Um Método Baseado em *Fingerprint* de Sinais e Aprendizado de Máquina para Identificação de Estações Terrenas Interferentes

Josinaldo Azevedo[1], **Paulo C. S. Vidal**[1], **Ronaldo R. Goldschmidt**[1]

[1]Instituto Militar de Engenharia (IME)
Rio de Janeiro – RJ – Brasil

{josinaldo, vidal, ronaldo.rgold}@ime.eb.br

Abstract. *Satellite network communications are essential all over the world, and usually they are the only way to bring connectivity to hard-to-reach areas. These networks use wireless links and are affected by harmful interference. A relevant problem is to identify the source of such interference. The main technique for identifying the location of the interference source is the satellite geolocation. This technique reveals a list of suspected earth stations that probably emitted the interference. This work proposes a method that can reduce this list, by using classification models applied to Radio Frequency Fingerprint features extracted from the signals. Such method obtained accuracy above 98% in experiments with real data involving 64800 signal instances and 6 earth stations.*

Resumo. *Redes via Satélite são essenciais no mundo e, por vezes, se apresentam como o único meio de conectar regiões de difícil acesso. Tais redes utilizam comunicação sem fio e são afetadas por sinais interferentes, o que torna relevante identificar a origem desses sinais. A principal técnica para identificar a origem de sinais interferentes é a geolocalização, que apresenta um grupo de estações terrenas suspeitas. Este trabalho propõe um método que pode reduzir o número de estações elencadas pela técnica da geolocalização, ao aplicar modelos de classificação a características de Radio Frequency Fingerprint extraídas dos sinais. O método proposto obteve acurácia superior a 98% em experimentos com dados reais envolvendo 64.800 instâncias de sinais e 6 estações terrenas.*

1. Introdução

Os satélites artificiais são usados como meio de comunicação por mais de 50 anos e se tornaram uma parte essencial da infraestrutura de telecomunicações do mundo [Pratt and Allnutt 2020]. As redes baseadas nesta tecnologia fazem parte da comunicação sem fio, que é uma peça chave no desenvolvimento humano. Permitir que a comunicação ocorra de forma confiável, ágil e independente da localização geográfica é uma demanda histórica na sociedade [Laignier and Fortes 2009].

Esta demanda torna-se ainda mais desafiadora quando o objetivo é conectar as regiões mais remotas do planeta, como em diversos locais do Brasil (e.g.: Amazônia e Sertão Nordestino). Nem sempre é possível atender tais áreas com soluções tradicionais como a fibra óptica, por exemplo. E para prover conectividade nestes casos, utilizam-se as redes de comunicação via satélite.

Atualmente, vive-se em um mundo com enorme quantidade de sinais transmitidos por diversos dispositivos interligados em redes de comunicação sem fio e, a cada dia, ocorre um incremento no número e na variedade de sinais propagados [Huang et al. 2020]. Este fenômeno também impacta as redes de comunicação via satélite proporcionando um ambiente de radiocomunicação com congestionamento e sujeito às interferências prejudiciais ao seu funcionamento [Smith et al. 2021].

Neste contexto, considera-se interferência (i.e., sinal interferente) qualquer sinal não autorizado presente no espectro de radiofrequência (RF). Sinais interferentes causam diversos problemas, tais como: descontinuidade de serviços essenciais, perda de receita e a mobilização de diversos profissionais para solucionar o problema. Combater a interferência é uma tarefa fundamental para preservar a eficiência de uma rede de satélites.

A detecção de sinais interferentes e suas fontes tem sido bastante estudada nas últimas décadas [Henarejos et al. 2019]. Atualmente, a geolocalização é uma das técnicas mais utilizadas mundialmente para detectar fontes interferentes (i.e., estações emissoras de sinais interferentes) nas comunicações via satélite geoestacionário e segue recomendação da União Internacional de Telecomunicações (UIT) [UIT 2010]. Dado um sinal interferente *s*, a técnica de geolocalização consiste em apresentar a provável região de localização (denominada elipse de incerteza) da fonte emissora de *s*.

A partir da elipse de incerteza é identificada uma lista de estações que, por estarem dentro da região delimitada pela elipse, são consideradas suspeitas da emissão de *s* [Hao et al. 2020]. A etapa seguinte consiste em enviar equipes a campo para inspecionar essas estações e identificar quais delas é, de fato, a estação responsável por *s*. Esta etapa requer custos e tempo de deslocamento significativos, sobretudo nos casos de inspeção em locais remotos e de difícil acesso. Assim, quanto maior o número de estações suspeitas a serem inspecionadas, maior o custo e o tempo para execução desta etapa de inspeção.

Diante do cenário exposto, este trabalho traz a hipótese: *"É possível identificar a fonte responsável pela emissão de um sinal interferente presente no espectro destinado a redes de satélites por meio de características inerentes a este sinal"*. Caso exista um método de identificação de estações suspeitas a partir dos sinais por elas emitidos, esta hipótese será validada e o referido método poderá ser utilizado para reduzir os esforços e, consequentemente, os custos da etapa de inspeção de estações suspeitas.

A fim de buscar evidências para validar a hipótese acima, esta pesquisa propõe um método que extrai o *fingerprint* de cada sinal e o fornece a modelos de classificação (gerados a partir de algoritmos de Aprendizado de Máquina) que têm como objetivo indicar a provável estação emissora do sinal interferente. O *fingerprint* é um conjunto de recursos, como frequência e amplitude, que são extraídos do sinal capturado para identificar e verificar os dispositivos [Soltanieh et al. 2020].

