

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
INSTITUTO DE BIOCÊNCIAS DE BOTUCATU

Utilização de machine learning para avaliação dos efeitos não
tóxicos em fibroblastos murinos expostos *in vitro* a
nanopartículas de poliestireno

FERNANDA ELIAS MENDES

PROF. DR. DANIEL PEREZ VIEIRA

Trabalho de Conclusão de Curso apresentado ao
Instituto de Biotecnologia, Campus de Botucatu,
UNESP, para obtenção de Bacharel em Ciências
Biomédicas.

BOTUCATU – SP
2025

M538u

Mendes, Fernanda

Utilização de machine learning para avaliação dos efeitos não tóxicos em fibroblastos murinos expostos in vitro a nanopartículas de poliestireno / Fernanda Mendes. -- Botucatu, 2025

29 p. : tabs., fotos

Trabalho de conclusão de curso (Bacharelado - Ciências Biomédicas) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu

Orientador: Daniel Vieira

1. Toxicologia. 2. Nanopartículas. 3. Aprendizado de máquina. I. Título.

FERNANDA ELIAS MENDES

**UTILIZAÇÃO DE MACHINE LEARNING PARA AVALIAÇÃO DOS
EFEITOS NÃO TÓXICOS EM FIBROBLASTOS MURINOS EXPOSTOS
IN VITRO A NANOPARTÍCULAS DE POLIESTIRENO**

Trabalho de Conclusão de Curso,
apresentado a Universidade Estadual
Paulista, como parte das exigências para
a obtenção do título de Bacharel, do curso
de Graduação em Ciências Biomédicas.

São Paulo, 9 de dezembro de 2025.

BANCA EXAMINADORA



Documento assinado digitalmente

DANIEL PEREZ VIEIRA

Data: 17/12/2025 15:00:41-0300

Verifique em <https://validar.it.gov.br>

Prof. Dr. Daniel Perez Vieira
Instituto de Pesquisas Energéticas e Nucleares (IPEN)



Documento assinado digitalmente

FELIPE LOPES BRUM DA SILVEIRA

Data: 17/12/2025 12:21:28-0300

Verifique em <https://validar.it.gov.br>

Prof. Dr. Felipe Lopes Brum
Instituto UFPB de Desenvolvimento da Paraíba (IDEP)

AGRADECIMENTOS

Agradeço ao Daniel Perez, meu orientador, pela generosidade com a qual me recebeu em seu laboratório, pela disponibilidade e grande apoio na realização deste trabalho.

Aos meus colegas do Laboratório de Radiobiologia, especialmente Esther, Thais e Suleyna, por toda a ajuda, trocas de conhecimento e pelas conversas que tornaram meu dia a dia no laboratório mais agradável e prazeroso.

À minha família, sobretudo aos meus pais, Paulo e Luciane, pelo apoio e incentivo constantes.

Aos amigos que fiz este ano pela USP e às minhas amigas de São Paulo, por estarem sempre por perto e por me apoiarem nos momentos de angústia. Vocês tornaram este período imensamente mais leve e divertido.

Ao Daniel Baracho, pelo carinho e confiança que sempre demonstrou.

E, por fim, aos meus amigos de Botucatu, que foram casa durante os últimos três anos de graduação, e com quem tive a alegria de compartilhar uma das melhores épocas da minha vida.

SUMÁRIO

1. INTRODUÇÃO	7
2. METODOLOGIA.....	10
2.1 Cultura celular	10
2.2 Nanosferas de Poliestireno (nPS).....	10
2.3 Ensaio de Citotoxicidade/Viabilidade Celular	11
2.4 Avaliação temporal de parâmetros quantitativos, forma e intensidade dos núcleos das células expostas a nPS.....	11
2.5 Análise dos resultados	12
2.6 Modelo de Random Forest e Gradient Boosting para detectar reações celulares a nPS.....	12
3. RESULTADOS E DISCUSSÃO	15
3.1 Comparação entre os modelos de machine learning Gradient Boosting e Random Forest utilizando todas as 91 variáveis.....	15
3.2 Comparação entre os modelos de machine learning Gradient Boosting e Random Forest para nPS de 10 ou 15nm usando apenas as variáveis mais bem classificadas.....	20
4. CONCLUSÕES	25
REFERÊNCIAS BIBLIOGRÁFICAS	27

RESUMO

O crescimento da produção global de plásticos consolidou-se como uma grande preocupação ambiental. Micro e nanopartículas liberadas a partir do acúmulo de seus resíduos já foram identificadas em diversos organismos e ecossistemas, evidenciando sua ampla dispersão no ambiente. Entre os polímeros mais utilizados mundialmente destaca-se o poliestireno (PS), que, apesar de corresponder a 10% da produção mundial de plásticos, resulta em maior acúmulo de partículas no ambiente quando comparado a outros tipos de polímeros, como polipropileno e polietileno, por ser amplamente utilizado na produção de consumíveis comuns (i.e embalagens e utensílios descartáveis). Estudos *in vitro* sobre nanopartículas de poliestireno (nPS) concentram-se majoritariamente em efeitos citotóxicos imediatos, avaliados por ensaios de viabilidade, produção de espécies reativas de oxigênio (ROS), dano ao DNA e alterações mitocondriais. Contudo, mudanças sutis na morfologia e na fluorescência celular, indicativas de processos precoces e não necessariamente letais, permanecem pouco investigadas, tornando abordagens como High-Content Screening (HCS) e machine learning (ML), estratégias promissoras para identificá-las. Nesse contexto, este estudo utiliza técnicas de HCS e ML para analisar alterações não tóxicas em fibroblastos murinos (NIH/3T3) expostos *in vitro* a nPS. Os resultados visam aprofundar a compreensão dos riscos associados a esse material e a fortalecer a discussão sobre seus impactos ambientais e de saúde pública.

Palavras-chave: nanoplásticos; machine learning (ML), High-Content Screening (HCS).

1. INTRODUÇÃO

Exercícios teóricos de idealizar sociedades ocidentalizadas e modernas sem o uso de plásticos são difíceis. A onipresença desses materiais resulta de suas propriedades intrínsecas, como elevada durabilidade, leveza, facilidade de fabricação e, principalmente, de seu baixo custo (WILLIAMS; RANGEL-BUITRAGO, 2022). Essas características, em conjunto, levaram à profunda incorporação desses polímeros sintéticos em praticamente todos os aspectos da vida cotidiana, desde embalagens e produtos de consumo até materiais utilizados na construção civil e em processos industriais (MITRANO; WICK; NOWACK, 2021).

Estima-se que, em 2024, a produção mundial de plásticos ultrapassou 430 milhões de toneladas - volume que cresce anualmente -, das quais apenas 41,2 milhões de toneladas foram efetivamente recicladas (PLASTICS EUROPE, 2025). A combinação entre a baixa eficiência dos sistemas de reciclagem e o aumento contínuo da demanda por produtos plásticos, grande parte deles projetada para uso único e com reciclabilidade limitada (WALKER; FEQUET, 2023), intensifica o acúmulo desses materiais no ambiente, sobretudo nos ecossistemas marinhos e de água doce. Nesse contexto, Borrelle et al. identificaram que entre 19 a 23 Mt dos resíduos plásticos gerados globalmente em 2016 alcançaram ambientes aquáticos, com projeções indicando que esse volume pode atingir até 53 Mt anuais até 2030.

