

# Classificação de Tumores a Partir de Dados Tabulares e Imagens Utilizando Redes Neurais

PEDRO HORTÊNCIO MOREIRA ROSA

UFG – Universidade Federal de Goiás  
INF – Instituto de Informática  
Goiânia - GO

**Resumo:** Este artigo propõe dois modelos de classificação de tumores detectados em tecido mamário: um classificador para dados tabulares compostos de descrições das células e outro classificador para as imagens do tecido colhidas através de exames laboratoriais (histopatológicos). Para o problema relacionado a dados tabulares, foi utilizado o conjunto de dados *Breast Cancer Wisconsin (Diagnostic)*; já a classificação de imagens se baseia em um recorte de amostras do conjunto *Breast Cancer Histopathological Database (BreakHis)*. O modelo de classificação de dados tabulares, uma Rede Perceptron Multicamadas, obteve 98.6% de acurácia no conjunto de testes, com um total de dois erros absolutos. O modelo de classificação de imagens, composto majoritariamente por uma Rede Neural Convolucional, obteve 87.2% de acurácia no conjunto de testes.

**Palavras Chaves:** Classificação de tumores de mama, Rede Perceptron Multicamadas, Rede Neural Convolucional.

## 1 Introdução

Atualmente, o câncer é um dos maiores problemas de saúde enfrentados pela sociedade. Estima-se que 10 milhões de pessoas, mundialmente, morreram de câncer apenas em 2020 [1]. Foram diagnosticados 2,26 milhões de novos casos de câncer de mama no mesmo período em todo o mundo.

No Brasil, foram diagnosticados 309.750 neoplasias malignas em homens e 316.280 em mulheres [2]. O câncer de mama, em especial, representou 29,7% dos diagnósticos em mulheres em 2020 e 16,4% dos óbitos relacionados a câncer em 2019.

O diagnóstico precoce do câncer é tido como um elemento crucial no tratamento e prognóstico da doença, segundo a OMS [3]: “Early diagnosis of symptomatic cancers is relevant in all settings and the majority of cancers. Cancer programmes should be designed to reduce delays in, and barriers to, diagnosis, treatment and care.”

A proposição de sistemas que auxiliem no diagnóstico, possibilitando a realização de diagnósticos com mais rapidez e conferindo maior produtividade aos profissionais de saúde envolvidos, é exposta em [2]:

“Esperamos que as estimativas 2020-22 promovam um olhar além da técnica, sejam mais voltadas para as boas práticas de controle do câncer e sirvam como um catalisador que estimule gestores, profissionais de saúde, pesquisadores, comunicadores e a sociedade em geral para a reflexão acerca da questão: **precisamos buscar a qualificação e sustentabilidade dos sistemas de informação sobre câncer, assim como demais sistemas em saúde, para que os resultados obtidos criem os meios adequados para enfrentar os desafios de hoje** e subsidiem, no amanhã, o desenvolvimento de ações eficazes para a saúde pública no Brasil.”

Graças aos avanços em inteligência artificial, temos a oportunidade de transformar interpretações de exames e dados de pacientes de uma tarefa qualitativa para uma ação quantitativa e reprodutível. Isso se dá graças à capacidade de generalização de métodos modernos de inteligência artificial, como as redes neurais convolucionais, utilizadas para reconhecimento de padrões complexos em imagens.

Há previsões de que modelos que utilizam inteligência artificial irão revolucionar o modo como lidamos com diagnósticos de tumores, levando a uma

diminuição da necessidade de radiologistas e patologistas [4].

De fato, já há registros de modelos que superaram diagnósticos humanos. Um algoritmo treinado por pesquisadores da Google, em 2019, superou 6 radiologistas na detecção de câncer de pulmão [5].

Este artigo propõe, então, dois algoritmos para classificação de tumores em tecido mamário, criados utilizando dois formatos de dados: tabulares e imagens.

## 2 Metodologia

Ambos os classificadores propostos foram construídos com base no paradigma de redes neurais artificiais. O classificador de dados tabulares é composto de uma rede perceptron multicamadas (conhecida como Rede Neural Densa e Rede Neural Totalmente Conectada, do tipo *feedforward*). Já o classificador que tem como base imagens é composto por uma Rede Neural Convolucional cujo vetor de características resultante alimenta uma Rede Perceptron Multicamadas.

