

## Caracterização e Classificação de Conjuntos Desbalanceados de Dermoscopias

### Characterization and Classification of Imbalanced Dermoscopic Datasets

### Caracterización y Clasificación de Conjuntos Dermatoscópicos Desequilibrados

Newton Spolaôr<sup>1</sup>, Huei Diana Lee<sup>2</sup>, Weber Shoity Resende Takaki<sup>1</sup>, Leandro Augusto Ensina<sup>3</sup>, Antonio Rafael Sabino Parmezan<sup>1</sup>, Matheus Maciel<sup>4</sup>, Claudio Saddy Rodrigues Coy<sup>5</sup>, Feng Chung Wu<sup>6,7</sup>

1 Doutor, Laboratório de Bioinformática, Universidade Estadual do Oeste do Paraná – LABI/UNIOESTE, Foz do Iguaçu (PR), Brasil.

2 Professora Associada-III Doutora do Centro de Engenharias e Ciências Exatas, Laboratório de Bioinformática, Universidade Estadual do Oeste do Paraná – LABI/UNIOESTE, Foz do Iguaçu (PR), Brasil.

3 Mestre, Laboratório de Bioinformática, Universidade Estadual do Oeste do Paraná – LABI/UNIOESTE, Foz do Iguaçu (PR), Brasil.

4 Bacharel, Laboratório de Bioinformática, Universidade Estadual do Oeste do Paraná – LABI/UNIOESTE, Foz do Iguaçu (PR), Brasil.

5 Professor Titular Doutor da Faculdade de Ciências Médicas, Universidade Estadual de Campinas – FCM/UNICAMP, Campinas (SP), Brasil.

6 Professor Associado-III Doutor do Centro de Educação, Letras e Saúde, Laboratório de Bioinformática, Universidade Estadual do Oeste do Paraná – LABI/UNIOESTE, Foz do Iguaçu (PR), Brasil.

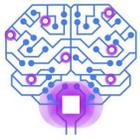
7 Professor Doutor da Faculdade de Ciências Médicas, Universidade Estadual de Campinas – FCM/UNICAMP, Campinas (SP), Brasil.

Autor correspondente: Huei Diana Lee

E-mail: huei.lee@unioeste.br

## Resumo

**Objetivo:** Investigar técnicas de inteligência computacional para caracterizar e classificar conjuntos desbalanceados de lesões dermatoscópicas. **Métodos:** O método desenvolvido contempla técnicas para pré-processamento de imagens, extração de atributos (características), sobreamostragem, seleção de atributos, e construção e avaliação de classificadores. Vinte configurações do método foram avaliadas em 274 dermoscopias públicas com 48 melanomas e 226 nevos. **Resultados:** Atingiu-se a maior acurácia



média, 83,57%, após reduzir o número de características em pelo menos 48,86%. Em geral, a sobreamostragem melhorou a sensibilidade média. **Conclusão:** Os melhores resultados do método na caracterização e classificação de um conjunto desbalanceado de dermatoscopias foram promissores e competitivos com algumas referências recentes.

**Descritores:** Neoplasias Cutâneas; Informática Médica; Inteligência Artificial

## Abstract

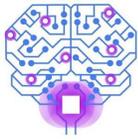
**Objective:** To investigate computational intelligence techniques to characterize and classify imbalanced datasets of dermoscopic lesions. **Methods:** The developed method includes techniques for image pre-processing, feature extraction, oversampling, feature selection, and classifier building and evaluation. We assessed 20 method configurations in 274 public dermoscopies with 48 melanomas and 226 nevi. **Results:** We reached the highest average accuracy, 83.57%, after reducing the feature number by at least 48.86%. In general, the oversampling technique improved the average sensitivity. **Conclusion:** The best method results in the characterization and classification of an imbalanced dermoscopic dataset were promising and competitive with some recent references.

**Keywords:** Skin Neoplasms; Medical Informatics; Artificial Intelligence

## Resumen

**Objetivo:** investigar técnicas de inteligencia computacional para caracterizar y clasificar conjuntos desequilibrados de lesiones dermatoscópicas. **Métodos:** El método desarrollado incluye técnicas de preprocesamiento de imágenes, extracción de atributos, sobremuestreo, selección de atributos y construcción y evaluación de clasificadores. Se evaluaron 20 configuraciones del método en 274 dermatoscopias públicas con 48 melanomas y 226 nevos. **Resultados:** La mayor precisión promedio, 83,57%, se alcanzó después de reducir el número de características en al menos un 48,86%. En general, el sobremuestreo mejoró la sensibilidad promedio. **Conclusión:** Los mejores resultados del método en la caracterización y clasificación de un conjunto dermatoscópico desequilibrado fueron prometedores y competitivos con algunas referencias recientes.