Nesses locais, eles são transportados pelas correntes, acumulam-se nos sedimentos e interagem com organismos de diferentes níveis tróficos. Esses polímeros persistem por longos períodos no ambiente, sendo continuamente submetidos a processos físicos, químicos e biológicos que favorecem sua desintegração (POOJA et al., 2023). A exposição prolongada a esses fatores resulta na formação de partículas progressivamente menores, denominadas de micro (1-5mm) e nanoplásticos (<100 nm). Os critérios de classificação de suas dimensões ainda estão em debate na literatura científica (WALKER; FEQUET, 2023), mas já se sabe que além da fragmentação natural, essas partículas também podem derivar da liberação direta de produtos de consumo (ZANGMEISTER et al., 2022), fato este que amplia sua disseminação nos sistemas aquáticos.

Em escala global, os principais polímeros produzidos são o polietileno (PE, 36% do total), o polipropileno (PP, 21%), e o policloreto de vinila (PVC, 12%), seguidos pelo tereftalato de polietileno (PET), poliuretano (PUR) e poliestireno (PS), cada um representando menos de 10% da produção total (GEYER; JAMBECK; LAW, 2017). Apesar de sua menor representatividade quantitativa, o PS tem se destacado como contaminante emergente, especialmente em áreas costeiras e centros urbanos (KIK; BUKOWSKA; SICIŃSKA, 2020), na medida em que sua alta fragmentação favorece a formação de micro e nanopartículas (KELPSIENE; CEDERVALL; MALMENDAL, 2023). Além disso, suas propriedades físico-químicas, como a estrutura superficial porosa e hidrofóbica, conferem-lhe notável capacidade de adsorver compostos orgânicos persistentes, como pesticidas e hidrocarbonetos aromáticos policíclicos. Essa afinidade faz com que o PS atue como vetor importante para o transporte desses contaminantes dentro dos ecossistemas, aumentando sua biodisponibilidade e potencial de toxicidade.

Em escala micro e nanométrica, o PS demonstra capacidade de atravessar barreiras biológicas e acumular-se em diferentes tecidos, fenômeno já evidenciado por estudos que identificaram microplásticos em tecido adiposo e órgãos humanos como pulmões, fígado, rins (BANERJEE; BILLEY; SHELVER, 2021). A internalização dessas partículas pode ocorrer por múltiplas vias, incluindo endocitose e difusão passiva, o que possibilita sua interação direta com componentes celulares e a interferência em processos fisiológicos essenciais (MAHMUD et al., 2024). Uma vez no interior das células, o PS tem sido associado à indução de estresse oxidativo, danos ao DNA e alterações na expressão gênica, sugerindo potenciais efeitos citotóxicos e genotóxicos (SKABA et al., 2025).

Todavia, a caracterização dos efeitos biológicos induzidos por essas partículas ainda representa um desafio metodológico. Embora estudos *in vitro* sejam amplamente utilizados para investigar respostas celulares frente à exposição às nPS, a interpretação desses resultados é frequentemente limitada pela natureza simplificada dos ensaios tradicionais e pela ausência de ferramentas capazes de captar a complexidade das respostas biológicas. Testes clássicos de viabilidade celular, como os baseados na redução de sais de tetrazólio (MTT, MTS, XTT), permanecem amplamente utilizados, mas sua aplicação a nanopartículas é marcada

por inconsistências. Em especial, há interferência direta entre as partículas e os reagentes, o que leva à superestimação da viabilidade celular e compromete a confiabilidade dos resultados. Além disso, tais ensaios fornecem apenas medidas pontuais, incapazes de refletir o comportamento dinâmico e multifatorial das células diante da exposição prolongada ao material particulado (KONG et al., 2011).

Nesse cenário, as técnicas de HCS surgem como alternativa robusta e eficiente para avaliar as respostas celulares complexas à exposição a nanopartículas em modelos *in vitro* (ZHANG et al., 2023). Essa abordagem combina microscopia automatizada e análise quantitativa de imagem, permitindo mensurar simultaneamente diversos parâmetros morfológicos e funcionais em células individuais. Diferentemente dos ensaios convencionais, o HCS permite mapear fenômenos celulares interdependentes, gerando um retrato muito mais completo das respostas biológicas. A natureza multiparamétrica da técnica oferece um potencial inestimável para identificar mudanças sutis, inclusive em condições consideradas não citotóxicas.

Considerando o volume e a complexidade dos dados gerados pelos experimentos de HCS, torna-se essencial empregar abordagens analíticas avançadas que superem as limitações das estatísticas convencionais. Nesse contexto, técnicas de ML destacam-se pela capacidade de reconhecer padrões não lineares e correlações sutis entre variáveis celulares. Entre os algoritmos disponíveis, o Random Forest (RF) aprimora a robustez dos modelos ao gerar múltiplas árvores de decisão independentes, reduzindo o risco de sobreajuste (*overfitting*) (ZHANG et al., 2025), o que o torna ideal para dados de alta dimensionalidade e variabilidade experimental. Já o Gradient Boosting (GB) constrói modelos de forma sequencial, refinando as previsões ao corrigir iterativamente os erros das etapas anteriores (BLOCKKEEL et al., 2023).

Portanto, o presente trabalho propõe integrar a análise multiparamétrica do HCS a esses métodos de ML para avaliar como ocorre, ao longo do tempo, a interação entre nPS e células de fibroblastos murinos *in vitro*. Essa abordagem busca identificar e compreender alterações celulares associadas à exposição prolongada, na tentativa de oferecer uma visão mais aprofundada das respostas a nível celular e contribuir para os estudos de modelos experimentais mais realistas e preditivos em toxicologia.

2. METODOLOGIA

Nesta seção, são apresentados os procedimentos experimentais adotados para a condução deste estudo. Todos os métodos foram realizados de forma padronizada, visando assegurar a reprodutibilidade dos resultados e a confiabilidade das análises subsequentes.

2.1 Cultura celular

Fibroblastos murinos, linhagem NIH/3T3, foram cultivados em 5 mL de meio RPMI 1640, pH 7,4 (Gibco, Grand Island, EUA), com suplementação de 10% de soro fetal bovino (Cultilab, Campinas, Brasil), 2 g/L de bicarbonato de sódio (NaHCO_3 , Sigma-Aldrich, Brasil) e 1% de solução de penicilina e estreptomicina (10.000 U/10.000 $\mu\text{g/mL}$, Gibco, Carlsbad, EUA), em frascos plásticos estéreis para cultura celular com 25 cm^2 de área de cultivo (Corning, Tewksbury, EUA), mantidos em incubadoras com atmosfera e temperatura controladas (5% de CO_2 , 37°C), até que a confluência estimada por inspeção visual atingisse de 60 a 70%, com trocas de meio a cada 48 horas. Após esse período, o meio de cultura foi removido, as células aderidas foram lavadas com solução salina tamponada com fosfato (PBS) estéril a 37°C e tratadas com solução de tripsina a 0,25% contendo 0,02% de EDTA tetrassódico (Sigma-Aldrich, Brasil) para obtenção de suspensões celulares completamente dissociadas. Apenas as frações viáveis, determinadas pela ausência de incorporação do corante vital azul de tripano a 0,4% (Life Technologies, Carlsbad, EUA), foram consideradas para subcultura ou experimentação.

2.2 Nanosferas de Poliestireno (nPS)

Partículas de poliestireno de diâmetro de 15 nm e de 10 nm foram adquiridas do fornecedor Lab261 (Palo Alto, CA). As esferas estavam em suspensões com concentração de 10 mg/mL, com quantidade não especificada de surfactante e 2 mM de azida sódica como ativo antimicrobiano. As suspensões foram sonicadas em banho de ultrassom por 5 minutos à temperatura ambiente para garantir a dissociação e, em seguida, diluídas em meio de cultura completo (contendo soro e antibióticos) para

obtenção das concentrações de teste. Antes de serem adicionadas aos poços, cada solução foi agitada em vortex por 10 segundos para garantir a homogeneidade.