Os modelos de classificação gerados em experimentos envolveram seis estações emissoras de sinal em uma rede de satélite geoestacionário, sessenta e quatro mil e oitocentas amostras de sinais e oito algoritmos de aprendizado de máquina. Os resultados apresentaram acurácia superior a 98%, fornecendo evidências experimentais de validade da hipótese levantada.

O texto encontra-se organizado em mais quatro seções. A Seção 2 apresenta os fundamentos teóricos necessários para compreensão deste trabalho. A Seção 3 resume os

trabalhos relacionados e destaca suas principais diferenças e similaridades em relação a esta pesquisa. A Seção 4 descreve o funcionamento do método proposto. A Seção 5 descreve os experimentos realizados e as análises dos resultados gerados. As considerações finais deste trabalho estão na Seção 6 e englobam as principais contribuições da pesquisa desenvolvida, assim como algumas sugestões de trabalhos futuros.

2. Fundamentação Teórica

Esta seção apresenta os elementos que compõem a infraestrutura de uma rede de comunicação via satélite, a técnica de geolocalização utilizada na identificação de fontes interferentes, e o conceito de *fingerprint* no contexto da radiofrequência.

2.1. Rede de Comunicação via Satélite

Toda rede de comunicação via satélite possui três elementos básicos [Pratt and Allnutt 2020]: a estação terrena, o satélite e a faixa de RF utilizada.

Uma estação terrena, em geral, fica localizada em grandes centros urbanos e estará interligada à infraestrutura das operadoras de telecomunicações. Ela terá o papel de conectar-se à outra estação terrena que, por sua vez, estará instalada na localidade remota. O satélite é um um transceptor de RF, que é um receptor ligado a um transmissor, usando diferentes frequências de rádio para transmitir e receber o sinal de uma estação terrena, amplificá-lo e retransmiti-lo para outra estação terrena [Pratt and Allnutt 2020]. A estação terrena tem como principal papel atender o tráfego das regiões que não possuem outro meio de comunicação. Nesta pesquisa, a conexão ocorre na faixa de RF em banda C (3,625 GHz a 4,2 GHz).

2.2. Técnica de Geolocalização

A técnica de geolocalização é aplicada no tratamento de interferências em sistemas de comunicação via satélite. Ela indica a região mais provável que a fonte interferente pode estar, mas não aponta especificamente uma estação [Hao et al. 2020].

As abordagens tradicionais de geolocalização para detecção de interferência são baseadas principalmente em medições de diferença de tempo de chegada (*Time-Difference-Of-Arrival* - TDOA) e diferença de frequência de chegada (*Frequency-Difference-Of-Arrival* - FDOA). Como ela apenas indica a provável região que uma determinada fonte interferente pode estar localizada, existe a possibilidade de conter mais de uma estação terrena. Com isto, uma equipe técnica é enviada ao local para visitar uma a uma até descobrir a responsável pelo sinal interferente.

2.3. *Fingerprint* para Radiofrequência

Conforme apresentado em [Gahlawat 2020], artefatos com tecnologia sem fio apresentam características únicas nas suas transmissões, inclusive equipamentos que compõem as redes com comunicação via satélite, pois cada dispositivo tem suas próprias assinaturas e isso se deve principalmente aos vários componentes não lineares usados. A não linearidade se deve aos amplificadores de potência, resistores e demais componentes do transmissor, por exemplo. Em um canal sem fio ruidoso, esses são os únicos recursos de um dispositivo que não mudam e não podem ser replicados facilmente, e é por isso que são bons para identificar os dispositivos, pois são específicos do hardware.

A afirmação anterior também é encontrada em [Deng et al. 2017], onde é descrito que a extração de impressão digital de RF (*Radio Frequency Fingerprint* - RFF) é uma tecnologia que pode identificar o transmissor de rádio através da análise do sinal radioelétrico transmitido. O RFF extraído a partir das características das ondas eletromagnéticas emitidas pelo transmissor é única, pois é a camada física que possibilita a identificação, e onde se extrai o *fingerprint* (impressão digital) do dispositivo sem fio. Isto ocorre graças as imperfeições de hardware no circuito analógico.

3. Trabalhos Relacionados

A Tabela 1 apresenta uma comparação entre oito pesquisas e este trabalho para identificação do sinal em diferentes faixas de RF ou tipos de redes empregadas pelos autores nos artigos citados (vide coluna mais à direita na Tabela). Cabe destacar que todos os trabalhos empregam técnicas de Aprendizado de Máquina e não houve uma padronização com relação as métricas de avaliação utilizadas por eles.

Para comparar os trabalhos analisados e esta pesquisa, os itens avaliados foram divididos nos seguintes tópicos:

a) Existe comparação entre técnicas de AM - se a pesquisa utilizou mais de um algoritmo de Aprendizado de Máquina;

b) Coleta de medidas num cenário real de campo - analisa se o estudo foi num cenário real ou simulado, e se utilizou equipamentos de medição profissionais e calibrados;

c) Análise de características radioelétricas (*fingerprint*) - mostra se o trabalho utilizou o conceito de *Radio Frequency Fingerprint*;

d) Identifica o transmissor - verifica se o experimento identifica o equipamento responsável pela interferência;

e) Análise do sinal na rede satélite - identifica se o trabalho analisa sinais de RF em uma rede satélite;

f) Identifica a estação terrena - se a pesquisa identifica a estação terrena como transmissor interferente (caso particular para uma rede satélite).

Tabela 1. Comparação entre os trabalhos relacionados e a proposta.

