Rede Adversária Generativa Semi-Supervisionada para Falsificação de Sinais Modulados Utilizados em Simulação de Ataque a Modelos de Reconhecimento Automático de Modulações

Myke D. M. Valadão[1], Diego A. Amoedo[2], Samuel A. Tavares[3], Beatriz A. Santos[3],
Antonio M. C. Pereira[1], Rafael S. Furtado[1] Celso B. Carvalho[1], André L. A. da Costa[3], Waldir S. S. Júnior[1]
[1]Universidade Federal do Amazonas (DTEC/CETELI/UFAM), AM-Brazil
[2]Agência Nacional de Telecomunicações (ANATEL), AM-Brazil
[3]Universidade Federal de Uberlândia (FEELT/UFU), MG-Brazil
Emails: myke.medeiros@gmail.com, diegoalves@anatel.gov.br, {antoniopereira, ccarvalho_, rafaelfurtado, waldirjr}@ufam.edu.br, {samuel.tavares, alacosta, beatriz.santos}@ufu.br

***Resumo*— Diante do problema da falta de dinâmica na alocação de usuários no espectro de frequência, o reconhecimento automático de modulações foi uma das soluções para fornecer informações a priori e auxiliar no processo de sensoriamento de espectro. Entretanto, com o aprimoramento de redes generativas adversárias sérias questões de segurança entraram no debate mais recente. É proposto, então, a utilização de uma SGAN para gerar sinais falsificados de diversas modulações e simular ataques a modelos de reconhecimento de modulações presentes na literatura. Com o método proposto conseguimos enganar esses modelos em mais de 70% dos casos.**

***Palavras-Chave*— Rádio Cognitivo, Rede Adversária Generativa Semi-Supervisionada, Reconhecimento Automático de Modulações.**

***Abstract*— Due to the problem of lack of dynamics in the allocation of users in the frequency spectrum, the automatic recognition of modulations was one of the solutions to provide a priori information and assist in the spectrum sensing process. However, security issues have recently been introduced with the enhancement of generative adversarial networks. We propose, then, the use of an SGAN to generate different modulated signals spoofed and simulate attacks on recognition models in the literature. With the proposed method we deceived these models in more than 70% of the cases.**

***Keywords*— Cognitive Radio, Semi-Supervised Generative Adversarial Network, Automatic Recognition of Modulations.**

I. Introdução

Com o aumento da demanda por serviço de comunicação remota, o acesso ao espectro de frequência, que é um recurso limitado, ficou cada vez mais acirrado. O rádio cognitivo foi um dos recursos utilizados para sensoriar o espectro e alocar dinamicamente usuários secundários (US) em espaços do espectro parcialmente ou totalmente não utilizados pelos usuários preferenciais, usuários primários (UP) [1], [2]. O reconhecimento automático de modulações (RAM) foi uma das soluções para fornecer informações a priori e auxiliar no processo de sensoriamento de espectro [3]. Portanto, saber identificar a presença de UP em bandas de frequência é de vital importância para o melhor aproveitamento do espectro de frequência.

Entretanto, com o desenvolvimento e aprimoramento de redes adversárias generativas (GAN, do inglês *generative adversarial network*) sérias questões de segurança entraram no debate mais recente [4], [5]. Com essas redes, diversos tipos de dados podem ser criados, copiados e modificados [6]. Observando do ponto de vista do sensoriamento de espectro e da correta identificação de modulações, utilizando variações dessas redes é possível um usuário malicioso (UM) se passar por UP imitando suas características e ocupando buracos espectrais destinados a outros usuários [7].

Atualmente os modelos de RAM são treinados para reconhecer a modulação do sinal recebido, mas eles não são capazes de dizer se o sinal recebido é de um UP ou um UM imitando as características de um UP. Então, com GANs é possível que esses UMs consigam se passar por UPs e, assim, comprometerem a efeciência do rádio cognitivo. Atualmente na literatura a maioria dessas redes é utilizada para aumento de base de dados (do inglês, *data augmentation*) [8], [9], não para gerar sinais que possam ser utilizados de maneira maliciosa.