**Descriptorios:** Neoplasias Cutâneas; Informática Médica; Inteligencia Artificial



## Introdução

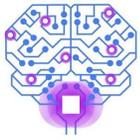
A visão computacional, por meio de tópicos como análise e processamento de imagens digitais, tem sido fundamental em soluções para o apoio ao diagnóstico precoce de doenças como o câncer de pele <sup>(1-8)</sup>. A exemplo de outros países, no Brasil estima-se que o câncer de pele seja o mais incidente no triênio 2020-2022 <sup>(9)</sup>. Parte desses casos são melanoma, um perigoso tumor que surge a partir de células produtoras de melanina.

A análise e processamento de imagens pode apoiar especialistas da saúde, por exemplo, na tomada de decisão inerente à descrição e à rotulação de lesões malignas e benignas em exames dermoscópicas <sup>(7,8)</sup>. Para tanto, tarefas de Aprendizado de Máquina (AM) como a extração de atributos (características) de dermoscopias e a classificação delas têm sido adotadas <sup>(10,11)</sup>. A Seleção de Atributos (SA) importantes, dentre os que foram previamente extraídos, também tem se destacado pelo potencial de combater efeitos da “maldição da dimensionalidade” que afetam a descoberta de conhecimento <sup>(12)</sup>.

Muitas vezes, o aprendizado ocorre sobre conjuntos de exemplos (imagens médicas ou instâncias) com desbalanceamento de classes contendo, por exemplo, um número de lesões malignas (rótulo ou classe positiva) muito menor que o de benignas (negativa). Nesse contexto, algoritmos de AM podem ter dificuldade em (1) categorizar novas imagens da classe menos comum, em geral de alta relevância clínica, e em (2) atingir sensibilidade satisfatória no apoio a Dermatologia e outros domínios que lidam com doenças preocupantes, como o câncer. A sobreamostragem (*oversampling*) <sup>(13)</sup> é uma alternativa notória de combate ao desbalanceamento que adiciona exemplos do rótulo minoritário ao conjunto, com apoio de instâncias já existentes.

O uso conjunto de *oversampling* e SA é incipiente na análise de dermoscopias <sup>(5,6)</sup>. Em vários processos com essa combinação, estima-se a importância das características escolhidas com apoio de um classificador. Uma alternativa nesse cenário é a abordagem filtro para SA, independente de classificação, que pode identificar atributos apropriados para posterior aprendizado com menor custo computacional por distintos algoritmos <sup>(11,12)</sup>.

O objetivo deste trabalho consiste em investigar técnicas de inteligência computacional para caracterizar e classificar conjuntos desbalanceados de lesões dermoscópicas. Um método baseado em sobreamostragem e SA filtro foi implementado e aplicado em 274 imagens de pele, das quais somente 48 pertencem à classe minoritária



(melanoma). Os melhores resultados na diferenciação entre imagens malignas e benignas foram promissores e competitivos perante algumas referências recentes.

## Métodos

Nesta seção estão apresentados os materiais usados pelo método desenvolvido, além das configurações e dos procedimentos abordados em sua avaliação experimental.

## Material

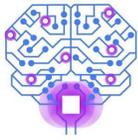
O conjunto de imagens empregado consiste em 274 dermoscopias adquiridas a partir do repositório *International Skin Imaging Collaboration* (ISIC) (<https://www.isic-archive.com>). Dentre elas, 48 contêm melanoma e 226 nevos <sup>(14,15)</sup>. Ressalta-se que a existência de uma quantidade maior de nevos neste conjunto (1) aproxima a proporção observada no ISIC e (2) considera que a maioria das pessoas no mundo real tem nevos <sup>(16)</sup>, em frequência maior que a incidência de melanoma. Cada dermoscopia está ainda associada a uma máscara que delimita a lesão correspondente. Não há necessidade de aprovação de comitê de ética neste trabalho porque as imagens citadas são públicas.

As ferramentas computacionais Weka 3.8.5 <sup>(11)</sup> e os ambientes de programação R 4.1.1 (<https://cran.r-project.org/bin/windows/base>) – com apoio de bibliotecas como Caret <sup>(17)</sup> – e MathWorks MATLAB 2011 (<https://www.mathworks.com/products/matlab.html>) apoiaram a implementação das técnicas usadas no método descrito a seguir.

## Método Desenvolvido

O método deste trabalho contém cinco fases: (1) pré-processamento de imagens; (2) extração de atributos (características); (3) sobreamostragem; (4) SA; (5) construção e avaliação de classificadores. Algumas variações do método descritas na próxima seção são referências experimentais que dispensam as técnicas das Fases (3) e ou (4).