Foi utilizado, em ambos os casos, o TensorFlow (através do framework Keras) para a construção dos modelos, por meio da linguagem de programação Python.

### 2.1 Breast Cancer Wisconsin (Diagnostic) Data Set

A fonte de dados tabulares para a criação do modelo foi o conjunto de dados *Breast Cancer Wisconsin (Diagnostic)* disponibilizado no UCI Machine Learning Repository [6].

As características presentes foram computadas a partir de imagens provenientes de punção aspirativa por agulha fina (PAAF) da massa mamária. As características descrevem o núcleo celular presente na imagem. Os autores do conjunto de dados calcularam 10 características, com valores reais, para cada um dos núcleos celulares:

**Raio:** média das distâncias do centro para os pontos no perímetro.

**Textura:** desvio padrão de valores em escala de cinza.

**Perímetro**

**Área**

**Suavidade (*smoothness*):** variação local na medida do raio.

$\frac{\text{perimetro}^2}{\text{area}} - 1.0$

**Compacidade** (*compactness*):  $\frac{\text{perimetro}^2}{\text{area}} - 1.0$   
**Concavidade** (*concavity*): severidade de porções côncavas do contorno.  
**Pontos côncavos**: número de porções côncavas do contorno.  
**Simetria**  
**Dimensão fractal**:  $\text{coastline approximation} - 1$

Essas 10 características são repetidas 3 vezes, com valores de média, erro padrão e *worst* (maior valor computado). Assim, são ao todo 30 valores para cada imagem. Há também a presença do diagnóstico (maligno ou benigno) e a ID da imagem.

O conjunto de dados possui 569 amostras, divididas em 212 malignas e 357 benignas. Os valores foram estandardizados (variância unitária e média zerada). A divisão de dados de treinamento está informada na Tabela 1. É observado um leve desbalanceamento entre as classes, com mais amostras de imagens com tumores benignos. Como os dados são relativamente simples, não foi observada uma penalização do modelo em tempo de treinamento ou de teste.

Para evitar o uso de um terceiro conjunto de dados, o processo de validação do modelo foi feito através do processo *K-Folds*, em sua variante estratificada, onde o conjunto de treino é dividido em *K* conjuntos, cada um com uma porção separada para validação. No cenário deste artigo, *K* foi definido como 5.

	Benigno (1)	Maligno (0)
Treino	268	158
Teste	89	54

Tabela 1 - Divisão de amostras do conjunto *Breast Cancer Wisconsin*.

### 2.1.1 Modelo Proposto (Wisconsin)

O modelo proposto para a resolução do problema é uma rede neural artificial densa, composta de 4 blocos:

Bloco 1 - Camada densa com 100 unidades + camada de ativação com a função *rectified linear unit* ReLU + camada de *dropout*

configurado como 0,2.

Bloco 2 - Camada densa com 50 unidades + camada de ativação com a função *rectified linear unit* ReLU + camada de *dropout* configurado como 0,1.

Bloco 3 - Camada densa com 30 unidades + camada de ativação com a função *rectified linear unit* ReLU

Bloco 4 - Camada de saída, com uma unidade e função de ativação sigmóide.

O modelo foi compilado utilizando o otimizador Adam [7] com taxa de aprendizado ajustada para 0,0005, a função *loss* configurada como *binary cross entropy*. O modelo foi treinado por 20 épocas, com um *batch size* de 64.

A presença de camadas de *dropout*, o *learning rate* pequeno e o treinamento por apenas 20 épocas foram esforços para evitar que o modelo apresentasse *overfitting*.

## 2.2 Breast Cancer Histopathological Database (BreakHis)

Em relação ao classificador de imagens, foi utilizado um recorte do conjunto *Breast cancer Histopathological Database*, disponibilizado pela Universidade Federal do Paraná [8].

O conjunto de dados completo possui 9109 imagens de tumores de mama, distribuídos em diferentes fatores de ampliação: 40X, 100X, 200X e 400X, como ilustrado na Figura 1.

