



APRENDIZADO DE MÁQUINA APLICADO PARA RECONHECIMENTO PROBABILÍSTICO DE VAPORES QUÍMICOS COM SENSORES MQ

Lucas S. C. Teixeira¹; Roberto S. S. Filho¹; Laura C. Guimarães¹; Queli A. R. Almeida¹; Bianca Ortiz-Silva²

¹ Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro, Campus Duque de Caxias, Rio de Janeiro, Brasil.

² Universidade Federal do Rio de Janeiro, Núcleo de Enfrentamento e Estudos de Doenças Infecciosas Emergentes e Reemergentes, Rio de Janeiro, Brasil.

E-mail para correspondência: bianca.ortiz@ccsdecania.ufrj.br

Palavras-Chave: biossegurança, sensores MOS, nariz eletrônico

Introdução

Sistemas de monitoramento com sensores de baixo custo geram grande quantidade de leituras, mas o simples acúmulo de números não produz, por si só, uma informação útil para a segurança química. O desafio central passa a ser interpretar simultaneamente sinais que variam entre dispositivos, sofrem influência do tempo de aquecimento e podem responder a mais de uma substância. Essa dificuldade é particularmente importante nos sensores semicondutores de óxido metálico, denominados sensores MOS, nos quais a interação do vapor com uma camada aquecida modifica a resistência elétrica do material. Esses componentes são acessíveis e sensíveis, porém apresentam sensibilidade cruzada, isto é, um mesmo sensor reage a diferentes vapores e não identifica sozinho uma substância específica (WILSON; BAIETTO, 2009).

O IFORG foi inicialmente desenvolvido como uma plataforma para registrar alterações atmosféricas em armários de reagentes e evoluiu para uma infraestrutura de aquisição distribuída. O termo distribuído significa que a coleta não depende de um único equipamento: diferentes protótipos podem medir setores distintos e encaminhar os dados para uma central. Cada protótipo utiliza o ESP32, um microcontrolador programável com conexão sem fio e entradas para sinais analógicos, associado aos sensores MQ-2, MQ-3, MQ-4, MQ-5, MQ-7 e MQ-135. O sistema registra o horário, a identificação do equipamento, a etiqueta experimental e as leituras de cada canal, formando um banco que pode ser posteriormente analisado sem interromper as coletas. Essa evolução amplia o trabalho apresentado no CBQ 2025, no qual o monitoramento eletrônico foi aplicado à gestão de risco em armários de reagentes (TEIXEIRA et al., 2025).

Neste trabalho, a baixa seletividade individual foi tratada como fonte de informação. As seis respostas simultâneas formam uma assinatura multivariada dos sensores utilizados, isto é, um conjunto de valores cuja combinação descreve melhor o evento do que qualquer canal isolado. Um nariz eletrônico é justamente um sistema que reúne um arranjo de sensores, uma etapa de aquisição e um método de reconhecimento de padrões para comparar essas assinaturas. Estudos com sensores MOS demonstram que algoritmos de aprendizado de máquina, programas capazes de aprender relações a partir de exemplos etiquetados, podem diferenciar compostos orgânicos voláteis mesmo quando os sensores possuem respostas sobrepostas (HUANG; DOH; BAE, 2021; OATES et al., 2022; SINGH et al., 2024).

O objetivo foi desenvolver e avaliar uma camada computacional de aprendizado de máquina para o IFORG, sem depender inicialmente da instalação do modelo nos protótipos. Dados de exposições controladas foram usados para comparar classificadores, isto é, algoritmos que atribuem cada conjunto de leituras a uma classe conhecida. Em seguida, o modelo foi



aplicado a uma aula de recristalização que permaneceu totalmente fora do treinamento. Buscou-se verificar se a normalização individual dos sensores, a análise conjunta dos canais e a organização temporal das leituras permitiriam: reconhecer padrões associados aos reagentes controlados; aproximar uma substância não treinada da família química mais semelhante; resumir milhares de valores em hipóteses compreensíveis; e criar uma base para alertas, documentação de eventos e diagnóstico do funcionamento dos próprios sensores.

Material e Métodos

A plataforma foi formada por nós sensores, isto é, protótipos de ESP32 equipados com MQ-2, MQ-3, MQ-4, MQ-5, MQ-7 e MQ-135, destinados a coletar os sinais e enviá-los pela rede sem fio. Os dados foram recebidos por um nó servidor, isto é, outro ESP32 configurado como central para reunir os pacotes, disponibilizar um painel em navegador e registrá-los no Google Planilhas. O conversor analógico-digital, ou ADC, transformou a tensão elétrica de cada sensor em números inteiros de 0 a 4095. Esses valores brutos foram preservados como medida principal, pois mantêm a informação original e permitem aplicar diferentes tratamentos posteriormente. Para cada combinação entre protótipo e canal foi calculada uma linha de base em ar limpo, definida como a resposta de referência antes das exposições, e as demais leituras foram expressas também como razão e variação percentual em relação a essa condição.