Na Fase (1), as 274 dermoscopias adquiridas foram submetidas a uma segmentação simples em que se recorta cada imagem no retângulo delimitado pelas coordenadas mínimas e máximas da máscara correspondente <sup>(14,15)</sup>. Assim, as lesões completas e alguns trechos de pele na vizinhança foram mantidos para análise posterior.



Na Fase (2), 88 atributos de textura foram extraídos de cada imagem segmentada. Ao todo, são 13 descritores de Haralick <sup>(18)</sup> e 75 de Laws <sup>(19)</sup>, implementados em <sup>(8)</sup>.

A sobreamostragem inerente à Fase (3) foi realizada por meio do algoritmo *Synthetic Minority Over-sampling Technique* (SMOTE) <sup>(13)</sup>. Sua ideia, ilustrada em <sup>(5,6)</sup>, é criar exemplos artificiais da classe menos frequente  $C$  levando em conta o espaço de atributos, diferentemente de abordagens de *data augmentation* que atuam diretamente em imagens de todos os rótulos <sup>(10)</sup>. Para tanto, no SMOTE se identifica, para cada exemplo  $S$  da classe  $C$ , os  $K$  vizinhos mais próximos de mesmo rótulo. Após,  $E$  exemplos são randomicamente selecionados dentre os  $K$  disponíveis para gerar  $E$  novas instâncias. Cada uma delas é definida pela interpolação entre  $S$  e um dos seus  $E$  vizinhos. Ao final, o número de instâncias de  $C$  aumenta em  $E \times 100\%$ .

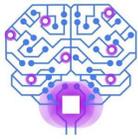
Na Fase (4), como realizado em <sup>(14,15)</sup>, aplicou-se o algoritmo filtro *Correlation-based Feature Selection* (CFS) <sup>(11)</sup>. O intuito do CFS é encontrar um subconjunto de características altamente correlacionadas com a classe e pouco redundantes entre si.

Na Fase (5) foram usados os algoritmos de classificação *Random Forest* (RF), *Support Vector Machines* (SVM) e *eXtreme Gradient Boosting* (XGB) <sup>(11,17)</sup>. Os modelos foram construídos e avaliados conforme detalhado na próxima seção.

## Avaliação Experimental

Foram implementadas duas abordagens de Validação Cruzada (VC) estratificada no conjunto de 274 imagens representado por 88 características. Na VC tradicional esse conjunto foi dividido em dez pares de *folds* (partições) de treinamento e teste <sup>(11)</sup>. Por sua vez, na VC aninhada (*nested*) adicionou-se um segundo nível de validação para cada *fold* de treinamento  $F$  do primeiro nível <sup>(20)</sup>. Para tanto,  $F$  foi subdividido em dez pares de *folds* de treinamento e validação. As partições de treinamento do segundo nível serviram para estimar em  $F$  um metaparâmetro: grupo de atributos selecionados. Para determiná-lo, realizou-se SA em cada partição de treinamento do segundo nível, levando a dez subconjuntos de descritores. A frequência de cada item na união desses subconjuntos permitiu definir o grupo com os  $P\%$  descritores mais frequentes.

SMOTE, CFS, RF e SVM foram usados com os valores de parâmetro *default* do Weka <sup>(11)</sup>, como  $E=1$  e  $K=5$  para SMOTE. Este algoritmo foi ainda avaliado com  $K=1$  e 3,



dados que se identificou em estudos relacionados que quantidades reduzidas de vizinhos poderiam contribuir com o desempenho de classificação <sup>(5)</sup>. A implementação XGBTree do algoritmo XGB na biblioteca Caret foi também considerada com valores *default* <sup>(17)</sup>.

Neste trabalho foram investigadas 20 configurações do método desenvolvido que combinam os seguintes cenários com e sem sobreamostragem e SA:

- Caso sem SMOTE (base) e com SMOTE para  $K=1, 3$  e  $5$  – quatro variações;
- Caso sem SA na VC tradicional, com SA na mesma abordagem (CFS tradicional) e na aninhada (CFS aninhado) para  $P=100\%$ ,  $75\%$  e  $50\%$  – cinco variações.

Nas configurações do tipo CFS tradicional foi aplicado o subconjunto de características escolhido no *iésimo fold* de treinamento  $F$  como entrada para construir modelos RF, SVM e XGB na mesma partição. Por sua vez, nas do tipo CFS aninhado foi empregado o grupo de atributos estimado em  $F$ , localizado no primeiro nível, para gerar classificadores com os três algoritmos de aprendizado no mesmo *fold*.

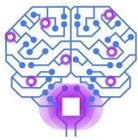
Nos cenários com sobreamostragem e SA para ambos os esquemas de validação, *oversampling* foi aplicado antes do CFS em *folds* de treinamento. Especificamente na modalidade aninhada, a criação de exemplos sintéticos ocorreu em cada um dos níveis.