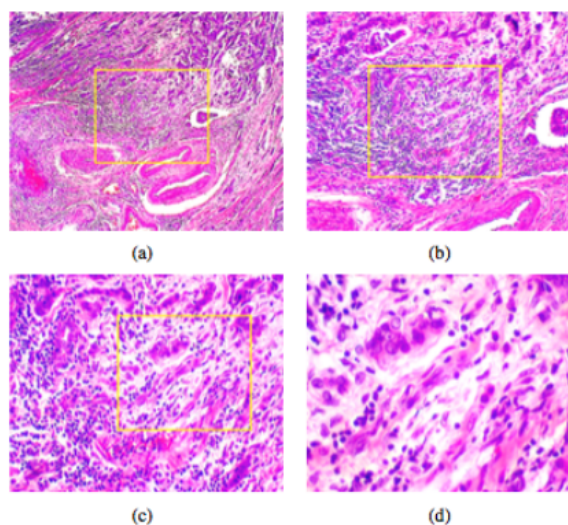


Imagem 1 - Tumor maligno (colorizado com HE) com diferentes fatores de ampliação: (a) 40X, (b) 100X, (c) 200X, and (d) 400X. [1]

As imagens possuem 700 x 460 pixels e 3 canais RGB com profundidade de 8-bit.

O recorte utilizado para este artigo foram apenas as imagens referentes ao fator de ampliação 400X. O recorte contém um total de 1820 imagens, sendo 588 representando amostras benignas e 1232 representando amostras malignas. A distribuição entre os conjuntos está especificada na Tabela 2.

	Benigno (0)	Maligno (1)
<b>Treino</b>	441	924
<b>Teste</b>	147	308

Tabela 2 - Distribuição de dados do conjunto BreakHis.

Foram reservados 272 imagens do conjunto de treino para validação. A divisão foi feita em tempo de treinamento, utilizando a função *Image Data Generator* do framework Keras.

Devido à divisão de amostras, com uma classe contendo o dobro de amostras da outra, houve dificuldades em encontrar um equilíbrio durante o treinamento do modelo. Como o total de amostras também não é abundante, foi adotado a técnica de *data augmentation* para aumentar, virtualmente, o universo de amostras de treinamento e validação. As seguintes modificações foram feitas aleatoriamente às imagens:

**Alteração no brilho** (*brightness range*): variando de 0,95 a 1,05, sendo 1 o valor da imagem sem alteração.

**Alteração no canal de cores** (*channel shift range*): alteração na coloração em até 30 pontos.

**Alteração na proximidade** (*zoom range*): 0,95 a 1,05, sendo 1 o valor da imagem sem alteração.

**Espelhamento horizontal e vertical** (*horizontal and vertical flip*): aleatoriamente espelha, vertical ou horizontalmente, a imagem. Esta foi a única alteração realizada no conjunto de validação.

Após as alterações, realizadas pela classe *Image Data Generator* do framework Keras, as imagens são escaladas através da divisão de todos os valores por 255. As alterações podem ser vistas na Figura 2. As imagens sem alteração estão representadas na Figura 3.

### 2.2.1 Modelo Proposto

A exemplo do problema relacionado a dados tabulares, foi utilizada uma rede neural artificial densa para a classificação das imagens. Entretanto, as imagens são processadas por uma rede convolucional antes de serem alimentadas à parte densa. O modelo proposto **não** é sequencial. Os blocos que compõe o modelo são:

**Bloco 2:** camada convolucional 2D, com 64 filtros e kernel de tamanho (3, 3) + camada de ativação composta por função ReLU + camada de *pooling* do tipo valor máximo (Max Pooling 2D), com *padding* ajustado como *same* (a saída possui a mesma altura e largura da entrada).

**Bloco 2:** idêntico ao Bloco 1, sendo a única diferença o fato de que a camada Max Pooling 2D possui o *padding* ajustado como *valid*.

**Bloco 3:** composto por 3 camadas que se conectam à camada Max Pooling 2D do Bloco 2. A função deste bloco é simplesmente transformar os outputs do Bloco 2 em tensores 1D. Cada camada executa essa ação de uma maneira diferente. *Global Max Pooling 2D* e *Global Average Pooling 2D* são variações de uma mesma técnica proposta por [9] e formam duas camadas. A terceira é simplesmente uma camada *Flatten*, que concatena o tensor 2D em um tensor 1D, que possui um *dropout* de 0,2.

**Bloco 4:** concatenação das 3 camadas do Bloco 3 + *dropout* de 0,2.

**Bloco 5:** camada de saída, com duas unidades e função de ativação *softmax*.

O modelo foi compilado com o otimizador Adam ajustado com *learning rate* de 0.01, função *loss* como *binary crossentropy* e métricas de monitoramento: acurácia, AUC e Recall.