Autor	Itens Avaliados						Rede/Sinal Avaliado
	a	b	c	d	e	f	
[Kennedy et al. 2008]	Não	Sim	Sim	Sim	Não	Não	Celular
[Brik et al. 2008]	Não	Não	Sim	Sim	Não	Não	WIFI
[Deng et al. 2017]	Não	Não	Sim	Sim	Não	Não	Rádios Anykey AKDS700
[Bitar et al. 2018]	Não	Sim	Sim	Sim	Não	Não	WIFI, Zigbee e Bluetooth
[Henarejos et al. 2019]	Não	Sim	Sim	Sim	Sim	Não	Celular e Satélite - GEO
[Li et al. 2020]	Sim	Não	Sim	Sim	Não	Não	Gerador de sinal
[Huang et al. 2020]	Não	Sim	Não	Não	Sim	Não	Sinal estação Terrena - GEO
[Gahlawat 2020]	Sim	Não	Sim	Sim	Sim	Não	GNSS (Referência GPS) - LEO
Proposta	**Sim**	**Sim**	**Sim**	**Sim**	**Sim**	**Sim**	**Sinal estação Terrena - GEO**

Os trabalhos indicados na Tabela 1 encontram-se resumidos a seguir.

[Kennedy et al. 2008] assume que a resposta de frequência do receptor permanece constante e as únicas diferenças nas amostras de banda base digital produzidas pelo rádio do receptor são devidas aos diferentes componentes dos elementos transmissores, bem como ruído de canal e interferência. Esta abordagem mostrou ser capaz de distinguir oito transmissores idênticos com 97% de precisão em 30dB SNR (Relação Sinal Ruído).

Em [Brik et al. 2008] foi implementada e avaliada uma técnica para identificar placas de interface de rede, através da leitura de quadros do protocolo IEEE 802.11, por meio de análise passiva de RF. Demonstraram experimentalmente a eficácia na diferenciação entre mais de 130 placas de rede idênticas com precisão superior a 99%.

Em [Deng et al. 2017] foi analisado o RFF de acordo com a estrutura física do transmissor de rádio. Um sistema de aquisição de sinal foi projetado para capturar os sinais para avaliar sua abordagem, onde os sinais são gerados a partir de três rádios Anykey AKDS700. O método proposto alcançou a acurácia de 93,73%.

No trabalho de [Bitar et al. 2018] foi estudado a utilização das Redes Neurais Convolucionais (do inglês *Convolutional Neural Network* - CNN) para o problema de identificação de dispositivos sem fio coexistentes. Eles distinguiram as transmissões simultâneas das redes WIFI, Zigbee e Bluetooth, que utilizam a mesma faixa de RF, e obtiveram uma precisão de 93%.

Na pesquisa de [Henarejos et al. 2019] foi proposto o uso de técnicas de Aprendizado Profundo (do inglês *Deep Learning*) para apresentar um sistema para gerenciamento de interferências em sistemas satélites e redes celulares (LTE, UMTS e GSM), com acurácia superior a 90%. Descreveram dois subsistemas capazes de detectar a presença de interferência, mesmo em alta relação sinal para interferência (do inglês, *Signal Interference Ratio* - SIR), e classificação de interferência em diversos padrões de rádio.

O trabalho de [Li et al. 2020] teve o objetivo de identificar dispositivos sem fio através da análise do RFF. Para tal, foram utilizados características não lineares dos componentes internos de hardware de um transmissor contidos num banco de dados. As pequenas diferenças e imprecisões no processo de fabricação determinam a característica única contida no sinal transmitido. A pesquisa obteve uma acurácia de 99% na distinção e identificação dos dispositivos estudados.

Em [Huang et al. 2020] foi proposto um novo método para detecção de sinal no espectro de banda larga de redes satélite, sem a preocupação de identificar o transmissor, com base em Redes Neurais Convolucionais. Obtiveram uma precisão de 98,32%.

A pesquisa que mais se aproximou da proposta deste trabalho foi verificada em [Gahlawat 2020]. Nela foi proposto utilizar a técnica de *fingerprint* para fins de segurança. Os sinais dos satélites que disponibilizam as referências de GPS (*Global Navigation Satellite System* - GNSS) podem ser substituídos por outras referências de forma maliciosa. Esta ação pode levar os sistemas que utilizam tal referência a fornecerem uma resposta errônea, e apresentarem resultados indesejados. O trabalho propõe que os sinais sejam filtrados e selecionados segundo as identificações proporcionadas pela técnica do *fingerprint*. Isto praticamente eliminaria a chance do sinal do GPS ser substituído por uma outra fonte, e tornaria mais segura a utilização desta rede.

Até onde foi possível investigar, não foram encontrados trabalhos que abordassem

a utilização do *fingerprint* para o estabelecimento de uma identidade inequívoca dos sinais transmitidos por uma estação terrena no âmbito de uma rede de comunicação via satélite.

4. Método Proposto

A fim de verificar a validade da hipótese levantada neste trabalho, foi utilizado o método representado graficamente na Figura 1. Basicamente, este método consiste em extrair o *fingerprint* (i.e., instâncias de sinais) previamente coletado junto a diferentes estações terrenas, pré-processar essas instâncias para, em seguida, treinar modelos de classificação e avaliar a capacidade desses modelos em, dado um sinal arbitrário, identificar corretamente a estação terrena responsável pela emissão desse sinal. Os próximos parágrafos apresentam um detalhamento conceitual de cada uma das etapas do método proposto e das informações processadas por ele.