2.3 Ensaio de Citotoxicidade/Viabilidade Celular

Para determinar as concentrações tóxicas e não tóxicas de nPS, as células foram tripsinizadas e semeadas (10.000 células/poço, 100 µL/poço) em uma placa de 96 poços e deixadas aderir por 24 horas em incubadora. Após esse período, o meio foi removido e 100 µL das soluções controle e das suspensões de nPS no meio foram adicionados em quadruplicata e mantidos em cultura por mais 24 horas. Os poços de controle positivo foram tratados com 10% de dimetilsulfóxido (DMSO, Sigma-Aldrich, Brasil) diluído em meio de cultura. Os controles negativos receberam, além do meio de cultura, NaCl na concentração final de 0,045%. Além disso, células cultivadas em meio RPMI 1640 sem a adição de qualquer composto foram utilizadas como controle basal (CC). Após 24 horas de incubação, o meio de cultura foi removido e as células foram lavadas com 100 µL de PBS (37°C) para remover eventuais nPS livres. Uma solução de MTS (2 mg/mL em PBS, Promega, Madison, WI) e PMS (0,92 mg/mL em PBS, CAS 299-11-6, Sigma-Aldrich) em meio foi adicionada (100 µL/poço). A leitura da absorbância (490 nm) foi realizada em um leitor de microplacas (Multiskan EX, Labsystems) 2 horas após a incubação.

2.4 Avaliação temporal de parâmetros quantitativos, forma e intensidade dos núcleos das células expostas a nPS

Placas de 96 poços foram preparadas conforme descrito e receberam diferentes concentrações (0,02; 0,01 e 0,005 mg/mL) de nPS dispersas em meio de cultura em octuplicata. O meio líquido também continha os corantes fluorescentes bis-benzimida (Sigma-Aldrich, CAS 875756-97-1) e SYTOX® Green (Thermo Fisher Scientific, número de catálogo S7020) nas concentrações 11 mM e 0,4 µM, respectivamente. A bis-benzimida (Hoechst 33342) é um corante pan-nuclear capaz de difundir-se através da membrana citoplasmática das células e que emite fluorescência em comprimentos de onda entre 430 e 460 nm (“azul”) quando excitada entre 340–360 nm. O SYTOX® Green, por outro lado, só consegue se difundir através de membranas de células em processo de morte em curso ou concretizado e emite fluorescência em 523 nm (“verde”) mediante excitação a 504 nm. As células foram

incubadas na câmara interna do incubador do equipamento para HCS (High Content Screening) INCell Analyzer 2500 HS (Cytiva Lifesciences), que manteve a temperatura a 37°C e atmosfera com 5% de CO₂ durante o experimento de 24 horas. A cada 60 minutos, foram adquiridas imagens das células em 9 campos por poço, em comprimentos de onda que permitiam a detecção dos sinais de bis-benzimida e do SYTOX® Green, totalizando 41.472 imagens (96 poços × 9 campos × 2 comprimentos de onda × 24 horas) por experimento. Foi utilizada a objetiva com aumento de 4X (Nikon 4X/o.20, Plan Apo, CFI/60), e as aquisições foram analisadas com o software INCarta 1.17 (Cytiva Lifesciences), que quantificou diversos indicadores quantitativos, de formato e de intensidade dos núcleos das células dos experimentos.

2.5 Análise dos resultados

A viabilidade celular de cada grupo experimental foi calculada em função do CC (% dos controles). Os valores de citotoxicidade calculados para cada amostra e controles foram comparados por meio da análise de variância unidirecional (One-way ANOVA) e as diferenças entre os grupos foram avaliadas utilizando a correção de Bonferroni. As concentrações e os diâmetros considerados tóxicos foram aplicados em novos experimentos, nos quais os dados foram ajustados à função sigmoideal dose-resposta para calcular a concentração inibitória de 50% das células (IC₅₀%). As análises foram realizadas utilizando o software GraphPad Prism® versão 10.2.0.

2.6 Modelo de Random Forest e Gradient Boosting para detectar reações celulares a nPS

Foram reunidas as aquisições de dois experimentos independentes em uma única tabela de valores separados por vírgulas (CSV). Utilizando o software Microsoft Excel 2019, calcularam-se as medianas dos controles (CC; células controle) para cada hora de experimento, o Desvio Absoluto da Mediana (MAD) e, por fim, o escore-Z robusto (RZS; *Robust Z-Score*). A tabela com os dados normalizados foi então analisada utilizando o software Orange Data Mining 3.39.0. Para a classificação dos dados, foram adicionadas duas novas variáveis categóricas: “Well Label”, contendo o diâmetro e as concentrações de nPS utilizadas nos respectivos poços, e “Feno”, relacionada ao fenótipo esperado após a presença ou ausência das exposições, conforme mostrado na Tabela 1.

Classificação	Fenótipo esperado ("Feno")
Viáveis	Células viáveis, expostas apenas ao meio de cultura (controle negativo)
Mortas	Células com morte induzida por DMSO a 10% no meio de cultura (controle positivo)
Mortas - Plástico	Células expostas a nPS a 0,02 mg/mL que apresentaram redução da viabilidade celular (ensaio MTS)
Sub Hi - Plástico	Células expostas a nPS a 0,01 mg/mL, sem indução significativa de toxicidade (ensaio MTS), consideradas em condição "subletal alta", próxima à concentração letal
Sub Lo - Plástico	Células expostas a nPS a 0,005 mg/mL, sem indução significativa de toxicidade (ensaio MTS), consideradas em condição "subletal baixa", mais distante da concentração letal

Tabela 1: Explicação da coluna "Feno" (variável categórica) usada para a classificação das amostras e para relacionar as variáveis morfológicas e de fluorescência ao seu significado biológico.

A remoção dos outliers foi realizada utilizando o algoritmo isolation forest (com 5% de contaminação) para reduzir os efeitos de valores atípicos, mesmo após a normalização pelo método RZS. Posteriormente, as amostras foram randomizadas e 10% do conjunto de dados foi completamente separado para testes futuros dos modelos treinados, não participando das análises finais a partir desse ponto. Os 90% restantes foram novamente divididos em 90/10: 90% dos dados foram usados para o treinamento do modelo e 10% para testes, uma vez que se optou por avaliar o desempenho dos modelos utilizando conjuntos de dados de treinamento e teste separados, em vez do método de validação cruzada.

As 91 variáveis numéricas, a variável “Tempo” e a variável categórica “Feno” foram aplicadas aos modelos RF e GB em uma primeira rodada de treinamento, na qual as variáveis foram classificadas de acordo com sua importância para cada modelo. Subsequentemente, as variáveis foram removidas dos modelos uma a uma (em ordem crescente de importância) até que a proporção entre valores reais e preditos (reportada na matriz de confusão resultante do modelo) caísse para menos de 90% em pelo menos uma das classes (“Feno”). Ao atingir esse parâmetro, as variáveis restantes foram mantidas em cada modelo, que foram então salvos como treinados e usados para avaliar os dados do arquivo contendo os 10% iniciais, separados do grupo original e impedidos de terem sido previamente expostos ao modelo.

3. RESULTADOS E DISCUSSÃO

Nesta seção, são apresentados e discutidos os resultados obtidos neste estudo.