Assim como proposto em [10], com a chegada de redes do 5G/6G, foi necessário se atentar a como essas redes se tornarão seguras para seus usuários. Os autores em [10] utilizaram GAN para simular ataques e com isso fortalecer o sistema de detecção contra ataques. Sobre aplicação, existem dois times, o *Red Team*, que é responsável por ataques, e o *Blue Team*, que é o time responsável por tornar a aplicação mais segura. Fazendo um paralelo, é possível utilizar as informações adquiridas na utilização de GANs por UMs para elaborar métodos de teste de vulnerabilidade a ser seguidos para aplicações de RAM. Também, seria possível criar mecanismos para impedir que as aplicações de defesa sejam facilmente modificadas ou burladas.

Neste contexto, é proposta a utilização de uma SGAN (do

inglês, *semi-supervised generative adversarial network*) para gerar sinais falsificados de diversas modulações e simular ataques a modelos de RAM presentes na literatura. Com o método proposto os modelos testados foram enganados por mais de 70% dos sinais modulados falsificados gerados pela SGAN a 18 dB, o modelo proposto por [11] foi enganado por aproximadamente 80% dos sinais falsificados gerados a 18 dB. Além dos modelos presentes na literatura, uma avaliação de desempenho foi feita no próprio modelo discriminador gerado, já que se tratou de um treinamento semi-supervisionado. Fazendo uma comparação com reais sinais de PU com mesma SNR (do inglês, *signal to noise ratio*), chegamos perto de reais valores de acurácia alcançados por esses modelos [11], [12], [13]. Podemos concluir, então, que essas redes são poderosas ferramentas para ações de UMs do ponto de vista de eficiência de espectro.

A. Contribuições do Artigo

As contribuições científicas desta pesquisa são descritas conforme a seguir: primeiramente, existem pouquíssimas pesquisas que envolvem a utilização de redes generativas para falsificação de sinais modulados aplicados a questões de segurança no espectro de frequência, o que levanta questionamentos sobre métodos de como se previnir desse tipo de ataque; segundo, GANs usualmente são utilizadas de maneira não-supervisionada, nesse artigo é proposto o treinamento de maneira semi-supervisionada, podendo usar o discriminador como modelo de classificação também; por fim, foram usados, para simulação de ataque, três modelos presentes na literatura que utilizaram mesma base de dados para treinamento, e o desempenho alcançado demonstrou que os sinais falsificados gerados alcançaram altos níveis de correlação com sinais reais.

II. Trabalhos Relacionados

Os autores em [4] propõem a utilização de uma GAN para gerar sinais de redes sem fio sintéticos de maneira que possam ser confundidos com sinais reais. Eles usaram um modelo pré-treinado baseado em aprendizado profundo para classificar os sinais falsos recebidos. Eles levaram em consideração, nessa abordagem, também a probabilidade de sucesso no ataque em relação a localização do transmissor adversário (que envia o sinal falsificado). Em relação aos resultados, os autores conseguiram, no melhor cenário, atingir cerca de 76% de probabilidade de sucesso no ataque.

Em [14] os autores propõem ataques de interferência no espectro de frequência aplicando *adversarial machine learning*. Nesse ataque o UM, que não tem conhecimento do classificador do transmissor, possui informações de canal e detecta a decisão do transmissor. Com essas informações o UM treina seu próprio modelo classificador que é muito semelhante ao classificador do transmissor. Essa abordagem provê ao UM informações confiáveis de predição correta do canal por parte do transmissor, possibilitando, assim, que o UM realize ataques de interferência nesses canais ocupados por usuários.

Em [9] os autores fazem *data augmentation* com a utilização de uma CGAN (do inglês, *conditional generative adversarial network*) para, então, aumentarem a acurácia do modelo CNN (do inglês, *convolutional neural network*) proposto por [15]. Apesar de não ser uma abordagem voltada para segurança do espectro de frequência, como é o proposto por nós, a metodologia de geração de sinais, nesse caso, é semelhante a por nós proposta. Eles conseguiram uma melhora considerável nas métricas de avaliação do modelo, o que demonstra que o método de geração de sinais por uma CGAN gerou sinais falsos com alto nível de correlação com os sinais reais.

De maneira similar a proposta por [9], os autores em [16] também propõem a utilização de uma GAN no processo de *data augmentation* com intuito de aumentar a acurácia de modelos de detecção de sinais no espectro de frequência. Foram treinados dois modelos de aprendizado de máquina, *random forest* e *support vector machine*, para validar que o método proposto consegue imitar com níveis aceitáveis de correlação os sinais reais usados como base original. Os resultados obtidos demonstram um aumento na acurácia dos dois modelos com o *data augmentation*, o que indica que os sinais gerados pela GAN são semelhantes aos sinais reais.