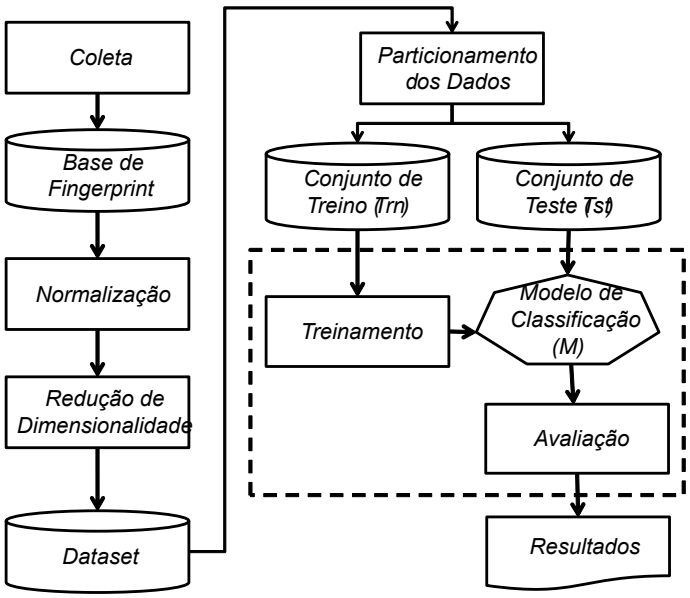

Figura 1. Método para Extração e Classificação do *Fingerprint*.

A etapa de *Coleta* consiste em receber sinais emitidos por diferentes estações terrenas. Para cada sinal s emitido por uma estação terrena r, o processo de *Coleta* obtém o *fingerprint* de s. O *fingerprint* de s corresponde a instâncias $s_{t_1}, s_{t_2}, ..., s_{t_n}$ de s, de tal forma que cada instância s_{t_i} é um conjunto ordenado de valores $p_1, p_2, p_3, ..., p_m$, onde cada p_k representa a magnitude (expressa em dBm) de uma amostra (i.e., um ponto da envoltória) de s_{t_i}. Cada s_{t_i} é obtido no instante de tempo t_i, de tal forma que o intervalo de tempo decorrido entre a coleta de quaisquer instâncias consecutivas s_{t_i} e $s_{t_{i+1}}$ seja constante e igual a um δ_T (i.e., $\delta_T = t_{i+1} - t_i$), previamente definido pelo analista responsável pela execução do processo de *Coleta*. Para cada instância s_{t_i} de um sinal s, o processo de *Coleta* armazena os valores $p_1, p_2, p_3, ..., p_m$ das amostras de s_{t_i} associando a eles a identificação da estação terrena r responsável pela emissão de s. Esses dados são armazenados no repositório denominado *Base de Fingerprint*.

Após concluída a etapa de *Coleta*, o *fingerprint* dos sinais contidos na *Base de Fingerprint* passa por um processo de *Normalização*. Tal processo tem como objetivo assegurar que as amostras das instâncias de sinal coletadas previamente tenham a mesma

ordem de grandeza. Cabe ao analista de dados a escolha do método de normalização de dados a ser utilizado nesta etapa.

Em seguida, os dados normalizados passam pela etapa de *Redução de Dimensionalidade*. Conforme o próprio nome indica, tal etapa realiza uma transformação nos dados de tal forma que a dimensão do conjunto de entrada (no caso, m - número de amostras de cada *fingerprint*) seja reduzida para um valor p, $p < m$. De forma análoga à etapa anterior, o analista de dados é o responsável por escolher o método de redução de dimensionalidade a ser aplicado.

Após a etapa de *Redução de Dimensionalidade*, é gerado um *dataset* para ser dividido nos conjuntos de Treino (Trn) e de Teste (Tst), de forma que $Trn \cap Tst = \emptyset$. Tal particionamento é realizado pela etapa chamada *Particionamento dos Dados*. A proporção de instâncias de sinal de cada estação terrena alocada nos conjuntos Trn e Tst é um parâmetro previamente definido pelo analista de dados.

A etapa de *Treinamento* consiste em gerar um *Modelo de Classificação* M que seja capaz de identificar a estação terrena (classe) responsável pela emissão de cada sinal cuja instância seja submetida a ele. Para tanto, aplica um algoritmo de aprendizado de máquina sobre os dados de Trn. Por fim, a etapa de *Avaliação* testa a capacidade de M identificar corretamente a estação terrena de onde partiu o sinal de cada instância armazenada em Tst. Diferentes algoritmos de classificação podem ser experimentados e avaliados nesta etapa, cabendo sua escolha também ao analista de dados.

É importante notar que, uma vez selecionado o modelo de classificação M' com melhor desempenho gerado pelo método proposto nesta Seção, M' pode ser usado na prática de forma a procurar contribuir para reduzir os custos decorrentes das inspeções de campo realizadas a partir da lista de estações suspeitas gerada pela técnica de geolocalização. Para tanto, uma instância do sinal interferente deve ser apresentada a M' para que ele indique qual, dentre as estações suspeitas apontadas pela técnica de geolocalização, deve ser priorizada na inspeção de campo.

5. Experimentos e Resultados

Os experimentos realizados neste trabalho em busca de evidências de validade da hipótese levantada tiveram o suporte de uma prestadora de serviços de telecomunicações e da Agência Nacional de Telecomunicações (ANATEL). Todos seguiram o método descrito na seção anterior e envolveram a análise dos dados coletados de seis sinais transmitidos por seis estações terrena, onde cada estação transmitiu um sinal. Os sinais coletados foram recebidos de um satélite geoestacionário com cobertura no Brasil e possuíam as mesmas características de RF, tais como: modulação, largura de banda, nível de transmissão, FEC (*Forward Error Correction*) e demais atributos de transmissão.