3.1 Comparação entre os modelos de machine learning Gradient Boosting e Random Forest utilizando todas as 91 variáveis

Todas as 91 variáveis foram agrupadas com base em suas semelhanças intrínsecas, mostradas na Tabela 2.

Grupo	Variáveis
células_dimensões	Cells Nuc/Cell Area wv3; Cells Area wv3; Cells Major Axis Angle wv3; Cells Minor Axis wv3; Cells Diameter wv3
células_fl_distribuição	Cells Intensity Spreading wv3; Cells Skewness wv1; Cells High Grey Level Run Emphasis wv1; Cells Entropy wv1; Cells Kurtosis wv1; Cells Energy wv1; Cells Compactness wv3; Cells Low Grey Level Run Emphasis wv1
células_fl_intensidade	Cells Pearson correlation (Cell) wv3; Cells Pearson correlation (Cyto) wv1; Cells Intensity CV (Cell) wv3; Cells Bckg Intensity wv3; Cells Cyto/Bckg Intensity wv3; Cells Total Intensity-Bckg (Cyto) wv3; Cells Intensity-Bckg (Cell) wv3; Cells Intensity-Bckg(Cyto) wv3; Cells Nuc/Cyto Intensity wv3; Cells Pearson correlation (Cell) wv1; Cells Intensity CV (Cyto) wv3; Cells Run Length Non uniformity wv1; Cells Grey Level Non uniformity wv1; Cells Chord Ratio wv3; Cells Max Intensity wv3; Cells Total Intensity-Bckg (Cell) wv3; Cells Intensity SD (Cell) wv3; Cells Intensity SD (Cyto) wv3; Cells Pearson correlation (Cyto) wv3; Cells Light Flux wv3; Cells Global Bckg

	Intensity wv3; Cells Total Intensity (Cyto) wv3; Cells Cyto Intensity/Global Bckg wv3; Cells Intensity (Cyto) wv3; Cells Cell/Bckg Intensity wv3; Cells Total Intensity (Cell) wv3; Cells Count wv3; Cells Intensity (Cell) wv3
células_forma	Cells Form Factor wv3; Cells Elongation wv3; Cells Gyration Radius wv3; Cells Major Axis wv3
células_espacial	Cells cg Y wv3; Cells cg X wv3
núcleos_dimensões	Nuclei Area wv1; Nuclei Diameter wv1; Nuclei Perimeter wv1
núcleos_fl_distribuição	Nuclei Energy wv1; Nuclei Compactness wv1; Nuclei High Grey Level Run Emphasis wv1; Nuclei Low Grey Level Run Emphasis wv1; Nuclei Run Length Non uniformity wv1; Nuclei Light Flux wv1; Nuclei Chord Ratio wv1; Nuclei Skewness wv1; Nuclei Entropy wv1; Nuclei Grey Level Non uniformity wv1
núcleos_fl_intensidade	Nuclei Intensity-Bckg wv1; Nuclei Total Intensity-Bckg wv1; Nuclei Intensity/Global Bckg wv1; Nuclei Kurtosis wv1; Nuclei Intensity SD wv1; Nuclei Intensity wv1; Cells Perimeter wv3; Nuclei Pearson correlation wv1; Nuclei Intensity CV wv1; Nuclei Total Intensity wv1; Nuclei Global Bckg Intensity wv1
núcleos_numéricos	Nuclei Count wv1

núcleos_forma	Nuclei Gyration Radius wv1; Nuclei Major Axis Angle wv1; Nuclei Form Factor wv1; Nuclei Elongation wv1; Nuclei Major Axis wv1; Nuclei Minor Axis wv1
núcleos_espacial	Nuclei Neighbor Count (MIN) wv1; Nuclei Displacement wv1; Nuclei Neighbor Count (Lune) wv1; Nuclei Spacing (MIN) wv1; Nuclei cg X wv1; Nuclei cg Y wv1; Nuclei Spacing (Lune) wv1; Nuclei Neighbor Count (Gabriel) wv1; Nuclei Spacing (Gabriel) wv1; Nuclei Spacing (SOI) wv1; Nuclei Neighbor Count (SOI) wv1
tempo	T

Tabela 2: Todas as variáveis (91) adquiridas para treinar os modelos GB e RF, agrupadas com base em sua similaridade.

As métricas de avaliação do GB e RF para os dados são mostradas na Tabela 3.

Modelo	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
GB	0.997	0.956	0.956	0.956	0.956	0.944	0.987	0.179
RF	0.992	0.926	0.926	0.926	0.926	0.906	0.980	0.493

Tabela 3: Coeficientes de avaliação do GB e RF para os dados. AUC: Área Sob a Curva; AC: Acurácia de Classificação; F1: Pontuação F1; Prec: Precisão; Rev: Revocação; MCC: Coeficiente de Correlação de Matthews; Esp: Especificidade; LogLoss. Negrito: melhores coeficientes classificados.

O HCS produz grandes quantidades de dados multiparamétricos coletados de células individuais (KREWSKI et al., 2020; MALINOWSKA et al., 2022) ou de populações celulares em uma placa, o que pode ser desafiador para análise por métodos estatísticos tradicionais. Algoritmos de ML podem identificar padrões e relações complexas nesses dados, permitindo uma compreensão mais abrangente das respostas celulares a estímulos, como diferentes concentrações de nPS. Por exemplo, o ML pode ser utilizado para classificar células em grupos fenotípicos distintos com base em sua morfologia, perfis de expressão proteica ou outras características extraídas de imagens de HCS.

Modelos de aprendizado de máquina baseados em floresta, como RF e GB, oferecem capacidades robustas para analisar dados biológicos complexos e de alta dimensionalidade gerados por HCS, melhorando a interpretabilidade e o poder preditivo dos resultados experimentais. Esses métodos são reconhecidos por seu melhor desempenho em relação a modelos lineares e podem gerenciar melhor as complexidades dos fenótipos celulares e interações moleculares (FRÖHLICH et al., 2012).

O RF é particularmente eficiente no tratamento de relações não lineares e interações entre variáveis, pois agrega as previsões de inúmeras árvores de decisão. Já os métodos de GB constroem modelos de forma sequencial, corrigindo os erros das iterações anteriores, o que leva a previsões altamente precisas e estáveis (XIANG; ZHOU; TAN, 2023). Sua aplicação pode aprimorar a avaliação dos efeitos de compostos e concentrações, acelerando o processo de análise e melhorando o controle de qualidade na descoberta de fármacos (KESLER et al., 2017). As tecnologias modernas de triagem de alto rendimento oferecem capacidade preditiva estável mesmo com grandes volumes de dados e permitem o perfilamento de milhares de moléculas expostas a diversos alvos biológicos (WALSH et al., 2021). Além disso, esses algoritmos avançados podem integrar diferentes tipos de dados, como características genômicas e informações de sensibilidade a fármacos, possibilitando a construção de modelos preditivos abrangentes para aplicações em medicina de precisão (PRUMMER, 2012).

Além da descoberta de fármacos, a utilidade desses métodos se estende a diversas tarefas de diagnóstico biológico e médico, nas quais seus menores custos

computacionais e suas capacidades robustas de otimização se mostram altamente vantajosos (DAVID et al., 2019). Por exemplo, técnicas baseadas em GB demonstram boa adequação para tarefas como diagnóstico de doenças, previsão de progressão e avaliação de respostas a tratamentos (MATLOCK et al., 2018; YILDIZ; KALAYCI, 2024).