O modelo foi treinado por até 50 épocas com um *batch size* definido em 50. O limite não foi atingido em nenhuma iteração de treinamento, pois foi configurada uma *callback* que interrompe o treinamento caso haja estagnação da *loss* calculada sobre o conjunto de validação durante 10 épocas. Foi implementada, também, uma *callback* que diminui o *learning rate* em 50% (até um limite de 0.0001) caso haja estagnação da *loss* calculada sobre o conjunto de validação durante ao menos 5 épocas.

### Amostras de Imagens do Conjunto de Treino

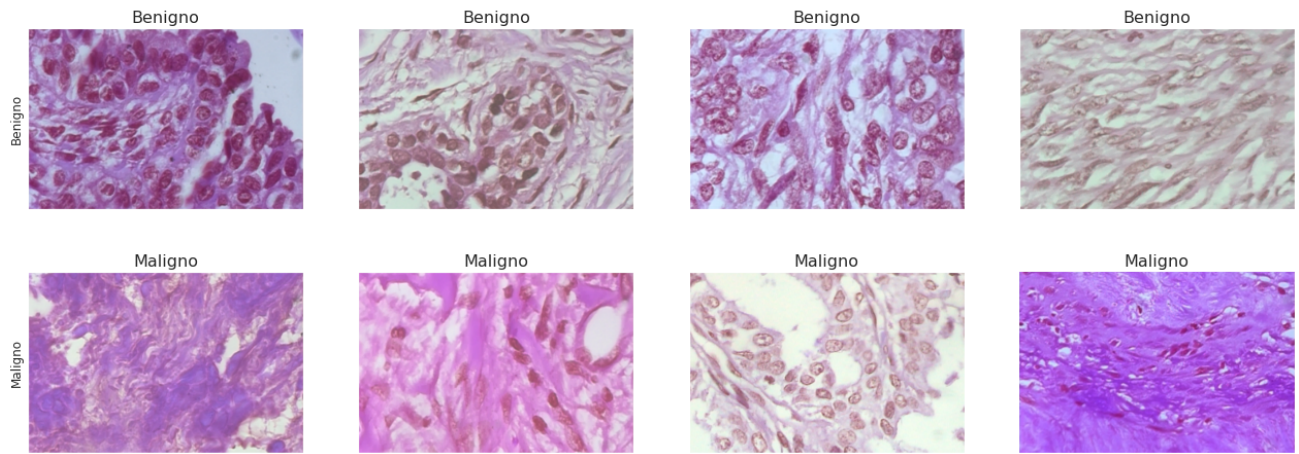


Figura 2 - Amostras de imagens do conjunto BreakHis sem alterações.

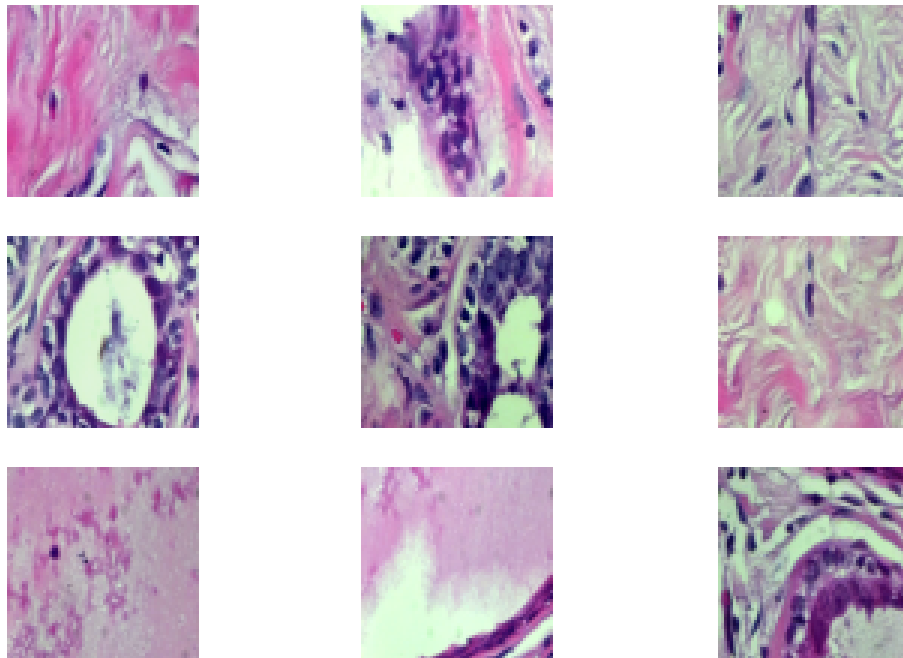


Figura 3 - Imagens alteradas via *Image Data Generator*. Possuem todas as alterações listadas na seção 2.2.