A etapa de *Coleta* ocorreu em janeiro de 2022 e é continuação da pesquisa reportada em [Azevedo et al. 2021] apresentada no SBBD-DSW/2021. Durante sua execução, foram utilizados quatro intervalos de tempo δ_T entre as instâncias de cada sinal: 50ms, 100ms, 150ms e 200ms. Desta forma, foram produzidas quatro *bases* distintas de *fingerprint*[1], uma para cada δ_T. De cada sinal foram extraídas $n = 2.700$ instâncias, sendo

[1]Disponível em `https://drive.google.com/drive/folders/16Thra0IEwwwZet8FUf8SluEWZZnqEz5z?usp=sharing`

$m = 801$ amostras por instância. Sendo assim, cada uma das quatro *bases* foi composta por 6 sinais (cada um advindo de uma estação terrena distinta), $6 * n = 16.200$ instâncias e $6 * n * m = 12.976.200$ amostras. As quatro *bases* totalizaram 64.800 instâncias e 51.904.800 amostras conforme sumarizado na Tabela 2. A Figura 2 ilustra algumas amostras na envoltória de uma instância de um sinal coletada a $\delta_T = 50ms$. Maiores detalhes sobre o procedimento de *Coleta* seguido nesses experimentos podem ser obtidos em [Azevedo et al. 2021].

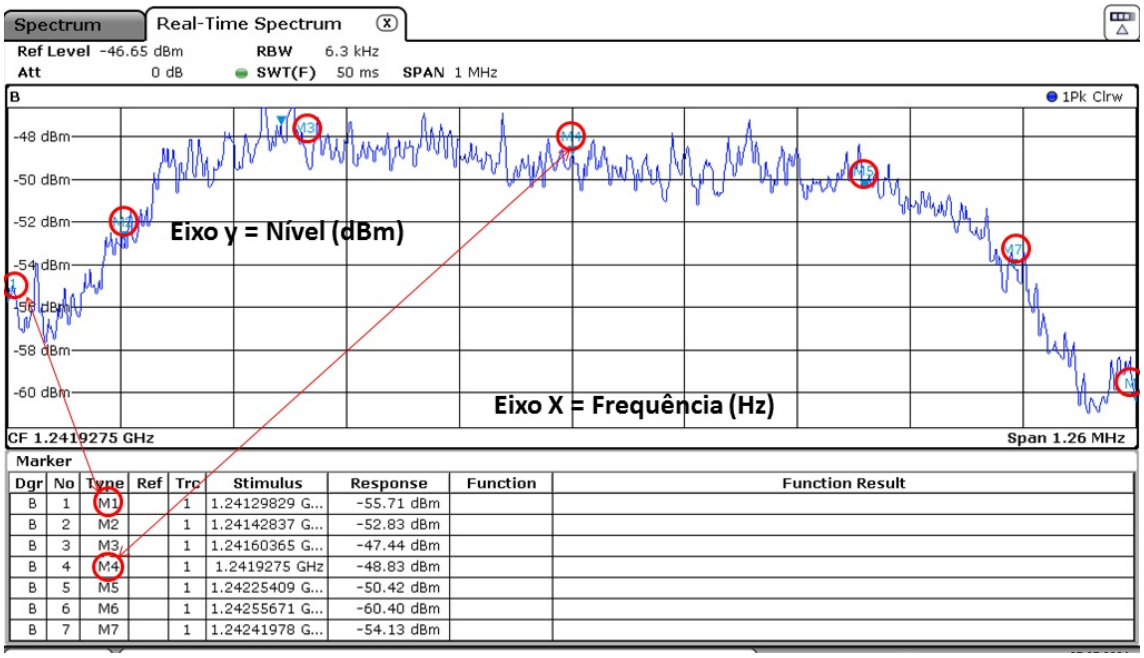

Figura 2. Exemplo de amostras de uma instância de sinal coletada a $\delta_T = 50$ ms.

Tabela 2. Resumo dos Dados Coletados para os Experimentos

Intervalo de Coleta (δ_T)	# Estações Terrenas	# Sinais Analisados	# Instâncias por Sinal	# Amostras por Instância	Total de Instâncias	Total de Amostras
50ms	6	6	2.700	801	16.200	12.976.200
100ms	6	6	2.700	801	16.200	12.976.200
150ms	6	6	2.700	801	16.200	12.976.200
200ms	6	6	2.700	801	16.200	12.976.200
				Total Coleta	**64.800**	**51.904.800**

Os dados de cada uma das quatro *bases* de *fingerprint* foram submetidos à etapa de *Normalização*, para garantir valores no intervalo $[0, 1]$. O método de normalização de dados utilizado foi o reescala linear [Aggarwal et al. 2015].

Para a etapa de *Redução de Dimensionalidade*, o método utilizado foi o PCA (*Principal Components Analysis*) [Faceli et al. 2021]. O PCA reduziu a quantidade de amostras por sinal de 801 para 3 componentes, preservando 70% de variância nos dados no espaço de dimensão reduzida.

Cada um dos quatro *datasets* obtidos ao final da execução das etapas anteriores foram aleatoriamente particionados em dois conjuntos[2], onde o Conjunto de Treino (Trn) foi composto por 10.800 instâncias (67% do total de instâncias dos seis sinais), e o Conjunto Teste (Tst) por 5.400 instâncias (33% do total de instâncias dos seis sinais). A fim de garantir o balanceamento de classes nos Conjuntos Trn e Tst, o critério de particionamento utilizado foi o de seleção aleatória estratificada [Faceli et al. 2021].

Para treinamento e avaliação dos modelos de classificação foram utilizadas as implementações dos algoritmos K-NN, *Random Forest*, *Multilayer Perceptron* (MLP), SVM, *Naïve Bayes*, Árvore de Decisão (C4.5), *Logistic Regression* e *Gradient Boosting*, disponíveis na suíte *Orange Canvas*[3]. Informações sobre os algoritmos utilizados, incluindo referências mais detalhadas a respeito deles, podem ser obtidas em [Faceli et al. 2021] e [Aggarwal et al. 2015]. Os valores adotados nos hiperparâmetros foram os valores *default* disponíveis na ferramenta. Todos os resultados apresentados são fruto de um processo de validação cruzada com 10 conjuntos. A plataforma computacional utilizada foi Intel(R) Core(TM) i5 CPU M480 @ 2.67 GHz e 6 GB de RAM e Sistema Operacional Windows.