Por utilizar refinamento sequencial, o GB pode ser mais adequado e apresentar desempenho superior em termos de precisão, taxa de verdadeiros positivos e taxa de falsos positivos. Ao concentrar o aprendizado nas instâncias que geram classificações incorretas, cada árvore sucessiva aprimora a capacidade do modelo de identificar corretamente os casos positivos, reduzindo previsões equivocadas - característica especialmente relevante em contextos nos quais o custo de falsos positivos é elevado, como na detecção de fraudes financeiras e no diagnóstico médico (XU et al., 2020). Além disso, o GB otimiza sua função objetivo ao reduzir iterativamente os erros residuais das iterações anteriores, processo fundamentado matematicamente na técnica de descida de gradiente (HUANG et al., 2022).

Comparados a outros métodos de ML, o RF e o GB geralmente apresentam melhor desempenho e interpretabilidade para a análise de dados de HCS. Modelos lineares tradicionais tendem a apresentar limitações ao capturar relações não lineares entre características celulares e desfechos, enquanto redes neurais podem ser computacionalmente custosas e menos transparentes (BOLDINI et al., 2023; YAN; XU, 2024; OMAR et al., 2024). Já os métodos baseados em florestas oferecem um equilíbrio favorável entre precisão e interpretabilidade, permitindo avaliar a contribuição relativa de cada variável no modelo. Essa característica possibilita identificar os parâmetros celulares com maior poder preditivo, fornecendo insights importantes sobre mecanismos biológicos subjacentes. Ademais, esses métodos demonstram robustez diante de ruídos e valores discrepantes, um ponto crítico na análise de dados de HCS, que frequentemente apresentam variações experimentais.

3.2 Comparação entre os modelos de machine learning Gradient Boosting e Random Forest para nPS de 10 ou 15nm usando apenas as variáveis mais bem classificadas

Após a classificação das variáveis, os modelos GB e RF tiveram seus conjuntos reduzidos apenas às variáveis relevantes para o treinamento de cada um. Todas as variáveis seleccionadas foram plotadas conforme apresentado na Figura 1.

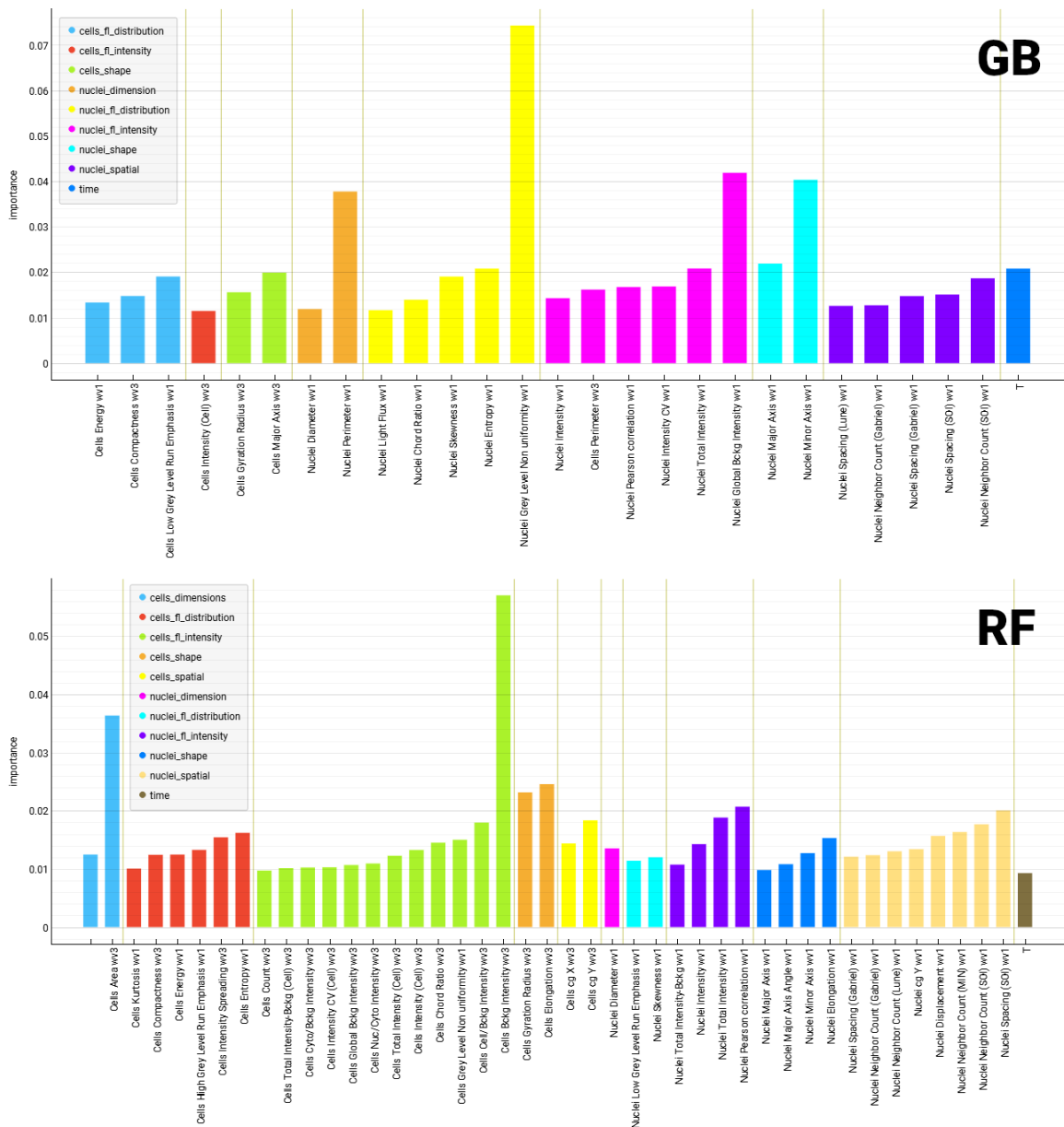


Figura 1: Variáveis classificadas para nPS de 10 ou 15nm como mais importantes para o GB e RF, agrupadas com base em similaridades.

As matrizes de confusão dos modelos contendo somente as variáveis classificadas, bem como suas respectivas métricas de avaliação, estão dispostas na Figura 2 e na Tabela 4, respectivamente. O modelo GB apresentou desempenho eficiente utilizando apenas 27 variáveis, enquanto o RF exigiu 44.

		Predicted				
		Death	PS – Death	PS – Sub Hi	PS – Sub Lo	Viable
Actual	Death	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %
	PS – Death	0.4 %	98.1 %	0.4 %	1.1 %	0.0 %
	PS – Sub Hi	0.2 %	0.4 %	98.6 %	0.8 %	0.0 %
	PS – Sub Lo	0.6 %	0.6 %	1.0 %	97.8 %	0.0 %
	Viable	0.0 %	0.0 %	0.0 %	0.0 %	100.0 %

		Predicted				
		Death	PS – Death	PS – Sub Hi	PS – Sub Lo	Viable
Actual	Death	100.0 %	0.0 %	0.0 %	0.0 %	0.0 %
	PS – Death	0.0 %	99.0 %	0.6 %	0.4 %	0.0 %
	PS – Sub Hi	0.2 %	0.6 %	98.4 %	0.8 %	0.0 %
	PS – Sub Lo	0.2 %	0.4 %	1.4 %	98.0 %	0.0 %
	Viable	0.0 %	0.0 %	0.0 %	0.2 %	99.8 %

Figura 2: Matrizes de confusão (RF e GB conforme indicado) dos modelos usando apenas variáveis classificadas.