III. Metodologia

A. Introdução

Usualmente, sistemas de RAM são capazes de reconhecer qual a modulação do sinal recebido, entretanto, esses sistemas não possuem a capacidade de determinar se esse sinal é de um real UP ou de um UM tentando se passar por um UP. Nós, então, para demonstrarmos a fragilidade desses modelos de RAM, é proposto a utilização de uma SGAN para gerarmos sinais modulados falsos e simularmos ataques de UMs com esses sinais. Para simularmos esses ataques, testamos os sinais falsos gerados em modelos de RAM presentes na literatura, além de usarmos o próprio modelo discriminador da SGAN como modelo de reconhecimento de modulações. Para vias de comparação, todos os modelos utilizados foram treinados com a mesma base de dados. Uma visão geral do sistema de treinamento é apresentado na Fig. 1.

B. Arquitetura da SGAN proposta

A SGAN é composta de dois componentes principais, o gerador e o discriminador. Um vetor ruído aleatório é transformado em um falso sinal modulado pelo gerador. A tarefa do discriminador é reduzir a *loss* até o ponto em que o gerador seja tão bom que o discriminador não saiba diferenciar um sinal modulado real de um sinal modulado falso, e, ao mesmo tempo, ficar tão acurado que possa também ser utilizado como modelo de reconhecimento de modulações.

O gerador proposto é composto inicialmente por uma camada *Dense* seguida de uma camada *Reshape*. Sequencialmente, a essas duas primeiras camadas, nós temos três blocos, cada um composto por *Batch Normalization*, *UpSampling2D*, *Conv2D* e *Activation*. Ao final desses três blocos temos outra *Batch Normalization* seguida de *Conv2D* e *Activation* para encerrar a arquitetura do gerador. A entrada do gerador é um

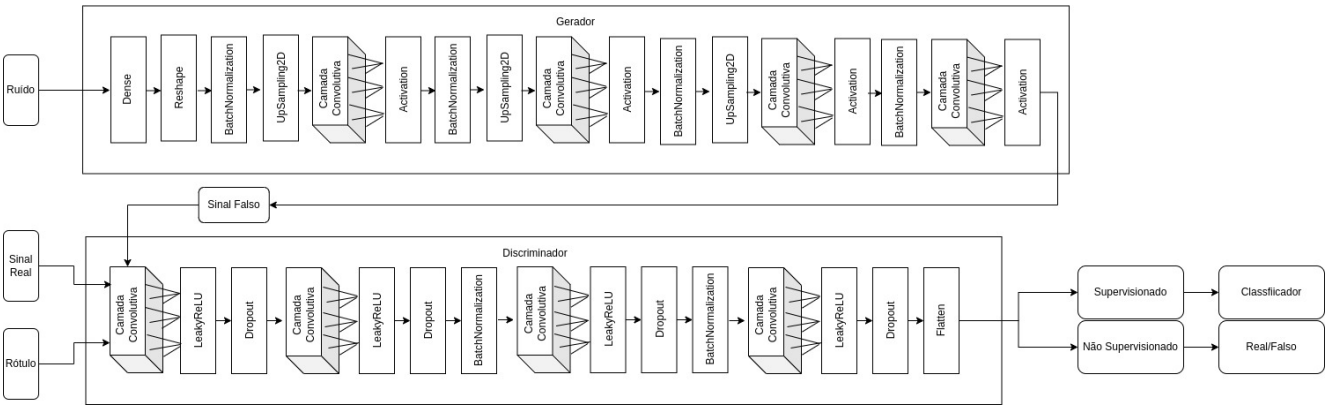

Fig. 1. Sistema de treinamento da SGAN proposta.

vetor ruído de dimensão X e a saída é um sinal modulado falso com a mesma dimensão que o sinal real da base de dados utilizada para treinamento do discriminador.

O discriminador proposto possui inicialmente dois blocos convolucionais compostos cada um por uma camada *Conv2D*, *LeakyReLU* e *Dropout*, seguidos por uma camada *Batch Normalization*. Ainda na sequência, temos um bloco convolucional seguido de outra camada *Batch Normalization*, e, finalizando, outro bloco convolucional seguido de uma camada *Flatten*. As entradas do discriminador são os sinais modulados reais e seus rótulos, e a dimensão da saída do discriminador é compatível com primeira camada *Dense* do gerador.