As Tabelas 3 e 4 apresentam os desempenhos dos modelos de classificação produzidos nos experimentos. As *métricas de desempenho* [Faceli et al. 2021] foram: Acurácia, Precisão, Revocação e F1. É importante destacar que os valores de Precisão, Revocação e F1 representam as médias aritméticas simples das respectivas métricas considerando as seis classes do problema (i.e., as seis estações terrenas). Para fins de comparação, o *baseline* considerado foi a lista de estações terrenas obtidas pela técnica de geolocalização.

Tabela 3. Resultado da Classificação das Coletas com Intervalos de 50/100ms.

	δ_I=50ms						δ_I=100ms					
	Tempo de Treino (s)	Tempo de Teste (s)	Acurácia	Precisão	Revocação	F1	Tempo de Treino (s)	Tempo de Teste (s)	Acurácia	Precisão	Revocação	F1
KNN	0.23	2.644	93.92%	81.76%	81.76%	81.76%	0.242	3.235	98.29%	94.87%	94.87%	94.87%
Random Forest	2.255	0.259	94.32%	82.96%	82.96%	82.96%	2.475	0.246	98.19%	94.56%	94.56%	94.56%
MLP Neural Network	169.959	0.54	94.35%	83.04%	83.04%	83.04%	120.629	0.512	98.44%	95.31%	95.31%	95.31%
SVM	9.496	2.084	94.33%	82.98%	82.98%	82.98%	6.114	1.248	98.51%	95.54%	95.54%	95.54%
Naive Bayes	0.155	0.232	90.10%	70.31%	70.31%	70.31%	0.159	0.044	95.82%	87.46%	87.46%	87.46%
Decision Tree (C4.5)	5.073	0.008	94.04%	82.13%	82.13%	82.13%	0.341	0.012	97.94%	93.83%	93.83%	93.83%
Logistic Regression	9.459	0.041	92.36%	77.09%	77.09%	77.09%	7.343	0.055	98.44%	95.31%	95.31%	95.31%
Gradient Boosting	135.44	1.656	94.33%	83.00%	83.00%	83.00%	153.916	1.827	98.27%	94.80%	94.80%	94.80%

Em uma visão geral, o primeiro ponto a ser destacado sobre os resultados é que todos os algoritmos produziram modelos com avaliações superiores a 70% em todas as métricas em todos os *datasets*, conforme pode ser observado no gráfico da Figura 3. Tal fato é um indício de que a aplicação de modelos de classificação no *fingerprint* dos sinais parece ser adequada para identificar estações responsáveis pela emissão de sinais. O resultado global é ainda melhor se o algoritmo Naïve Bayes for excluído da análise. Neste caso, os menores valores obtidos ficam em um patamar mínimo de 77%. Tais resultados são bem razoáveis pois, o Naïve Bayes, em geral, costuma apresentar desempenhos baixos, uma vez que na maioria dos problemas as variáveis não são independentes entre

[2]Disponíveis em `https://drive.google.com/drive/folders/1rowOZ_d4CeIGGqQr0l7WdUh_rDfbOxkH?usp=sharing`

[3]Disponível em `https://orangedatamining.com/`

Tabela 4. Resultado da Classificação das Coletas com Intervalos de 150/200ms.

	δ_T=150ms						δ_T=200ms					
	Tempo de Treino (s)	Tempo de Teste (s)	Acurácia	Precisão	Revocação	F1	Tempo de Treino (s)	Tempo de Teste (s)	Acurácia	Precisão	Revocação	F1
KNN	0.228	2.855	96.36%	89.07%	89.07%	89.07%	0.192	2.473	94.56%	83.69%	83.69%	83.69%
Random Forest	2.074	0.252	96.45%	89.35%	89.35%	89.35%	1.807	0.214	94.94%	84.83%	84.83%	84.83%
MLP Neural Network	93.046	0.499	96.72%	90.17%	90.17%	90.17%	208.361	0.417	94.96%	84.87%	84.87%	84.87%
SVM	5.968	1.242	96.91%	90.72%	90.72%	90.72%	5.174	1.116	95.38%	86.13%	86.13%	86.13%
Naive Bayes	0.162	0.045	94.79%	84.37%	84.37%	84.37%	0.13	0.031	92.62%	77.87%	77.87%	77.87%
Decision Tree (C4.5)	0.327	0.011	96.20%	88.61%	88.61%	88.61%	1.875	0.013	94.52%	83.56%	83.56%	83.56%
Logistic Regression	7.343	0.04	96.57%	89.70%	89.70%	89.70%	6.485	0.035	94.21%	82.63%	82.63%	82.63%
Gradient Boosting	143.387	1.772	96.50%	89.50%	89.50%	89.50%	124.299	1.622	95.06%	85.17%	85.17%	85.17%

si, o que contraria a premissa em que o algoritmo se baseia. Em contrapartida, é importante observar que as métricas de desempenho do próprio Naïve Bayes melhoram quando aumentam os tamanhos dos intervalos de coleta.