### 3 Resultados

Ambos os modelos tiveram resultados bastante satisfatórios, considerando que são razoavelmente simples e foram treinados com uma quantidade reduzida de dados. Em ambos os casos, houve uma dificuldade inicial devido à distribuição das amostras entre as classes. Isso ficou claro especialmente com o modelo de visão computacional, no conjunto de dados BreakHis.

#### 3.1 Wisconsin Dataset

O modelo foi treinado primeiramente em uma estratégia de validação cruzada. Foi utilizado o método *K-Folds*, com estratificação da variável-alvo (diagnóstico), de modo que cada subdivisão dos dados preservasse a distribuição de amostras. Foram realizadas 5 divisões distintas.

A acurácia média do modelo, ao fim dos 5 treinamentos, foi de 96,94%. O modelo, então, foi treinado novamente, desta vez utilizando todas as amostras disponíveis para treinamento. Os do relatório de classificação disponibilizado pela biblioteca *scikit-learn* podem ser vistos na Tabela 3.

Métrica	0	1	Média
Acurácia	-	-	0,99
Precisão	0,98	0,99	0,99
Recall	0,98	0,99	0,99
F1-Score	0,98	0,99	0,99

Tabela 3 - Métricas dadas pelo relatório de classificação. Classe 0: maligno. Classe 1: Benigno.

A acurácia no conjunto de testes sem arredondamento de duas casas foi de 98,60%. Foi realizada a execução do modelo **uma única vez** no conjunto de testes. A pontuação da curva ROC AUC, que mede a taxa de falsos positivos e falsos negativos, foi de 0.9851 e a curva pode ser vista na Figura 4.

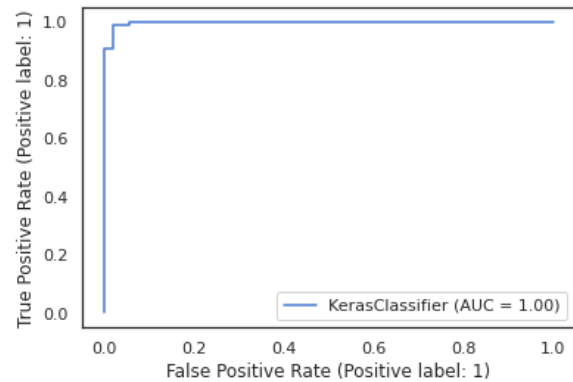


Figura 4 - Curva ROC AUC gerada a partir das predições do conjunto de testes.

Por fim, a matriz de confusão, disponível na Figura 5, informa um erro em cada classe.

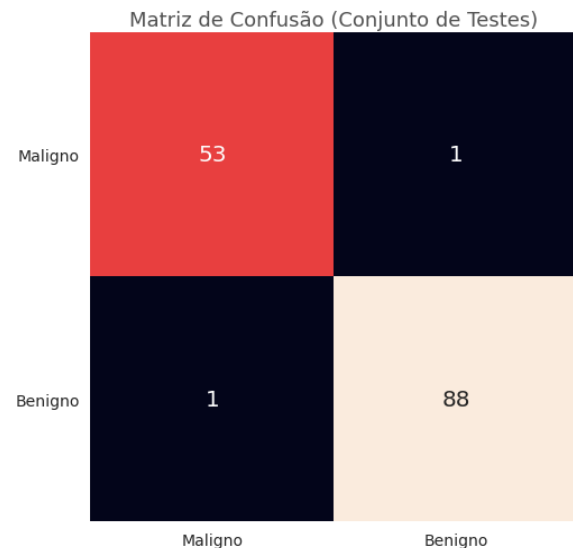


Figura 5 - Matriz de confusão gerada a partir da predição no conjunto de testes.

### 3.2 BreakHis

O modelo chegou ao final do treinamento após 38 épocas. A melhor época foi a de número 28, sendo os pesos restaurados após o fim do treinamento. Os resultados indicam uma acurácia no conjunto de validação de 0.8880 e 0.8518 no conjunto de treino. As métricas do treinamento podem ser vistas na Tabela 4 e os gráficos de *loss* e acurácia nas Figuras 5 e 7, respectivamente.