C. Método de simulação de ataque

O método de simulação de ataque é realizado com a utilização de modelos de RAM presentes na literatura [11], [12], [13], além do próprio modelo discriminador gerado pela SGAN. Os autores nesses trabalhos treinaram seus modelos de maneira supervisionada, utilizando a mesma base de dados. O modelo gerador é usado para gerarmos diversos sinais modulados falsos. Esses sinais falsos são as entradas dos modelos de RAM propostos na literatura e do modelo discriminador proposto. Ao final uma métrica é utilizada para avaliar o quanto os modelos foram enganados por esses falsos sinais gerados.

IV. Procedimento Experimental

A. Base de Dados

Foi utilizado para os experimentos a base de dados *RML2016.10a* que contém 11 modulações (8PSK, AM-DSB, AM-SSB, BPSK, CPFSK, GFSK, PAM4, QAM16, QAM64, QPSK e WBFM) dividida em 20 SNRs (-20 dB, -18 dB, -16 dB, -14 dB, -12 dB, -10 dB, -8 dB, -6 dB, -4 dB, -2 dB, 0 dB, 2 dB, 4 dB, 6 dB, 8 dB, 10 dB, 12 dB, 14 dB, 16 dB e 18 dB), totalizando 220000 sinais modulados [13]. Cada sinal é composto por duas componentes, em *fase (I)* e em *quadratura (Q)*, contendo cada componente 128 amostras [13].

B. Parametrização da SGAN

As redes generativas são compostas de dois componentes chaves, o gerador e o discriminador. O gerador a partir de um sinal de entrada ruído tenta aprender como deveria ser o sinal desejado e o discriminador aprender a diferenciar se o sinal gerado pelo gerador é falso ou real. Na Tabela I nos apresentamos a arquitetura da rede geradora, e na Tabela II apresentamos a arquitetura da rede discriminadora.

TABELA I
ARQUITETURA DA REDE GERADORA.

Camada	**Saída**
Dense	(*None*, 4096)
Reshape	(*None*, 1, 16, 256)
BatchNormalization	(*None*, 1, 16, 256)
UpSampling2D	(*None*, 2, 32, 256)
Conv2D	(*None*, 2, 32, 64)
Activation	(*None*, 2, 32, 64)
BatchNormalization	(*None*, 2, 32, 64)
UpSamplig2D	(*None*, 2, 64, 64)
Conv2D	(*None*, 2, 64, 32)
Activation	(*None*, 2, 64, 32)
BatchNormalization	(*None*, 2, 64, 32)
UpSampling2D	(*None*, 2, 128, 32)
Conv2D	(*None*, 2, 128, 16)
Activation	(*None*, 2, 128, 16)
BatchNormalization	(*None*, 2, 128, 16)
Conv2D	(*None*, 2, 128, 1)
Activation	((*None*, 2, 128, 1)

Além das arquiteturas das redes geradoras e discriminadora temos ainda outros parâmetros importantes. O tamanho do vetor ruído de entrada foi configurada em 100, o otimizador é o *Adam* com taxa de aprendizagem de $0,0002$ e taxa de decaimento de $0,5$ e a métrica de avaliação do treinamento é a acurácia. Foram configurados 5000 épocas com *batch size* igual a 256. A *loss* do discriminador supervisionado é a *categorical crossentropy* e da parte não supervisionada é a *binary crossentropy*.

C. Métricas

O desempenho é determinado pela acurácia na classificação de sinais falsos gerados pela SGAN por modelos de RAM presentes na literatura e o próprio modelo discriminador. Com

TABELA II
ARQUITETURA DA REDE DISCRIMINADORA.

Camada	Saída
Conv2D	(*None*, 1, 64, 32)
LeakyReLU	(*None*, 1, 64, 32)
Dropout	(*None*, 1, 64, 32)
Conv2D	(*None*, 1, 32, 64)
LeakyReLU	(*None*, 1, 32, 64)
Dropout	(*None*, 1, 32, 64)
BatchNormalization	(*None*, 1, 32, 64)
Conv2D	(*None*, 1, 16, 128)
LeakyReLU	(*None*, 1, 16, 128)
Dropout	(*None*, 1, 16, 128)
BatchNormalization	(*None*, 1, 16, 128)
Conv2D	(*None*, 1, 16, 256)
LeakyReLU	(*None*, 1, 16, 256)
Dropout	(*None*, 1, 16, 256)
Flatten	(*None*, 4096)

o valor da acurácia podemos determinar o quão eficiente os sinais falsos foram em enganar os modelos. Além dessa avaliação geral, uma avaliação feita para cada tipo de modulação com a variação da SNR.