O complemento do Erro da Classe Majoritária (ECM),  $(1 - 48/274) \times 100 = 82,48\%$ , foi uma referência experimental. Além da acurácia porcentual, as medidas Sensitividade e Especificidade obtidas nas dez partições de teste foram registradas.

As acurácias dos  $3 \times 20 = 60$  preditores inferidos com o uso de RF, SVM e XGB em 20 variações foram comparadas estatisticamente com a hipótese nula de que o desempenho desses modelos é equivalente. O teste de normalidade D'Agostino e Pearson, recomendado no GraphPad Prism (<https://www.graphpad.com/scientific-software/prism>), foi adotado para auxiliar na escolha do teste estatístico mais adequado.

## Resultados e Discussão

Na Tabela 1 são apresentadas as acurácias médias, dentre dez *folds* de teste, dos modelos construídos pelos três Algoritmos (A) de aprendizado adotados neste trabalho. O desvio padrão correspondente é exibido entre parênteses. Cada caso com acurácia melhor que o complemento do ECM (82,48%) é destacado em cinza.



**Tabela 1** – Acurácia porcentual obtido em cada configuração por três algoritmos

A	Sem SA tradicional				CFS tradicional			
	Base	K=1	K=3	K=5	Base	K=1	K=3	K=5
<b>RF</b>	83,54 (4,40)	83,19 (4,71)	83,54 (4,76)	82,08 (4,54)	83,57 (6,29)	80,25 (8,22)	79,95 (4,51)	78,45 (3,28)
<b>SVM</b>	82,12 (2,08)	81,04 (7,43)	81,76 (6,43)	81,42 (8,51)	82,49 (1,46)	81,01 (4,27)	81,03 (3,71)	81,38 (5,04)
<b>XGB</b>	83,21 (4,91)	81,76 (4,47)	80,28 (4,32)	79,55 (4,60)	81,72 (4,00)	82,09 (5,07)	80,99 (3,51)	79,89 (5,36)
A	CFS aninhado (P=100%)				CFS aninhado (P=75%)			
	Base	K=1	K=3	K=5	Base	K=1	K=3	K=5
<b>RF</b>	83,19 (5,00)	83,57 (5,01)	81,72 (4,68)	81,35 (6,17)	81,73 (6,25)	82,43 (4,96)	80,99 (4,61)	82,47 (4,82)
<b>SVM</b>	82,49 (1,46)	82,84 (6,93)	82,49 (4,78)	80,66 (5,13)	82,49 (1,46)	82,12 (5,58)	80,29 (4,62)	81,76 (6,90)
<b>XGB</b>	81,02 (5,36)	82,16 (5,88)	81,71 (4,41)	82,45 (5,24)	79,55 (5,76)	82,09 (3,36)	82,08 (5,68)	79,19 (2,49)
A	CFS aninhado (P=50%)							
	Base	K=1	K=3	K=5				
<b>RF</b>	81,73 (6,32)	79,89 (5,62)	79,18 (5,48)	79,52 (4,48)				
<b>SVM</b>	82,49 (1,46)	80,29 (4,96)	80,62 (4,43)	81,02 (4,15)				
<b>XGB</b>	80,99 (4,89)	77,34 (3,52)	79,89 (5,83)	80,30 (6,39)				

Em 13 classificadores, baseados em nove configurações distintas, se atingiu desempenho melhor que o complemento do ECM. Quase a metade deles (seis) foi gerada por meio de RF, incluindo os dois de maior acurácia média, 83,57%. Esses dois casos, doravante denominados M1 e M2, são baseados respectivamente nas variações (1) CFS tradicional sem SMOTE e (2) CFS aninhado com  $P=100\%$  e SMOTE ( $K=1$ ). Com SVM foram gerados seis modelos que sobressaíram na Tabela 1, alguns dos quais são relacionados aos menores desvios padrão. Um caso destacado foi obtido com XGB sem SA nem sobreamostragem.

Para complementar a análise, aplicou-se o teste não paramétrico de Friedman (nível de significância  $\alpha=0,05$ ), dado que os valores de acurácia não passaram pelo teste de normalidade. Porém, não foi possível rejeitar a hipótese nula de que o desempenho dos preditores é equivalente. Um aspecto positivo é que cenários baseados em SA foram considerados promissores, pois registraram desempenho similar às alternativas que



usam mais descritores. Com M1, por exemplo, se empregou de seis a dez atributos por *fold* e se reduziu em pelo menos 88,64% o conjunto de dados original (88 características). Com M2 se diminuiu a dimensionalidade em pelo menos 48,86%. Os atributos escolhidos por M1 e M2 foram publicados (<https://dx.doi.org/10.5281/zenodo.6985365>). Uma quantidade menor de características é preferível no domínio da aplicação para simplificar a descrição das imagens a serem rotuladas pelo método deste trabalho.