Modelo	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
GB	0.994	0.937	0.937	0.937	0.937	0.920	0.988	0.238
RF	0.993	0.933	0.933	0.933	0.933	0.915	0.982	0.483

Tabela 4: Coeficientes de avaliação do GB e RF para os dados, usando apenas as variáveis mais bem classificadas. AUC: Área Sob a Curva; AC: Acurácia de Classificação; F1: Pontuação F1; Prec:

Precisão; Rev: Revocação; MCC: Coeficiente de Correlação de Matthews; Esp: Especificidade; LogLoss. Negrito: melhores coeficientes classificados.

Por fim, os modelos refinados carregados apenas com variáveis relevantes foram usados para prever classes e grupos do conjunto de dados prontamente separados dos dados originais, compreendendo 10% deles. As curvas ROC e matrizes de confusão obtidas após o refinamento são mostradas na Figura 3.

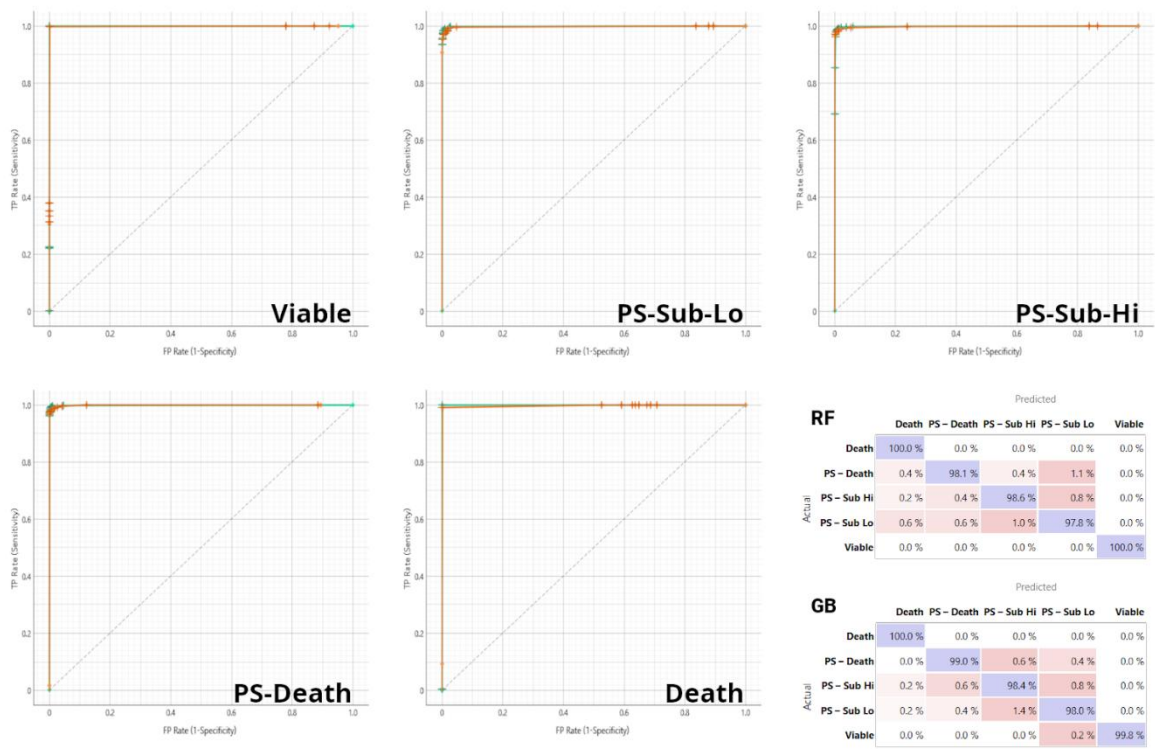


Figura 3: Curvas ROC de todos os fenótipos para GB (verde) e RF (laranja) após o refinamento do número de variáveis usadas para treinar os modelos. As probabilidades de Falso Positivo (FP) e Verdadeiro Positivo (VP) são representadas nos eixos **x** e **y**, respectivamente. Tabelas inferior direita: matrizes de confusão obtidas (RF e GB conforme indicado).

Para mostrar os verdadeiros e falsos positivos, as previsões do GB e RF foram plotadas em função dos valores transformados por RZS da contagem de células, conforme mostrado na Figura 4.

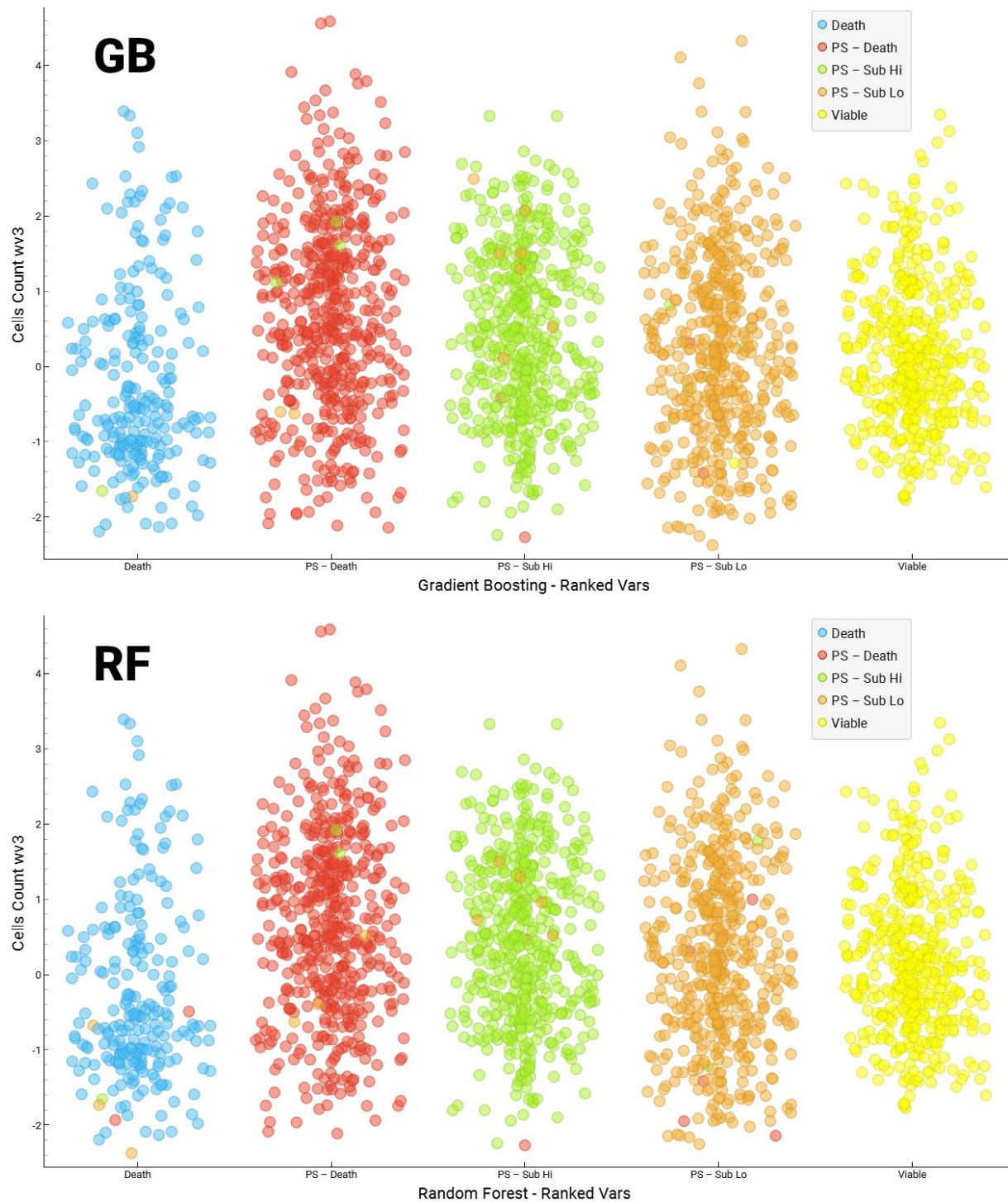


Figura 4: Previsões do GB e RF plotadas em função dos valores transformados por RZS da contagem de células, para mostrar os verdadeiros e falsos positivos.