As métricas relativas ao conjunto de testes estão presentes na tabela 5. O modelo obteve, no conjunto de testes, métricas similares ao conjunto de validação, o que indica não haver *overfitting*. De modo similar ao modelo anterior, a aferição no conjunto de testes ocorreu **apenas uma vez**.

Métrica	Treino	Validação
Loss	0.3974	0.3963
Acurácia	0.8518	0.8880
AUC	0.8934	0.9505
Recall	0.8518	0.8880

Tabela 4 - Métricas na melhor época (28) utilizada definitivamente no modelo.

Métrica	Pontuação
Acurácia	0.8725
ROC AUC	0.8436
Recall	0.9253
Precision	0.8906
F1-Score	0.9076

Tabela 5 - Métricas de avaliação das predições realizadas no conjunto de testes.

Os resultados obtidos se mostram comparáveis, e ligeiramente superiores, aos originalmente encontrados pelos autores originais do conjunto de dados [10]. A matriz de confusão proveniente das predições nos dados de teste está expressa na Figura 8.

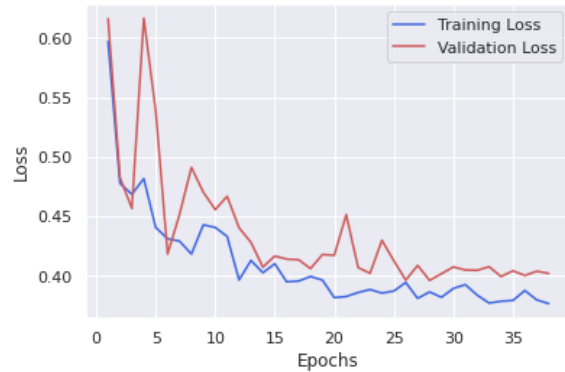


Figura 6 - Gráfico de *loss* durante as épocas de treinamento.

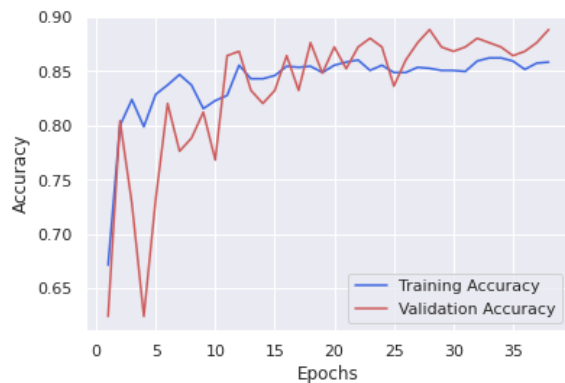


Figura 7 - Gráfico de acurácia durante as épocas de treinamento.

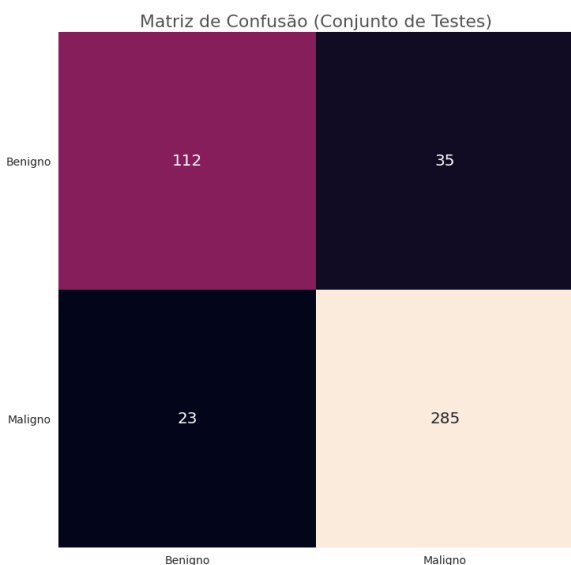


Figura 8 - Matriz de confusão do conjunto de testes do dataset BreakHis.

## 4 Conclusões

Ambos os modelos apresentados têm performance aceitável e problemas relacionados aos dados utilizados para o treinamento. O classificador de dados tabulares tem uma performance quase perfeita, com sua acurácia no conjunto de testes de 99%, porém seria necessário mais dados de avaliação para verificar de maneira mais extensiva suas limitações.