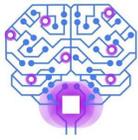
A sensibilidade média e o desvio padrão correspondente para as mesmas configurações da Tabela 1 são relatados na Tabela 2. Por sua vez, a especificidade média e o desvio padrão correspondente para as 20 variações do método são exibidos na Tabela 3. Os casos destacados na primeira tabela seguem realçados em cinza.

**Tabela 2** – Sensitividade obtida em cada configuração por três algoritmos

A	Sem SA tradicional				CFS tradicional			
	Base	K=1	K=3	K=5	Base	K=1	K=3	K=5
RF	0,21 (0,23)	0,31 (0,20)	0,36 (0,17)	0,31 (0,20)	0,25 (0,21)	0,39 (0,24)	0,37 (0,25)	0,35 (0,16)
SVM	0,00 (0,00)	0,38 (0,21)	0,34 (0,17)	0,36 (0,23)	0,00 (0,00)	0,02 (0,06)	0,00 (0,00)	0,06 (0,13)
XGB	0,33 (0,22)	0,33 (0,20)	0,33 (0,19)	0,31 (0,17)	0,21 (0,16)	0,35 (0,19)	0,33 (0,23)	0,35 (0,21)
A	CFS aninhado (P=100%)				CFS aninhado (P=75%)			
	Base	K=1	K=3	K=5	Base	K=1	K=3	K=5
RF	0,22 (0,26)	0,38 (0,18)	0,38 (0,18)	0,35 (0,21)	0,21 (0,21)	0,35 (0,23)	0,28 (0,23)	0,42 (0,16)
SVM	0,00 (0,00)	0,23 (0,25)	0,21 (0,16)	0,19 (0,23)	0,00 (0,00)	0,18 (0,24)	0,10 (0,17)	0,25 (0,24)
XGB	0,22 (0,26)	0,46 (0,20)	0,37 (0,25)	0,39 (0,20)	0,23 (0,20)	0,39 (0,16)	0,37 (0,22)	0,30 (0,22)
A	CFS aninhado (P=50%)							
	Base	K=1	K=3	K=5				
RF	0,29 (0,27)	0,35 (0,19)	0,37 (0,18)	0,35 (0,21)				
SVM	0,00 (0,00)	0,08 (0,14)	0,04 (0,08)	0,14 (0,16)				
XGB	0,26 (0,25)	0,28 (0,17)	0,39 (0,21)	0,41 (0,24)				

**Tabela 3** – Especificidade obtida em cada configuração por três algoritmos

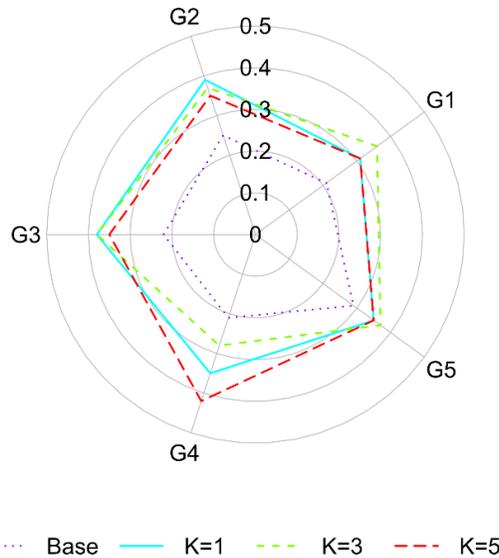
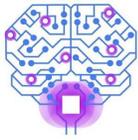
A	Sem SA tradicional				CFS tradicional			
	Base	K=1	K=3	K=5	Base	K=1	K=3	K=5



<b>RF</b>	0,97 (0,03)	0,94 (0,06)	0,94 (0,05)	0,93 (0,06)	0,96 (0,05)	0,89 (0,08)	0,89 (0,05)	0,88 (0,05)
<b>SVM</b>	1,00 (0,01)	0,90 (0,07)	0,92 (0,07)	0,91 (0,08)	1,00 (0,00)	0,98 (0,04)	0,98 (0,04)	0,97 (0,05)
<b>XGB</b>	0,94 (0,06)	0,92 (0,06)	0,90 (0,06)	0,90 (0,07)	0,95 (0,04)	0,92 (0,05)	0,91 (0,05)	0,89 (0,06)
<b>A</b>	<b>CFS aninhado (P=100%)</b>				<b>CFS aninhado (P=75%)</b>			
	<b>Base</b>	<b>K=1</b>	<b>K=3</b>	<b>K=5</b>	<b>Base</b>	<b>K=1</b>	<b>K=3</b>	<b>K=5</b>
<b>RF</b>	0,96 (0,04)	0,93 (0,05)	0,91 (0,06)	0,91 (0,06)	0,95 (0,04)	0,92 (0,04)	0,92 (0,05)	0,91 (0,04)
<b>SVM</b>	1,00 (0,00)	0,96 (0,05)	0,96 (0,04)	0,94 (0,04)	1,00 (0,00)	0,96 (0,05)	0,95 (0,05)	0,94 (0,06)
<b>XGB</b>	0,93 (0,06)	0,90 (0,09)	0,91 (0,06)	0,91 (0,07)	0,92 (0,05)	0,91 (0,06)	0,92 (0,05)	0,89 (0,04)
<b>A</b>	<b>CFS aninhado (P=50%)</b>							
	<b>Base</b>	<b>K=1</b>	<b>K=3</b>	<b>K=5</b>				
<b>RF</b>	0,93 (0,05)	0,89 (0,06)	0,88 (0,07)	0,89 (0,04)				
<b>SVM</b>	1,00 (0,00)	0,96 (0,06)	0,97 (0,05)	0,95 (0,04)				
<b>XGB</b>	0,92 (0,04)	0,88 (0,04)	0,88 (0,06)	0,89 (0,07)				