Todos os modelos treinados puderam prever adequadamente se um grupo celular fazia parte de um grupo exato ou não. O modelo mais ajustado após o refinamento das variáveis (GB; AUC = 1,000; Prec = 0,989; Esp = 0,997 e LogLoss = 0,120) precisou de menos variáveis para ser eficientemente treinado (27 em vez de 44 do RF), aparentemente tirou a maior parte de seu poder de predição de variáveis

ligadas à intensidade de fluorescência nuclear, distribuição e organização espacial dos núcleos nos campos de visão (6, 5 e 5 variáveis selecionadas, respectivamente). O modelo RF foi ajustado com o suporte principalmente das variáveis ligadas à intensidade de fluorescência celular, à distribuição da fluorescência e à distribuição espacial nuclear (12, 6 e 8 variáveis, respectivamente).

Apesar de ser construído como ligeiramente mais preciso que o RF, o GB usou menos variáveis para atingir sua precisão. A intensidade de fluorescência e sua distribuição nas imagens nucleares foram consideradas mais importantes para o treinamento do GB do que para o treinamento do RF - 20,69 e 17,55% no GB e 9,1 e 4,55% no RF. Diferenças na sensibilidade e atribuição de relevância das variáveis podem ser distintas entre os modelos, mesmo entre aqueles muito semelhantes, como GB e RF. Por outro lado, o treinamento do RF atribuiu muito mais relevância às variáveis relacionadas à intensidade de fluorescência das células - 27,28%, opondo-se a apenas 3,45% no GB.

Ambos os modelos possuem abordagens distintas de construção que influenciam significativamente a forma como cada variável contribui para o desempenho final, especialmente quando se considera a distribuição dos dados e outras particularidades do conjunto analisado.

O RF opera gerando múltiplas árvores de decisão de maneira independente, cada uma treinada em subconjuntos aleatórios de dados e de características. Embora essa estratégia reduza de forma eficaz a variância do modelo, as medidas de importância das variáveis podem apresentar viés em direção a preditores correlacionados - um efeito decorrente tanto da tendência do algoritmo em selecionar características correlacionadas durante o processo de divisão quanto do cálculo de importância baseado em permutação incondicional (BIAU; CADRE, 2017). Assim, em conjuntos de dados com forte correlação entre variáveis, o RF pode deixar de destacar os preditores mais informativos de maneira isolada, o que pode mascarar sua relevância individual.

Por outro lado, o GB constrói suas árvores de maneira sequencial, na qual cada nova árvore é explicitamente projetada para aprender e corrigir os erros cometidos pelas árvores precedentes. Esse mecanismo iterativo de correção de erros permite

que o modelo se concentre progressivamente em instâncias que foram previamente mal classificadas ou difíceis de prever, tornando as variáveis cruciais para resolver essas classificações incorretas específicas significativamente mais relevantes do que para o RF (WANG et al., 2025; STROBL et al., 2008). Além disso, as iterações tendem a identificar pontos fracos no modelo e aproveitar dados adicionais para superá-los incrementalmente, levando a enfatizar características relevantes para reduzir erros residuais. Esse refinamento sequencial e capacidade de correção de erros também torna o GB particularmente responsivo a algumas particularidades dos dados, como em cenários envolvendo conjuntos de dados desbalanceados, onde métodos convencionais podem ter dificuldades devido a distribuições de dados distorcidas (FERSINI; MESSINA; POZZI, 2014; HUANG et al., 2021). Os modelos GB podem ser adaptados usando funções de perda com balanceamento de classe para melhorar o desempenho (LYASHEVSKA et al., 2021)), o que pode significar que variáveis que ajudam na classificação precisa de classes minoritárias ou no tratamento de dados ruidosos tornam-se altamente relevantes à medida que o modelo refina iterativamente suas previsões. Além disso, os modelos GB são capazes de aprender relações não lineares e efeitos interativos profundos ao mudar continuamente o foco para observações problemáticas (FERSINI; MESSINA; POZZI, 2014). Essa capacidade de identificar e aproveitar interações complexas implica que características que participam desses relacionamentos intrincados ganham relevância aumentada dentro da estrutura do modelo, permitindo que o GB extraia poder preditivo mais nuances dos dados.

4. CONCLUSÕES

Os experimentos apresentados foram capazes de encontrar, entre dezenas de variáveis, as mais relevantes para treinar modelos de GB e RF para analisar células expostas a nPS com base em diferenças de intensidade e distribuição de fluorescência em núcleos e células, dimensões, descritores de forma e distribuição nos campos de visão adquiridos. Os modelos GB e RF treinados puderam diferenciar a morte celular causada por DMSO da causada por nPS. Os modelos também mostraram a capacidade de distinguir células viáveis de células expostas a nPS, abrindo espaço para um futuro sistema de detecção projetado para encontrar

nanoplásticos mesmo em concentrações não tóxicas, tornando mais preciso o monitoramento industrial e ambiental dos efeitos dos nanoplásticos.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] BANERJEE, A.; BILLEY, L. O.; SHELVER, W. L. Uptake and toxicity of polystyrene micro/nanoplastics in gastric cells: effects of particle size and surface functionalization. *PLoS One*, v. 16, p. 1–25, 2021. <https://doi.org/10.1371/journal.pone.0260803>.
- [2] BIAU, G.; CADRE, B. Optimization by gradient boosting. In: BONNEFOY, J.; FLEURET, F. (ed.). *Advances in contemporary statistics and econometrics*. Cham: Springer, 2017. p. 23–44. https://doi.org/10.1007/978-3-030-73249-3_2.
- [3] BLOCKEEL, H.; DEVOS, L.; FRÉNEY, B.; NANFACK, G.; NIJSSEN, S. Decision trees: from efficient prediction to responsible AI. *Frontiers in Artificial Intelligence*, v. 6, p. 1124553, 2023. <https://doi.org/10.3389/frai.2023.1124553>.
- [4] BOLDINI, D. et al. Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics*, v. 15, p. 1–13, 2023. <https://doi.org/10.1186/s13321-023-00743-7>.
- [5] BORRELLE, S. B. et al. Predicted growth in plastic waste exceeds efforts to mitigate plastic pollution. *Science*, v. 369, n. 6510, p. 1515-1518, 2020. <https://doi.org/10.1126/science.aba3656>.
- [6] DAVID, L. et al. Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research. *Frontiers in Pharmacology*, v. 10, p. 490337, 2019. <https://doi.org/10.3389/fphar.2019.01303>.
- [7] FERSINI, E.; MESSINA, E.; POZZI, F. A. Sentiment analysis: Bayesian ensemble learning. *Decision Support Systems*, v. 68, p. 26–38, 2014. <https://doi.org/10.1016/j.dss.2014.10.004>.
- [8] FRÖHLICH, E. et al. Action of polystyrene nanoparticles of different sizes on lysosomal function and integrity. *Particle and Fibre Toxicology*, v. 9, p. 1–13, 2012. <https://doi.org/10.1186/1743-8977-9-26>.

