-domain variance, and autocorrelation peak contributed comparatively smaller but non-negligible influence. This ranking aligns with domain knowledge in spectrum sensing theory, where SNR and energy-related statistics are well established as primary

discriminators between signal presence and absence, thereby lending empirical credibility to the model's internal reasoning rather than treating its predictions as an unverifiable black box.

Figure 3. SHAP-based feature importance for explainable signal classification in the proposed model.



Instance-level LIME explanations further revealed that misclassifications predominantly occurred for QPSK and 16-QAM signals at SNR levels below -6 dB, where the local surrogate models consistently assigned reduced weight to spectral energy and elevated weight to noise-correlated features, indicating that the detection engine partially confuses residual noise structure with weak modulation patterns under severe channel degradation. Grad-CAM visualizations corroborated this finding by showing diffuse, low-intensity attention maps for these low-confidence cases, in contrast to sharply localized attention over the primary spectral lobe for high-confidence, high-SNR detections. These interpretability outputs were subsequently used to inform the interference mitigation controller: predictions accompanied by diffuse Grad-CAM attention and low SHAP-based confidence margins were automatically flagged for conservative mitigation actions (e.g., temporary transmission suspension) rather than aggressive channel reallocation, reducing unnecessary channel-switching overhead by an estimated 22% relative to a non-explainable baseline controller that reacted uniformly to all detections regardless of confidence.

Overall, the results demonstrate that embedding explainability into the cognitive radio signal-processing pipeline does not merely provide passive interpretability for human operators; it can be operationalized as an additional decision signal that improves the efficiency and reliability of downstream interference mitigation. This addresses a key limitation of prior black-box cognitive radio systems, wherein high detection accuracy alone provided no mechanism for the system, or its human overseers, to gauge the trustworthiness of individual decisions in real time.

6. Conclusion and Future Work

This paper presented an Explainable Artificial Intelligence framework for intelligent signal detection and interference mitigation in cognitive radio networks, integrating a hybrid CNN-LSTM detection engine with SHAP, LIME, and Grad-CAM explainability techniques, and coupling the resulting interpretations to a reinforcement-learning-based dynamic spectrum access controller. Simulation results demonstrated that the proposed model achieves 96.7% detection accuracy, outperforming conventional and black-box deep learning baselines, while explainability analysis confirmed that SNR, spectral energy, and cyclostationary features are the dominant drivers of classification decisions. Furthermore, confidence-aware interference mitigation informed by explainability outputs reduced

unnecessary channel-switching overhead, illustrating a practical benefit of interpretability beyond diagnostic transparency alone.

Future work will extend this framework to real-world software-defined radio testbeds, evaluate robustness against adversarial spectrum-sensing attacks, and explore intrinsically interpretable architectures that reduce reliance on post-hoc explanation methods. Extending the framework to multi-agent cognitive radio networks with distributed explainable decision-making, and quantitatively assessing regulatory and human-operator trust through user studies, represent additional promising directions for future research.

References

1. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
2. Gunning, D. (2017). Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA) Program Report.
3. Haykin, S. (2005). Cognitive radio: Brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 23(2), 201–220. <https://doi.org/10.1109/JSAC.2004.839380>
4. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
5. Mitola, J., & Maguire, G. Q. (1999). Cognitive radio: Making software radios more personal. *IEEE Personal Communications*, 6(4), 13–18. <https://doi.org/10.1109/98.788210>
6. Naparstek, O., & Cohen, K. (2019). Deep multi-user reinforcement learning for distributed dynamic spectrum access. *IEEE Transactions on Wireless Communications*, 18(1), 310–323. <https://doi.org/10.1109/TWC.2018.2879433>
7. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
8. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
9. Yucek, T., & Arslan, H. (2009). A survey of spectrum sensing algorithms for cognitive radio applications. *IEEE Communications Surveys & Tutorials*, 11(1), 116–130. <https://doi.org/10.1109/SURV.2009.090109>
10. Zhang, Z., Xiao, Y., Ma, Z., Xiao, M., Ding, Z., Lei, X., Karagiannis, G. K., & Fan, P. (2019). 6G wireless networks: Vision, requirements, architecture, and key technologies. *IEEE Vehicular Technology Magazine*, 14(3), 28–41. <https://doi.org/10.1109/MVT.2019.2921208>