D. Resultados

O gráfico da Fig. 2 mostra o desempenho dos sinais falsos gerados pela SGAN em relação aos três diferentes modelos presentes na literatura utilizando a mesma base de dados de treinamento com a variação de SNR, além do desempenho do próprio modelo discriminador.

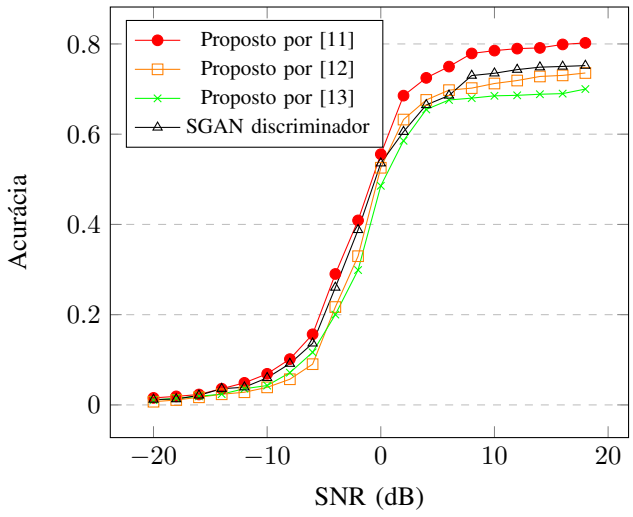

Fig. 2. Gráfico da acurácia com variação de SNR dos sinais falsificados gerados pela SGAN aplicados a três modelos diferentes de reconhecimento de modulações presentes na literatura, como também no modelo discriminador gerado.

Podemos notar no gráfico da Fig. 2 que nas SNRs mais altas os modelos foram enganados com mais facilidade, sendo o modelo proposto por [11] enganado por aproximadamente 80% dos sinais falsos gerados. Nota-se, também, que em baixas SNRs os sinais falsos não obtiveram tanto sucesso no ataque. Outra informação relevante é que o modelo discriminador foi mais enganado que os modelos propostos por [13] e [12].

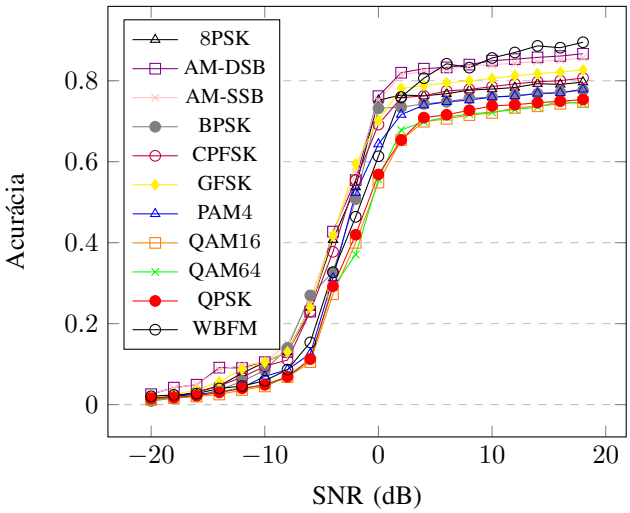

Fig. 3. Gráfico da acurácia com variação de SNR dos sinais falsos gerados por modulação testados no modelo proposto por [11].

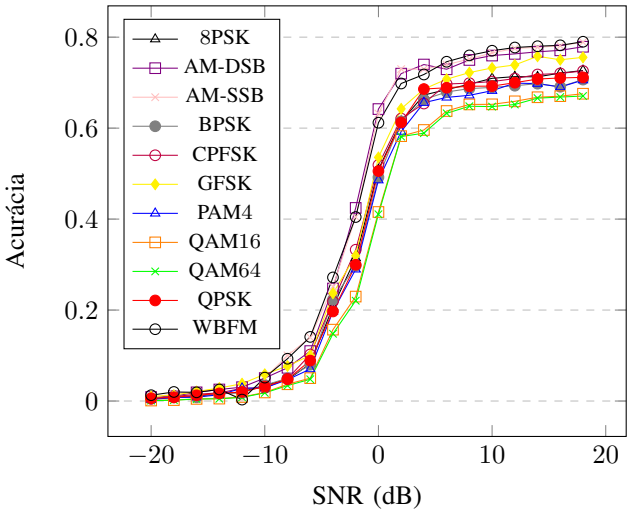

Fig. 4. Gráfico da acurácia com variação de SNR dos sinais falsos gerados por modulação testados no modelo proposto por [12].