Os ensaios controlados ocorreram no interior de uma capela química, equipamento de exaustão que captura vapores próximos à fonte, com os protótipos a cerca de 10-15 cm dos recipientes, temperatura ambiente de aproximadamente 25 °C, umidade relativa entre 60% e 70% e aquecimento prévio de 30 minutos. Foram avaliados os seguintes reagentes e seus volumes: éter etílico (1,0 mL), diclorometano (1,0 mL), acetona (1,25 mL), metanol (2,25 mL), acetato de etila (3,0 mL), ciclo-hexeno (3,25 mL) e etanol (5,0 mL), respeitando alterações registradas nas etiquetas. Linhas sem identificação segura foram excluídas. A aula foi marcada no banco como NAO_TREINO, bloqueada antes dos ajustes e usada somente na aplicação externa. O aprendizado reuniu 326 medições etiquetadas de S1F1, S4D2 e S7F2. Como havia diferenças de escala e canais zerados ou saturados, a análise principal foi concentrada no S4D2, com 109 registros distribuídos nas sete classes.

As séries foram divididas em 45 janelas móveis, isto é, blocos de quatro leituras que avançavam duas posições. De cada janela foram extraídas 64 características: médias, dispersões, amplitudes, inclinações, valores finais, respostas relativas ao ar limpo, participação de cada MQ, taxas de zero e saturação e quantidade de canais ativos. Compararam-se regressão logística, algoritmo que estima probabilidades por fronteiras matemáticas; Random Forest, conjunto de árvores de decisão que votam em conjunto; Extra Trees, variante com divisões mais aleatórias; e perceptron multicamadas, rede neural formada por camadas conectadas. A validação cruzada estratificada dividiu os dados em quatro partes, preservou a proporção das classes e alternou a parte usada para teste. O desempenho foi medido pela acurácia, percentual total de acertos, e pelo F1 macro, média que equilibra a qualidade das previsões entre todas as classes. Para a aula, ocorrida entre 13h40 e 15h45, foram analisadas 115 janelas não sobrepostas. A inferência off-line, aplicação do modelo em computador após a coleta, foi executada em Python sem modificar o firmware dos sensores.

Resultados e Discussão

A organização dos dados demonstrou que a principal contribuição inicial da inteligência artificial não foi apenas escolher o nome de um reagente, mas estabelecer um processo rastreável entre coleta, etiqueta, tratamento e previsão. Cada linha permaneceu associada ao



protótipo, horário e observação experimental, permitindo separar o que era exemplo de treinamento do que seria utilizado como teste externo. Essa separação evita vazamento de dados, situação em que o algoritmo recebe durante o treinamento informações do próprio conjunto usado para avaliá-lo e, por isso, aparenta desempenho maior do que realmente possui.

A auditoria dos canais identificou leituras zeradas, saturação próxima ao limite do ADC e diferenças de escala entre protótipos. Saturação significa que o canal atingiu o maior valor que o conversor consegue registrar, impedindo saber se o sinal continuaria aumentando. Esses comportamentos não foram tratados automaticamente como ausência ou presença intensa de vapor; foram incorporados como indicadores de qualidade. A normalização pela linha de base converteu cada sensor em uma resposta relativa ao seu próprio estado inicial. Assim, dois dispositivos não precisavam apresentar o mesmo número absoluto para que suas tendências fossem comparadas. Esse procedimento reduz parcialmente efeitos de fabricação, montagem, envelhecimento e alimentação elétrica, embora não substitua calibração química com atmosferas de concentração conhecida.

Tabela 1. Desempenho médio dos classificadores na validação cruzada estratificada $n = 109$.

Classificador	Acurácia média (%)	F1 macro
Regressão logística	62,5	0,585
Random Forest	60,4	0,538
Extra Trees	60,4	0,550
Rede neural MLP	58,1	0,517

A regressão logística apresentou a maior acurácia média, 62,5%, e F1 macro de 0,585. Random Forest e Extra Trees alcançaram 60,4% de acurácia, com F1 macro de 0,538 e 0,550, respectivamente, enquanto a rede neural alcançou 58,1% e F1 macro de 0,517. A proximidade entre os resultados indica que as assinaturas continham informação classificável, mas o tamanho reduzido do banco limitou a vantagem de modelos mais complexos. A regressão logística, mesmo sendo matematicamente mais simples, conseguiu estabelecer separações úteis entre as classes. A rede neural possui maior quantidade de parâmetros, isto é, valores internos ajustados durante o treinamento, e por isso ficou mais sensível ao pequeno número de janelas. Os resultados representam uma avaliação exploratória e não uma validação para medir exposição ocupacional.