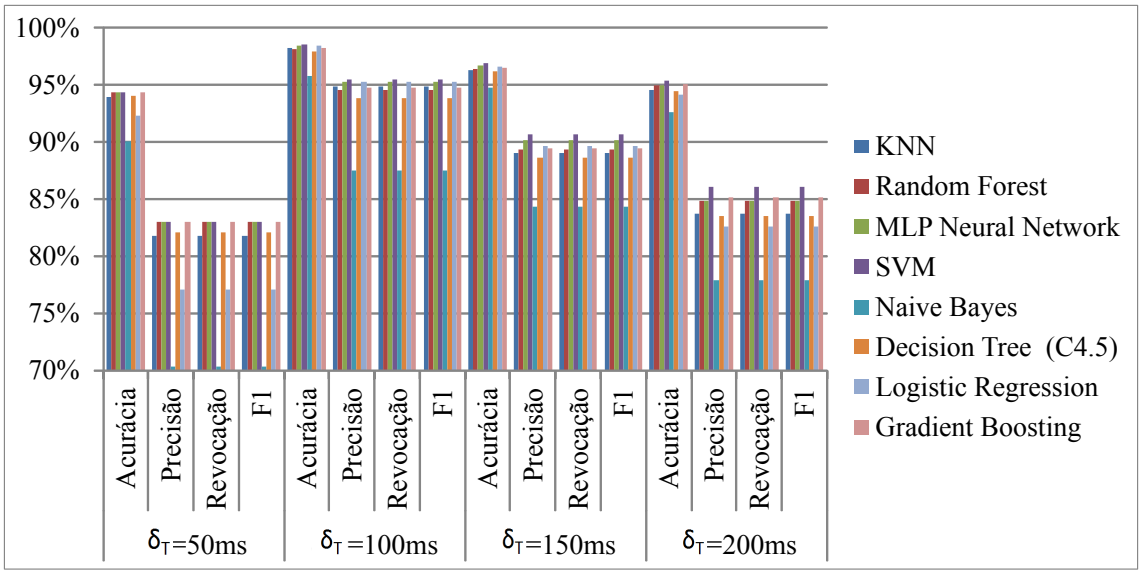

Figura 3. Indicadores de Desempenho dos Modelos de Classificação Agrupados por Intervalo de *Coleta*

O intervalo de coleta $\delta_T = 50ms$ apresentou os resultados mais modestos entre os 4 intervalos de coleta, sugerindo que tal intervalo de tempo pode não ser grande o suficiente para extrair *fingerprints* com características que permitam aos modelos de classificação alcançar o máximo de sua eficácia na identificação de estações terrenas responsáveis pela emissão de sinais interferentes.

O próximo intervalo de coleta, $\delta_T = 100ms$, obteve os melhores resultados nos indicadores avaliados dentre os quatro *datasets*. Com exceção do Naïve Bayes, os demais modelos obtiveram uma acurácia de aproximadamente 98% e ficaram acima de 93% para as demais métricas de desempenho. Os resultados da Revocação demonstraram uma boa capacidade dos modelos em conseguir identificar a grande maioria dos sinais emitidos por cada estação terrena. De forma análoga, a métrica de precisão, neste cenário, mostra que, em média, os modelos tiveram sucesso na identificação das estações responsáveis pela emissão da maior parte das instâncias de sinais avaliadas.

O intervalo de coleta $\delta_T = 150ms$ obteve a segunda colocação, quando considerados os resultados dos quatro *datasets*. Para a maior parte dos modelos deste intervalo, a acurácia permaneceu acima de 96%, uma diferença de menos de 2 p.p. em relação aos melhores resultados globais. O fato dos intervalos de coleta $\delta_T = 100ms$ e $\delta_T = 150ms$ liderarem com os melhores resultados obtidos sugere que o método proposto neste trabalho pode atingir melhores desempenhos com intervalos de coleta de durações intermediárias.

Já o intervalo de coleta $\delta_T = 200ms$, por sua vez, exibiu uma leve melhora nos resultados em relação ao intervalo $\delta_T = 50ms$ (o pior global), porém, sem superar os intervalos $\delta_T = 100ms$ (o melhor) e $\delta_T = 150ms$ (o segundo colocado). De forma análoga aos demais intervalos, os modelos gerados pelos algoritmos K-NN, *Random Forest*, MLP *Neural Network*, SVM, *Decision Tree*, *Logistic Regression* e *Gradient Boosting* no intervalo $\delta_T = 200ms$ permaneceram com indicadores de desempenho próximos entre si.

Por fim, é importante comentar que os dados coletados com o intervalo de $100ms$ levaram a uma acurácia de 98,51% para o modelo gerado pelo algoritmo SVM, e um valor acima de 95% para os demais, sendo esses os melhores resultados obtidos em todos os cenários. Embora o desempenho deste algoritmo tenha sido muito próximo aos dos demais algoritmos no mesmo intervalo, é importante ressaltar que mesmo ganhos pequenos de desempenho na identificação de estações interferentes podem significar economias na priorização das estações terrenas a serem inspecionadas. De toda forma, em geral, todos os resultados obtidos nos experimentos realizados apontam para a adequação do uso de modelos de classificação para identificar as estações responsáveis pela emissão de sinais a partir do *fingerprint* desses sinais. Reforçam, portanto, o conjunto de evidências que corroboram a validade da hipótese levantada neste trabalho.

6. Considerações Finais

Atualmente as redes com transmissão via Satélite são essenciais em todo o mundo e, em particular, no Brasil, onde se constituem no único meio de levar conectividade a diversas regiões de difícil acesso. Tais redes utilizam comunicação sem fio e são afetadas por sinais interferentes (e.g., interferências e uso indevido da faixa de radiofrequência). Um problema relevante neste contexto é identificar a estação terrena interferente (i.e., a estação responsável pela emissão desses sinais).

A principal técnica para identificar estações interferentes é baseada em geolocalização. Tal técnica se limita a apresentar uma relação de estações suspeitas de terem emitido o sinal interferente, cabendo à operadora de telecomunicações a onerosa tarefa de inspeção de cada estação a fim de identificar a responsável.