Também, vale ressaltar que a acurácia dos modelos de [13] e [12], testados com sinais modulados reais, é inferior ao modelo de [11], o que provavelmente impactou no experimento de simulação de ataque.

Os gráficos das Fig. 3, Fig. 4, Fig. 5 e Fig. 6 mostram o comportamento dos sinais falsos gerados por modulação. Podemos notar em todas as simulações de ataque que as modulações em quadratura foram menos eficazes em enganar os modelos. Percebe-se que as modulações QAM16 e QAM64 foram as que obtiveram menor sucesso em ser classificadas pelo modelo, provavelmente por se confundirem muito.

V. CONCLUSÕES

Neste artigo, foi proposto a utilização de uma SGAN para geração de sinais modulados falsificados com o intuito de enganar modelos RAM presentes na literatura, como também, o próprio modelo discrimador gerado pela SGAN, já que foi um treinamento semi-supervisionado. Notamos que em SNR

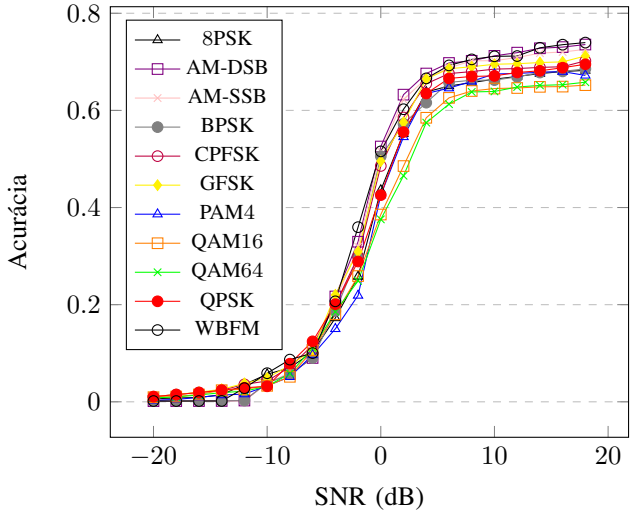

Fig. 5. Gráfico da acurácia com variação de SNR dos sinais falsos gerados por modulação testados no modelo proposto por [13].

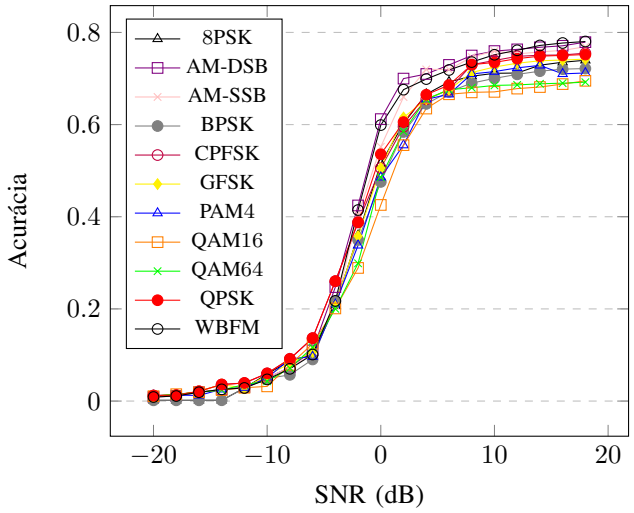

Fig. 6. Gráfico da acurácia com variação de SNR dos sinais falsos gerados por modulação testados no próprio modelo discriminador treinado pela SGAN.

acima de 5 dB os sinais falsos gerados conseguiraram alcançar acurácias competitivas com as acurácias desempenhadas pelos sinais reais nos modelos testados. No melhor caso, 18 dB, os sinais falsos gerados conseguiram enganar o modelo proposto por [11] em aproximadamente 80% dos casos. Do ponto de vista das modulações podemos notar que sinais modulados em quadratura obtiveram desempenho inferior as outras modulações. Conclui-se, então, que o sistema proposto é eficiente na tarefa de enganar modelos de RAM, o que abre uma discussão pertinente em como se denfender de UMs que possam a vir a utilizar tais ferramentas. Como trabalhos futuros, uma mudança na arquitetura da rede e aumento dos dados pode melhorar a acurácia na identificação de sinais em SNRs inferiores, como também, trabalhar em métodos de identificação de UMs no sistema.