Com apoio da Tabela 2, observa-se que RF superou SVM ao atingir sensibilidade média mínima de 0,21. De fato, em quatro dos seis preditores SVM destacados foram obtidos sensibilidade média e desvio padrão zero, o que implica que neles todas as imagens com melanoma (classe positiva) foram categorizadas como nevo (negativa). Esse cenário não é apropriado no domínio de aplicação deste trabalho. Ao relacionar esse achado com os resultados da Tabela 1 em termos de acurácia média, identifica-se que RF foi o melhor algoritmo de aprendizado nos experimentos realizados.

Com relação aos dois modelos RF de melhor acurácia média – Tabela 1 –, M2, baseado em CFS aninhado com  $P=100\%$  e SMOTE ( $K=1$ ), resultou em uma sensibilidade média maior que M1 – CFS tradicional sem sobreamostragem. Realmente, a sensibilidade média melhorou com *oversampling* na maioria dos casos avaliados. Na Figura 1, gerada com a ferramenta de <sup>(21)</sup>, é possível constatar este aspecto a partir da sensibilidade média dos 20 preditores RF. Nesse gráfico, os casos com diferentes valores de  $K$  sempre foram superiores ao caso Base. Algo similar ocorreu com SVM e XGB.

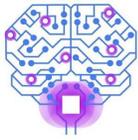


**Figura 1** – Ilustração da sensibilidade obtida por *Random Forest* nas variações sem SA tradicional (G1), CFS tradicional (G2) e CFS aninhado com  $P=100\%$  (G3),  $75\%$  (G4) e  $50\%$  (G5)

Embora *oversampling* está relacionada a menos casos destacados na Tabela 1 (cinco contra oito dos 13 realçados), ela contribuiu para se rotular corretamente mais melanomas, conseguindo melhor sensibilidade. Além disso, ainda que as especificidades médias – Tabela 3 – frequentemente reduziram com essa técnica, elas seguem relativamente altas. De fato, em muitos algoritmos de classificação se prefere categorizar a classe majoritária, o que pode explicar a especificidade média considerada satisfatória.

Convém justificar decisões tomadas na avaliação experimental deste trabalho. A técnica simples de segmentação se beneficiou da disponibilidade de máscaras no repositório ISIC para que as etapas seguintes do método pudessem focar a análise em cada lesão e em trechos próximos de pele. A extração de descritores de textura, como as de Haralick e Laws, tem levado a resultados competitivos em análise de imagens médicas com um custo computacional geralmente proporcional às dimensões da imagem (4,22). A sobreamostragem, representada pelo popular algoritmo SMOTE, foi adotada como alternativa à subamostragem do rótulo majoritário para tratar desbalanceamento, dado que o conjunto de dermoscopias investigado é relativamente pequeno.

CFS é recomendado no Weka para seleção filtro de subconjuntos de atributos (11). A seleção de subconjuntos destaca-se contra o *ranking* por possibilitar a identificação de um grupo de características com pouca redundância (12). RF e SVM são frequentemente



empregados pela comunidade de AM <sup>(5,14,15,6,8)</sup>, enquanto XGB tem dominado algumas competições recentes da mesma comunidade <sup>(10)</sup>. O uso de valores *default* nos parâmetros reduziu o número de variáveis a serem estudadas. Finalmente, a validação cruzada tradicional permitiu a escolha de um subconjunto com os melhores descritores especificamente para cada *fold*  $F$ , enquanto na modalidade aninhada se tratou o subconjunto como um metaparâmetro determinado a partir de diferentes partições de  $F$ .