O Random Forest foi escolhido para analisar a aula porque combina relações não lineares, isto é, relações que não podem ser representadas por uma única reta, e produz uma distribuição de probabilidades entre as classes. Essa distribuição é mais informativa do que uma resposta binária, pois uma atmosfera real pode conter misturas, resíduos de etapas anteriores ou substâncias ausentes do treinamento. A classe com maior probabilidade deve ser entendida como o padrão conhecido mais semelhante, e não como confirmação analítica. Desse modo, a IA atua como filtro de informação: reduz centenas de leituras a uma sequência de hipóteses que pode ser confrontada com o roteiro e com o horário das atividades.

Durante a aula de recristalização, o S4D2 apresentou aumento mediano de 132,8% no MQ-3 em relação ao ar limpo. O arquivo adicional do S2D2, utilizado somente como verificação independente da alteração ambiental e não como parte do treinamento principal, apresentou variações de 72,7% no MQ-2, 11,1% no MQ-3 e 45,5% no MQ-7. Entre as 115 janelas do S4D2, 85,2% foram associadas à classe CICLO_HEXENO e 14,8% à classe ACETATO_DE_ETILA. A sequência temporal indicou uma assinatura semelhante à de



hidrocarbonetos entre 13h40 e 13h54, aumento da compatibilidade com acetato de etila entre aproximadamente 13h55 e 14h12 e retorno posterior à região de hidrocarbonetos (Figura 1).

O roteiro da recristalização previa testes de solubilidade com água, etanol, acetona, acetato de etila e hexano. O hexano não estava entre as classes usadas no aprendizado, enquanto o ciclo-hexeno era o hidrocarboneto disponível no banco. Por isso, a saída CICLO_HEXENO foi interpretada como aproximação à família dos hidrocarbonetos, e não como indicação de que ciclo-hexeno estava necessariamente presente. Esse resultado funciona como teste de classe não conhecida: diante de um composto fora do treinamento, o modelo o aproximou de uma região quimicamente relacionada em vez de distribuir todas as previsões de forma aleatória, ao fazer isto o sistema avisa qual o nível de confiança da medição referente a sua correspondência com o aprendido no treinamento e em casos mais extremos onde não há uma correspondência aceitável é entendido como reagente desconhecido. A faixa associada ao acetato de etila coincidiu com um solvente previsto no procedimento, demonstrando que a análise temporal pode auxiliar na reconstrução aproximada das etapas de uma atividade real.

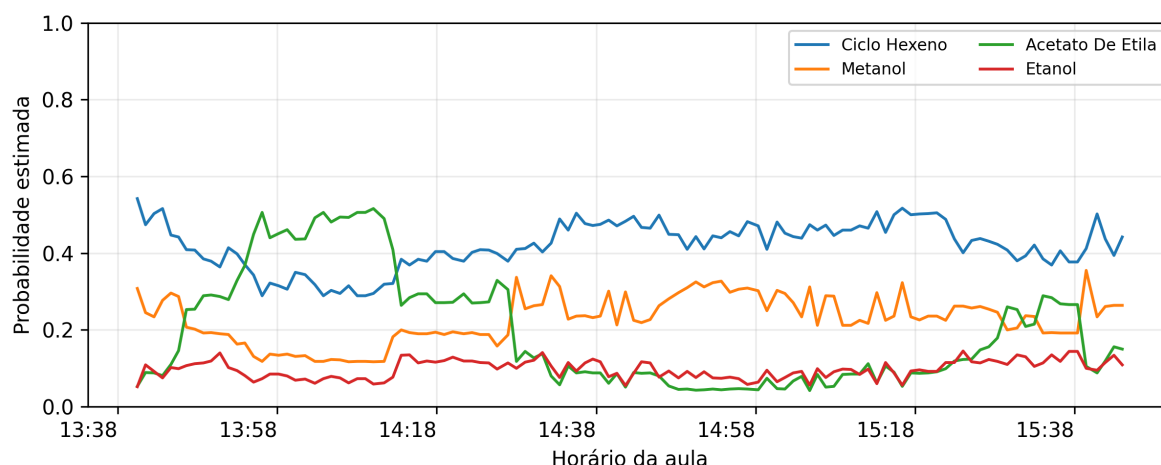


Figura 1. Probabilidades calculadas pelo Random Forest durante a aula. A classe ciclo-hexeno representa a assinatura de hidrocarboneto mais próxima do hexano, que não participou do treinamento.