Entretanto, entende-se que, dado o trabalho de se montar informações em dados tabulares, modelos que consomem imagens são mais desejáveis. Tendo em vista essa ótica, o classificador referente ao conjunto BreakHis tem desempenho razoável levando em conta outros classificadores treinados no mesmo conjunto de dados. A alta taxa de erros cometida na classe de tumores benignos (label 0) tem íntima relação com a proporção significativamente menor de amostras benignas. Isso significou uma maior necessidade de técnicas que simulam um maior número de imagens (como o processo de *data augmentation* utilizado), além de uma limitação na arquitetura do modelo. Amostras desbalanceadas implicam em uma tendência dos algoritmos de conferirem maior peso a uma das classes, levando a uma dificuldade de generalização e, conseqüentemente, a uma restrição ao uso de modelos mais robustos e avançados.

Técnicas modernas, como aprendizado por transferência (*transfer learning*), podem ser usadas para substituir a estrutura convolucional da rede neural,

conferindo uma maior capacidade de extração de características ao modelo. Entretanto, é necessário que haja mais imagens disponíveis para treinamento, pois, tendo em vista a complexidade dos modelos estado da arte utilizados no *transfer learning*, há uma forte tendência ao *overfitting* em modelos construídos utilizando esse paradigma.

Reforça-se a mensagem transmitida na introdução: inteligência artificial possui capacidade de revolucionar a saúde pública. Os algoritmos apresentados foram implementados sem grande esforço e utilizando pouco recurso computacional, sendo treinados em minutos, algo impossível há décadas atrás. Seguindo essa perspectiva e procurando resolver os empecilhos apontados, é perfeitamente possível, e até esperado, que algoritmos de inteligência artificial superem a acurácia humana em diagnósticos de tumores de mama.



## 5 Referências

- [1] Ferlay J. Ervik M. Lam F. Colombet M. Mery L. Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer; 2020 (<https://gco.iarc.fr/today>, acesso em Maio de 2021).
- [2] Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Estimativa 2020; (<https://www.inca.gov.br/estimativa/estado-capital/brasil>, acesso em Maio de 2021).
- [3] Organização Mundial de Saúde. Fact Sheets: Cancer; (<https://www.who.int/news-room/fact-sheets/detail/cancer>, acesso em Maio de 2021).
- [4] Obermeyer, Ziad. Emanuel, Ezekiel J. “Predicting the Future — Big Data, Machine Learning, and Clinical Medicine”. *The New England Journal of Medicine* 375, 1216-1219, (2016). <https://www.nejm.org/doi/full/10.1056/NEJMp1606181>.
- [5] Ardila, D., Kiraly, A.P., Bharadwaj, S. et al. “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography”. *Nature Medicine* 25, 954–961 (2019). <https://doi.org/10.1038/s41591-019-0447-x>.
- [6] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- [7] Kingma, D. P. & Ba, J. “Adam: A Method for Stochastic Optimization”. arXiv, 1412.6980 (2017). <https://arxiv.org/abs/1412.6980>.
- [8] Spanhol, F., Oliveira, L. S., Petitjean, C., Heutte, L., A Dataset for Breast Cancer Histopathological Image Classification, IEEE Transactions on Biomedical Engineering (TBME), 63(7):1455-1462, 2016. <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/> (Acesso em Maio de 2021).
- [9] Min Lin and Qiang Chen and Shuicheng Yan. “Network In Network”. arXiv, 1312.4400 (2014). <https://arxiv.org/abs/1312.4400>.
- [10] Spanhol, F., Oliveira, L. S., Petitjean, C., and Heutte, L., Breast Cancer Histopathological Image Classification using Convolutional Neural Network, International Joint Conference on Neural Networks (IJCNN 2016), Vancouver, Canada, 2016. <https://ieeexplore.ieee.org/document/7727519>.
- [11] Repositório GitHub: Cancer Detection. <https://github.com/pedrohortencio/cancer-detection>

## 6 Disponibilidade de Código

Os códigos fonte dos algoritmos, e, no caso do modelo de visão computacional, os pesos em formato H5 e o modelo salvo, estão disponíveis no GitHub [11]. Os arquivos referentes ao modelo foram fornecidos para que se possa, eventualmente, reutilizar o algoritmo sem que haja a necessidade de novo treinamento (o que geraria inconsistência nos resultados, tendo em vista o caráter aleatório da inicialização de pesos das redes neurais).