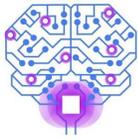
### Comparação com a Literatura

Ao considerar estudos realizados com descritores de textura sobre o mesmo conjunto de 274 dermoscopias <sup>(14,15)</sup>, – Tabela 4 –, observa-se que os modelos RF destacados anteriormente, M1 e M2, levaram a acurácia média melhor que todas as configurações de <sup>(14)</sup>. Além disso, a única variação superior nesta medida em <sup>(15)</sup> resultou em uma sensibilidade média de 0,27, inferior à de M2. Porém, quando se leva em conta o desvio padrão, nota-se que com as quatro abordagens se obteve desempenho similar.

**Tabela 4** – Desempenho dos melhores modelos deste trabalho, M1 e M2, e da literatura construídos sobre o mesmo conjunto de imagens médicas

	<b>M1</b>	<b>M2</b>	<sup>(14)</sup>	<sup>(15)</sup>
<b>Acurácia porcentual</b>	83,57 (6,29)	83,57 (5,01)	83,17 (7,52)	83,93 (5,85)
<b>Sensibilidade</b>	0,25 (0,21)	0,38 (0,18)	0,21 (0,16)	0,27 (0,25)
<b>Especificidade</b>	0,96 (0,05)	0,93 (0,05)	0,96 (0,07)	0,96 (0,05)

A comparação de M1 e M2 com <sup>(14,15)</sup> deve levar em conta que existem diferenças nas abordagens de VC e nas características adotadas. Em particular, em <sup>(14,15)</sup> as imagens descritas por 19 atributos de Haralick e de *Neighborhood Grey Tone Difference Matrix* <sup>(23)</sup> foram divididas em dez partições estratificadas. Após, sem sobreamostragem, se realizou CFS em cada *fold*, fazendo a união ou a intersecção dos subconjuntos obtidos dentre os *folds* para definir um grupo de atributos mais frequentes. O mesmo grupo foi usado nas dez partições para a geração dos preditores finais. Por sua vez, neste trabalho foram usados 88 descritores de Haralick e de Laws. Cada *fold* foi então representado por um subconjunto distinto no CFS tradicional (M1) e no CFS aninhado com SMOTE (M2).



A comparação indireta deste trabalho com outros da literatura em dermoscopia também é possível, com a ressalva de que o conjunto de imagens e outras variáveis experimentais diferem. O desempenho registrado por M1 e M2 em termos de acurácia porcentual e especificidade médias superou o atingido por alguns métodos revisados em <sup>(8)</sup>, mas foi superado pelo de outros <sup>(1)</sup>, incluindo o de métodos também baseados em sobreamostragem e SA <sup>(5,6)</sup>. Porém, em referências como <sup>(1,5,6)</sup> foram adotadas estratégias como aprendizado profundo (*deep learning*) ou SA dependente de classificador, cujo custo computacional é maior que o dos modelos da Tabela 4.

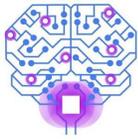
A principal deficiência dos processos implementados neste trabalho foi em termos de sensibilidade. Embora com *oversampling* se promoveu uma melhora nessa medida, ações adicionais devem ser investigadas futuramente para rotular mais melanomas corretamente. Algumas ideias nesse sentido são (1) a adoção de mais esquemas de pré-processamento de imagem <sup>(2)</sup>, (2) a extração e a avaliação de mais características, incluindo as que são baseadas em cor e em forma <sup>(3)</sup> e (3) a realização de *grid search* em parâmetros do XGB <sup>(17)</sup>.

## Conclusão

Neste trabalho foram estudadas técnicas para caracterizar e classificar conjuntos desbalanceados de lesões dermoscópicas. O método implementado, baseado em sobreamostragem e SA filtro, foi aplicado em um conjunto de imagens com melanomas e nevos. As melhores variações, M1 e M2, foram promissoras e competitivas, indicando que (1) a SA levou a desempenho similar ao de configurações com mais características e (2) a sensibilidade média melhorou com a sobreamostragem na maioria dos casos. Trabalhos futuros incluem a adoção de mais técnicas de pré-processamento.

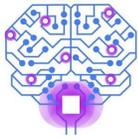
## Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo auxílio financeiro recebido da CAPES – Brasil para Matheus Maciel e Huei Diana Lee. A APB Grezzana pelo apoio.