Diante do exposto, o presente trabalho propôs um método que pode contribuir para reduzir a demanda de inspeção, ao identificar a estação interferente a partir das estações suspeitas elencadas pela técnica da geolocalização. O método proposto, principal contribuição deste trabalho, utiliza modelos de classificação aplicados a características de *Radio Frequency Fingerprint* extraídas dos sinais. Tal método alcançou acurácia superior a 98% em experimentos com dados reais envolvendo 64.800 amostras de sinais, coletadas em intervalos de 50, 100, 150 e 200 ms, a partir de 6 estações terrenas de uma rede de satélite geoestacionário, o que sugere a adequação do método desenvolvido.

Para obter a quantidade de sinais estudados foi utilizada uma infraestrutura de rede real de alta complexidade e alto custo, além de envolver vários profissionais, tanto do

governo quanto da operadora de telecomunicações. Foram utilizados seis sinais que correspondem a seis estações terrenas (modeladas como seis classes), mas com um total de 51.904.800 amostras, sendo 8.650.800 amostras de cada sinal de cada estação terrena. É importante observar que os resultados obtidos nos experimentos realizados demonstraram a viabilidade da proposta e fornecem argumentos para apoiar a obtenção de investimentos que permitam o prosseguimento e a ampliação desta pesquisa.

Cabe mencionar ainda que este trabalho envolveu as áreas de Redes (problema), Banco de Dados (dados coletados, tratados e persistidos em uma organização adequada à solução) e Aprendizado de Máquina (modelos de classificação), apresentando uma solução de pesquisa para um problema do mundo real oriundo da indústria/Governo.

Sugere-se para trabalhos futuros a análise de uma maior quantidade de estações e sinais, bem como das alterações que o *fingerprint* pode sofrer, seja pelo tempo em decorrência do desgaste natural dos equipamentos dos transmissores, ou pelas variações do clima (chuva, neve, granizo, etc). Também existe espaço para buscar a otimização dos hiper-parâmetros dos algoritmos utilizados na classificação (incluindo uma Etapa de *Calibragem*), e viabilizar a análise da sensibilidade.

Os autores também pretendem abrir uma frente de pesquisa voltada à investigação das razões pelas quais as medidas de desempenho obtidas na *Coleta* com intervalos de 100ms terem sido superiores às demais. Para tanto, levantam a hipótese de que as condições climáticas e os níveis de poluição na hora das medições possam ter favorecido os resultados observados.

Por fim, pretende-se realizar um estudo do desempenho de modelos de classificação binária por estação terrena, em substituição aos modelos de classificação utilizados neste trabalho. Tal abordagem deverá facilitar a inclusão de novas estações terrenas na solução, uma vez que não demandará sucessivos treinamentos de um único modelo de classificação de estações sempre que novas estações sejam consideradas. Além disso, uma terceira abordagem envolvendo modelos de clusterização de dados para detecção de anomalias está prevista entre os próximos passos da investigação.

Agradecimentos. Os autores agradecem à Embratel por gentilmente transmitirem os sinais analisados nesta pesquisa e à ANATEL pelo apoio na disponibilização dos equipamentos de medição profissionais.

Referências

Aggarwal, C. C. et al. (2015). *Data mining: the textbook*, volume 1. Springer.

Azevedo, J., Barcellos, A. L., Mendes, A. C., de Oliveira, D., Vidal, P. C., and Bedo, M. (2021). Sat-espec: Análise e coleta de dados da transmissão de estações terrenas de uma rede satélite. In *Anais do III Dataset Showcase Workshop*, pages 43–52. SBBD.

Bitar, N., Muhammad, S., and Refai, H. H. (2018). Wireless technology identification using deep convolutional neural networks. *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 1:1–6.

Brik, V. et al. (2008). Wireless device identification with radiometric signatures. *14th ACM international conference on Mobile computing and networking*, 1:116–127.

Deng, S. et al. (2017). Radio frequency fingerprint extraction based on multidimension permutation entropy. *International Journal of Antennas and Propagation*, 2017:.

Faceli, K. et al. (2021). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.

Gahlawat, S. (2020). Investigation of rf fingerprinting approaches in gnss. Master's thesis, Tampere University, Faculty of Information Technology, Finlândia. 2021-06-17.

Hao, C. et al. (2020). Interference geolocation in satellite communications systems: An overview. *IEEE Vehicular Technology Magazine*, 16(1):66–74.

Henarejos, P. et al. (2019). Deep learning for experimental hybrid terrestrial and satellite interference management. *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 1:1–5.

Huang, H. et al. (2020). Fcn-based carrier signal detection in broadband power spectrum. *IEEE Access*, 8:113042–113051.

Kennedy, I. et al. (2008). Radio transmitter fingerprinting: A steady state frequency domain approach. *2008 IEEE 68th Vehicular Technology Conference*, 1:1–5.

Laignier, P. and Fortes, R. (2009). *Introdução à história da comunicação*. Ed. E-papers.

Li, Y., Chen, X., Lin, Y., Srivastava, G., and Liu, S. (2020). Wireless transmitter identification based on device imperfections. *IEEE Access*, 8:59305–59314.

Pratt, T. and Allnutt, J. (2020). *Satellite Communication*. 2020 John Wiley & Sons Ltd, Virginia, USA, 3 edition.

Smith, R. et al. (2021). An o-ran approach to spectrum sharing between commercial 5g and government satellite systems. In *MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM)*, pages 739–744. IEEE.

Soltanieh, N. et al. (2020). A review of radio frequency fingerprinting techniques. *IEEE Journal of Radio Frequency Identification*, 4(3):222–233.

UIT (2010). Use of appendix 10 of the radio regulation to convey information related to emissions from both gso and non-gso space stations including geolocation information. 2022-05-31.