AGRADECIMENTOS

Parte dos resultados deste artigo foram financiados por ENVISION Indústria de Produtos Eletrônicos LTDA nos termos da Lei Brasileira Federal No. 8.387/91 (SUFRAMA).

REFERÊNCIAS

[1] M.D.M. Valadão, D.A. Amoedo, A. Costa, C. Carvalho, and W.S. Silva Junior. Deep cooperative spectrum sensing based on residual neural network using feature extraction and random forest classifier. *Sensors*, 21(21):7146, 2021.

[2] M.D.M. Valadão, D.A. Amoedo, A.M.C. Pereira, S.A. Tavares, R.S. Furtado, C.B. Carvalho, A.L.A. Costa, and W.S. Silva Júnior. Cooperative spectrum sensing system using residual convolutional neural network. In *Proc. IEEE Int. Conf. on Consumer Electronics (ICCE)*, pages 1–5. IEEE, 2022.

[3] S.S. Adjemov, N.V. Klenov, M.V. Tereshonok, and D.S. Chirov. Methods for the automatic recognition of digital modulation of signals in cognitive radio systems. *Moscow University Physics Bulletin*, 70(6):448–456, 2015.

[4] Y. Shi, K. Davaslioglu, and Y.E. Sagduyu. Generative adversarial network for wireless signal spoofing. In *Proc. ACM Workshop on Wireless Security and Machine Learning*, pages 55–60, 2019.

[5] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu. Generative adversarial network in the air: Deep adversarial learning for wireless signal spoofing. *IEEE Transactions on Cognitive Communications and Networking*, 7(1):294–303, 2020.

[6] V. Balakrishnan, D. Champion, E. Barr, M. Kramer, R. Sengar, and M. Bailes. Pulsar candidate identification using semi-supervised generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*, 505(1):1180–1194, 2021.

[7] E.C. Muñoz, H.J.E. Blanco, and I.P.P. Parra. Detection of malicious primary user emulation on mobile cognitive radio networks. In *Proc. Int. Conf. on Information Systems and Computer Science (INCISCOS)*, pages 144–149. IEEE, 2019.

[8] B. Tang, Y. Tu, Z. Zhang, and Y. Lin. Digital signal modulation classification with data augmentation using generative adversarial nets in cognitive radio networks. *IEEE Access*, 6:15713–15722, 2018.

[9] M. Patel, X. Wang, and S. Mao. Data augmentation with conditional GAN for automatic modulation classification. In *Proc. ACM Workshop on Wireless Security and Machine Learning*, pages 31–36, 2020.

[10] C.P. Xuan Qui, D. Hong Quang, P.T. Duy, D. Thi Thu Hien, and V. Pham. Strengthening IDS against evasion attacks with GAN-based adversarial samples in SDN-enabled network. In *Proc. RIVF Int. Conf. on Computing and Communication Technologies (RIVF)*, pages 1–6, 2021.

[11] Y. Chen, W. Shao, J. Liu, L. Yu, and Z. Qian. Automatic modulation classification scheme based on LSTM with random erasing and attention mechanism. *IEEE Access*, 8:154290–154300, 2020.

[12] Y. Wu, X. Li, and J. Fang. A deep learning approach for modulation recognition via exploiting temporal correlations. In *Proc. IEEE Int. Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5. IEEE, 2018.

[13] T.J. O'shea and N. West. Radio machine learning dataset generation with gnu radio. In *Proc. of the GNU Radio Conference*, volume 1, 2016.

[14] Y. Shi, Y.E. Sagduyu, T. Erpek, K. Davaslioglu, Z. Lu, and J.H. Li. Adversarial deep learning for cognitive radio security: Jamming attack and defense strategies. In *Proc. IEEE Int. Conf. on Communications Workshops (ICC Workshops)*, pages 1–6. IEEE, 2018.

[15] T.J. O'Shea, J. Corgan, and T.C. Clancy. Convolutional radio modulation recognition networks. In *Proc. Int. Conf. on Engineering Applications of Neural Networks*, pages 213–226. Springer, 2016.

[16] K. Davaslioglu and Y. E. Sagduyu. Generative adversarial learning for spectrum sensing. In *Proc. IEEE Int. Conf. on Communications (ICC)*, pages 1–6. IEEE, 2018.