## Referências

1. Bansal P, Garg R, Soni P. Detection of melanoma in dermoscopic images by integrating features extracted using handcrafted and deep learning models. *Comput Ind Eng.* 2022;168:108060.
2. Kaur R, GholamHosseini H, Sinha R. Hairlines removal and low contrast enhancement of melanoma skin images using convolutional neural network with aggregation of contextual information. *Biomed Signal Process Control.* 2022;76:103653.
3. Pathan S, Ali T, Vincent S, Nanjappa Y, David RM, Kumar OP. A dermoscopic inspired system for localization and malignancy classification of melanocytic lesions. *Appl Sci (Basel).* 2022;12(9):4243.
4. Popecki P, Jurczyszyn K, Ziętek M, Kozakiewicz M. Texture analysis in diagnosing skin pigmented lesions in normal and polarized light - a preliminary report. *J Clin Med.* 2022 Apr 29;11(9):2505.
5. Alazzam MB, Alassery F, Almulihi A. Diagnosis of melanoma using deep learning. *Math Probl Eng.* 2021;2021:1423605.
6. Javaid A, Sadiq M, Akram F. Skin cancer classification using image processing and machine learning. In: Zafar-Uz-Zaman M, Siddiqui NA, Iqbal M, et al., editors. *Proceedings of the 18th International Bhurban Conference on Applied Sciences and Technologies*; 2021; Islamabad, Pakistan. [New York]: Curran Associates; 2021. p. 439-44.
7. Valdés-Morales KL, Peralta-Pedrero ML, Cruz FJ, Morales-Sánchez MA. Diagnostic accuracy of dermoscopy of actinic keratosis: a systematic review. *Dermatol Pract Concept.* 2020;10(4):e2020121.
8. Lee HD, Mendes AI, Spolaôr N, Oliva JT, Sabino Parmezan AR, Chung WF, et al. Dermoscopic assisted diagnosis in melanoma: Reviewing results, optimizing methodologies and quantifying empirical guidelines. *Knowl-Based Syst.* 2018;158:9-24.
9. Instituto Nacional de Câncer (BR). Estimativa 2020: incidência de câncer no Brasil [Internet]. Rio de Janeiro: Instituto Nacional de Câncer; 2019 [cited 2022 Jul 13]. Available from: <https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//estimativa-2020-incidencia-de-cancer-no-brasil.pdf>.
10. Chollet F, Allaire JJ. *Deep learning in R*. Shelter Island: Manning publications; 2018. 335 p.
11. Witten IH, Frank E, Hall MA, Pal CJ. *Data mining: practical machine learning tools and techniques*. 4a. ed. Burlington: Morgan Kaufmann; 2016. 654 p.



12. Liu H, Motoda H. Computational methods of feature selection. Boca Ratón: Chapman & Hall/CRC; 2007. 411 p.
13. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16(1):321-57.
14. Grezzana APB, Lee HD, Spolaôr N, Wu FC. Extração e Seleção de Atributos para Processamento e Análise de Imagens Médicas. In: Pró-Reitoria de Pesquisa da Universidade de São Paulo, editor. Anais do Simpósio Internacional de Iniciação Científica e Tecnológica da USP; 2021; São Carlos, Brasil. São Paulo: Universidade de São Paulo; 2021. p. 1-1.
15. Grezzana APB, Lee HD, Spolaôr N, Wu FC. Segmentação, Caracterização e Classificação de Imagens Dermoscópicas Usando Seleção de Atributos. In: Comitê Assessor de Bolsas de Iniciação Científica da Universidade Estadual do Oeste do Paraná, editor. Anais do Encontro Anual de Iniciação Científica, Tecnológica e Inovação da Unioeste; 2021; Cascavel, Brasil. Cascavel: Universidade Estadual do Oeste do Paraná; 2021. p. 1-1.
16. Merck Sharp & Dohme. Nevos [Internet]. Rahway: Merck Sharp & Dohme; 2020 [cited 2022 Jul 13]. Available from: <https://www.msdsmanuals.com/pt-br/casa/dist%C3%BArbios-da-pele/tumores-cut%C3%A2neos-n%C3%A3o-cancerosos/nevos>.
17. Kuhn M, Johnson K. Applied predictive modelling. New York: Springer; 2013. 613 p.
18. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973;SMC-3(6):610-21.
19. Laws KI. Texture energy measures. In: Baumann LS, editor. Proceedings of the Defense Advanced Research Projects Agency Image Understanding Workshop; 1979; Los Angeles, United States. [Arlington]: Science Applications; [1979?]. p. 47-51.
20. Smit S, Hoefsloot HCJ, Smilde AK. Statistical data processing in clinical proteomics. J Chromatogr B Analyt Technol Biomed Life Sci. 2008;866(1):77-88.
21. Carvalho VAM, Spolaôr N, Cherman EA, Monard MC. A Framework for Multi-label Exploratory Data Analysis: ML-EDA. In: Ezzatti P, Delgado A, editors. Proceedings of the Latin American Computing Conference; 2014; Montevideu, Uruguai. [New York]: Curran Associates; 2014. p. 1-12.
22. Oliva JT, Lee HD, Spolaôr N, Coy CSR, Chung WF. Prototype system for feature extraction, classification and study of medical images. Expert Syst and Appl. 2016;63:267-83.
23. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern. 1989;19(5):1264-74.