- [9] GEYER, R.; JAMBECK, J. R.; LAW, K. L. Production, use, and fate of all plastics ever made. *Science Advances*, v. 3, n. 7, 2017. <https://doi.org/10.1126/sciadv.1700782>.
- [10] HUANG, T. et al. Extending models via gradient boosting: an application to Mendelian models. *Annals of Applied Statistics*, v. 15, p. 1126–1146, 2021. <https://doi.org/10.1214/21-AOAS1482>.
- [11] HUANG, W. et al. Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction. *Scientific Reports*, v. 12, p. 1–12, 2022. <https://doi.org/10.1038/s41598-021-04649-y>.
- [12] KELPSIENE, E.; CEDERVALL, T.; MALMENDAL, A. Metabolomics-based analysis in *Daphnia magna* after exposure to low environmental concentrations of polystyrene nanoparticles. *Environmental Science: Nano*, v. 10, p. 1858–1866, 2023. <https://doi.org/10.1039/D3EN00142C>.
- [13] KESLER, S. R. et al. Predicting long-term cognitive outcome following breast cancer with pre-treatment resting state fMRI and random forest machine learning. *Frontiers in Human Neuroscience*, v. 11, 2017. <https://doi.org/10.3389/fnhum.2017.00555>.
- [14] KIK, K.; BUKOWSKA, B.; SICIŃSKA, P. Polystyrene nanoparticles: sources, occurrence in the environment, distribution in tissues, accumulation and toxicity to various organisms. *Environmental Pollution*, v. 262, 2020. <https://doi.org/10.1016/j.envpol.2020.114297>.
- [15] KONG, B.; SEOG, J. H.; GRAHAM, L. M.; LEE, S. B. Experimental considerations on the cytotoxicity of nanoparticles. *Nanomedicine (London)*, v. 6, p. 929, 2011. <https://doi.org/10.2217/nnm.11.77>.
- [16] KREWSKI, D. et al. Toxicity testing in the 21st century: progress in the past decade and future perspectives. *Archives of Toxicology*, v. 94, p. 1–58, 2020. <https://doi.org/10.1007/s00204-019-02613-4>.

- [17] LYASHEVSKA, O. et al. Class imbalance in gradient boosting classification algorithms: application to experimental stroke data. *Statistical Methods in Medical Research*, v. 30, p. 916–925, 2021. <https://doi.org/10.1177/0962280220980484>.
- [18] MAHMUD, F. et al. Molecular and cellular effects of microplastics and nanoplastics: focus on inflammation and senescence. *Cells*, v. 13, p. 1788, 2024. <https://doi.org/10.3390/cells13211788>.
- [19] MALINOWSKA, J. M. et al. Integrating in vitro metabolomics with a 96-well high-throughput screening platform. *Metabolomics*, v. 18, p. 1–11, 2022. <https://doi.org/10.1007/s11306-021-01867-3>.
- [20] MATLOCK, K. et al. Investigation of model stacking for drug sensitivity prediction. *BMC Bioinformatics*, v. 19, p. 21–33, 2018. <https://doi.org/10.1186/s12859-018-2060-2>.
- [21] MITRANO, D. M.; WICK, P.; NOWACK, B. Placing nanoplastics in the context of global plastic pollution. *Nature Nanotechnology*, v. 16, p. 491–500, 2021. <https://doi.org/10.1038/s41565-021-00888-2>.
- [22] OMAR, E. D. et al. Comparative analysis of logistic regression, gradient boosted trees, SVM, and random forest algorithms for prediction of acute kidney injury requiring dialysis after cardiac surgery. *International Journal of Nephrology and Renovascular Disease*, v. 17, p. 197–204, 2024. <https://doi.org/10.2147/IJNRD.S461028>.
- [23] PLASTICS EUROPE. *Plastics – The Facts 2025*. Brussels: Plastics Europe, 2025.
- [24] POOJA, N. et al. An insight on sources and biodegradation of bioplastics: a review. *3 Biotech*, v. 13, p. 1–22, 2023. <https://doi.org/10.1007/s13205-023-03638-4>.
- [25] PRUMMER, M. Hypothesis testing in high-throughput screening for drug discovery. *Journal of Biomolecular Screening*, v. 17, p. 519–529, 2012. <https://doi.org/10.1177/1087057111431278>.

- [26] SKABA, D. et al. Nanoplastics and immune disruption: a systematic review of exposure routes, mechanisms, and health implications. *International Journal of Molecular Sciences*, v. 26, p. 5228, 2025. <https://doi.org/10.3390/ijms26115228>.
- [27] STROBL, C. et al. Conditional variable importance for random forests. *BMC Bioinformatics*, v. 9, p. 1–11, 2008. <https://doi.org/10.1186/1471-2105-9-307>.
- [28] WALKER, T. R.; FEQUET, L. Current trends of unsustainable plastic and micro(nano)plastic pollution. *TrAC – Trends in Analytical Chemistry*, v. 160, p. 116984, 2023. <https://doi.org/10.1016/j.trac.2023.116984>.
- [29] WALSH, I. et al. Recommendations for machine learning validation in biology. *Nature Methods*, v. 18, p. 1122-1127, 2020. <https://doi.org/10.1038/s41592-021-01205-4>.
- [30] WANG, Z. et al. Evaluating the strength properties of high-performance concrete in the form of ensemble and hybrid models using deep learning techniques. *Scientific Reports*, v. 15, p. 1-31, 2025. <https://doi.org/10.1038/s41598-025-10860-y>.
- [31] WILLIAMS, A. T.; RANGEL-BUITRAGO, N. The past, present, and future of plastic pollution. *Marine Pollution Bulletin*, v. 176, p. 113429, 2022. DOI: 10.1016/j.marpolbul.2022.113429.
- [32] XIANG, Q. et al. Toxicity effects of polystyrene nanoplastics with different sizes on freshwater microalgae *Chlorella vulgaris*. *Molecules*, v. 28, p. 3958, 2023. <https://doi.org/10.3390/molecules28093958>.
- [33] XU, Y. et al. Machine learning algorithms for predicting the recurrence of stage IV colorectal cancer after tumor resection. *Scientific Reports*, v. 10, p. 1-9, 2020. <https://doi.org/10.1038/s41598-020-59115-y>.
- [34] YAN, L. et al. XGBoost-enhanced graph neural networks: a new architecture for heterogeneous tabular data. *Applied Sciences*, v. 14, n. 13, p. 5826, 2024. <https://doi.org/10.3390/app14135826>.
- [35] YILDIZ, A. Y.; KALAYCI, A. Gradient boosting decision trees on medical diagnosis over tabular data. *arXiv*, 2024. <https://arxiv.org/pdf/2410.03705>.

- [36] ZANGMEISTER, C. D. et al. Common single-use consumer plastic products release trillions of sub-100 nm nanoparticles per liter into water during normal use. *Environmental Science & Technology*, v. 56, p. 5448-5455, 2022. <https://doi.org/10.1021/acs.est.1c06768>.
- [37] ZHANG, J. et al. Combining transcriptomic and metabolomic analyses to investigate the acute effects of microcystin-LR and nanoplastics of asian clams. *Water*, v. 15, n. 19, p. 3519, 2023. <https://doi.org/10.3390/w15193519>.
- [38] ZHANG, Z. et al. Predicting cell properties with AI from 3D imaging flow cytometer data. *Scientific Reports*, v. 15, p. 1-8, 2025. <https://doi.org/10.1038/s41598-024-80722-6>.