A aplicação evidenciou utilidades diferentes das abordadas em estudos centrados apenas na ventilação do laboratório. Primeiro, a IA executa triagem automática, isto é, seleciona os intervalos mais relevantes em uma série extensa para posterior inspeção humana. Segundo, produz uma memória digital do evento, reunindo duração, classe mais semelhante e comportamento dos canais. Terceiro, identifica padrões de falha instrumental, como sensor constantemente zerado ou saturado, auxiliando a manutenção do conjunto. Quarto, permite testar diferentes algoritmos em computador antes da implantação embarcada, nome dado à execução direta no microcontrolador. Essa estratégia evita reprogramar todos os protótipos durante o desenvolvimento e possibilita que, após validação, um modelo compacto seja transferido para o servidor ou para o ESP32, onde poderá emitir alertas locais com menor dependência de conexão externa.

As limitações permaneceram explícitas. As classes controladas resultaram de uma única bateria, não houve recuperação formal em ar limpo entre todos os reagentes e alguns canais apresentaram funcionamento parcial. O roteiro da aula registra os solventes utilizados, mas não fornece a concentração instantânea no ar nem o momento exato de cada manipulação. Portanto, as probabilidades não representam partes por milhão, unidade usada para concentração, e não devem ser convertidas em declaração de exposição. O sistema também não substitui



cromatografia, detectores certificados ou avaliação profissional. Seu valor atual é organizar dados, reconhecer semelhanças, localizar transições e gerar hipóteses verificáveis. O aprimoramento deverá incluir novas repetições independentes, tratamento de misturas e validação entre singular para cada protótipo.

Conclusões

O aprendizado de máquina acrescentou ao IFORG uma etapa de interpretação que não existia no registro convencional das leituras. A combinação dos seis canais foi tratada como assinatura, a linha de base individual reduziu diferenças de escala e os indicadores de qualidade impediram que falhas elétricas fossem confundidas automaticamente com resposta química. Na avaliação interna, a regressão logística apresentou a maior acurácia média, 62,5%, enquanto os métodos formados por árvores alcançaram desempenho próximo e forneceram probabilidades adequadas à análise temporal.

Na aplicação externa, preservada fora do treinamento, a IA indicou uma faixa compatível com acetato de etila e predominância de uma região de hidrocarbonetos. Como o hexano previsto no roteiro não integrava as classes, sua aproximação com o ciclo-hexeno demonstrou uma utilidade importante: indicar a família conhecida mais semelhante quando a substância exata não está disponível no banco. Essa saída não comprova identidade química, mas facilita a triagem de eventos e a comparação entre o registro eletrônico e a atividade realizada.

A execução em computador confirmou que a camada inteligente pode ser desenvolvida e avaliada antes de sua instalação no ESP32. O trabalho diferencia-se de uma avaliação de ventilação porque seu foco está no processamento dos dados, na comparação de algoritmos e na transformação das leituras em apoio à decisão. O conjunto atual sustenta a viabilidade da proposta como ferramenta de documentação, seleção de intervalos críticos e preparação de alertas. Para uso autônomo, ainda serão necessárias validações independentes, ampliação das classes, reconhecimento de misturas e critérios conservadores para rejeitar previsões incertas.

Agradecimentos

À FAPERJ, ao CNPq, à UFRJ e ao IFRJ pelo apoio institucional, científico e tecnológico.

Referências

- HUANG, Y.; DOH, I.-J.; BAE, E. Design and validation of a portable machine learning-based electronic nose. *Sensors*, v. 21, n. 11, 3923, 2021. Link: [acessar referência](#)
- OATES, M. J.; GONZÁLEZ-TERUEL, J. D.; RUIZ-ABELLÓN, M. C.; et al. Using a low-cost components e-nose for basic detection of different foodstuffs. *IEEE Sensors Journal*, v. 22, n. 14, p. 13872-13881, 2022. Link: [acessar referência](#)
- SINGH, S.; SAJANA, S.; VARMA, P.; et al. Metal oxide-based gas sensor array for VOCs determination in complex mixtures using machine learning. *Microchimica Acta*, v. 191, 196, 2024. Link: [acessar referência](#)
- TEIXEIRA, L. S. C.; FILHO, R. S. S.; ANNUNCIAÇÃO, G. P. M. M.; et al. Gestão de risco químico em armários de reagentes de laboratórios de pesquisa e ensino. *Anais do 64º Congresso Brasileiro de Química*, 2025. Link: [acessar referência](#)
- WILSON, A. D.; BAIETTO, M. Applications and advances in electronic-nose technologies. *Sensors*, v. 9, n. 7, p. 5099-5148, 2009. Link: [acessar referência](#)