Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Classificador Baseado em Comitê de Máquinas de Vetores de Suporte Aplicado a Imagens de Câncer de Mama

Darielson A. de Souza[1], Roberto I. M. F. Oliveira[2]
IMPA, Rio de Janeiro, RJ
Gilson A. Giraldi[3]
LNCC, Petrópolis, RJ

Resumo. Os índices elevados de casos de câncer de mama tornam necessário o desenvolvimento de tecnologias que sejam capazes de auxiliar a medicina em diagnósticos desta doença. Nesse contexto, o objetivo deste trabalho é o desenvolvimento de um modelo de comitês de SVMs (*Support Vector Machines*) e sua comparação com outros métodos de aprendizagem de máquina para classificação de imagens médicas de câncer de mama que foram adquiridas por meio de um microscópio de alta resolução. A classificação de comitês de SVMs se dá através do voto majoritário. Além do pré-processamento dos dados, fez-se várias execuções de treinamento/teste usando a validação *hold-out* com todos os modelos. A principal conclusão é que o modelo baseado em comitê de SVMs conseguiu ter um desempenho melhor que os demais métodos analisados, implementados usando os classificadores Vizinho mais próximo (KNN), Árvore de decisão, Floresta aleatória, CNN (*Convolutional Neural Network*) e Comitê de CNNs.

Palavras-chave. Imagens de câncer de mama, comitê, classificação de padrões, aprendizagem de máquina

1 Introdução

O câncer de mama é uma doença que se origina nas células mamárias, normalmente nos ductos (tubos que transportam o leite) ou nos lóbulos (glândulas produtoras de leite). É também um dos tipos mais comuns de câncer diagnosticados em mulheres em todo o mundo [2]. Além da gravidade desta doença, acrescenta-se o fato da mesma trazer desconfortos físicos, como mostra a referência [11], segundo a qual dor em mulheres com câncer de mama ocorre em cerca de 47% dos casos, aumentando com a evolução da doença. Por iniciar num tecido epitelial é configurado como um carcinoma, sendo o carcinoma ductal invasivo (*invasive ductal carcinoma* - IDC) um dos mais comuns.

O trabalho [18], explora o uso de técnicas de aprendizado profundo para diagnóstico de câncer de mama usando imagens de ultrassonografia. Os resultados mostraram uma melhora significativa na precisão diagnóstica em comparação com os métodos tradicionais, no caso, como árvore decisão, e vizinho mais próximo (KNN), por exemplo. Já o trabalho de [16], mostra um estudo de detecção de câncer de mama com redes neurais convolucionais (*convolutional neural networks* - CNNs) em imagens de mamografia. Os resultados demonstram que o modelo proposto alcançou alta precisão na detecção de tumores malignos. No trabalho [13], os pesquisadores exploraram a aplicação de algoritmos de aprendizado de máquina (SVM, CNN e Vizinho mais próximo (KNN)) para

[1]daryewson@gmail.com
[2]rimfo@gmail.com
[3]gilson@lncc.br

identificar subtipos de câncer de mama com base em perfis de expressão gênica. Os resultados revelaram a existência de subtipos distintos e forneceram pistas sobre as características moleculares associadas a cada subtipo. O artigo de [12] traz a previsão de resposta à terapia neoadjuvante no câncer de mama usando regressão logistica bayesiana. Os pesquisadores utilizaram dados clínicos e de imagem para treinar modelos preditivos e obtiveram resultados promissores na identificação de pacientes que responderiam positivamente ao tratamento. Alguns trabalhos da literatura, tais como [9] e [1], treinaram CNNs usando a mesma base de nosso artigo (vide [8]), para reconhecimento IDC (casos positivos contra casos negativos). A referência [9] aplica a arquitetura AlexNet ao reconhecimento IDC, enquanto que o trabalho [1] aplica um modelo de CNN dos próprios autores e a rede VGG-16 para o mesmo problema, ambos trabalhos com bons resultados.

Este artigo consiste no desenvolvimento de um modelo de comitê para a identificação de câncer de mama em imagens que foram adquiridas por meio de um microscópios de alta resolução, que podem ser acessadas pelo site [8]. O modelo proposto deve garantir uma acurácia significativa, a fim de auxiliar o profissional na leitura e interpretação de exames clínicos e consequentemente no diagnóstico. A principal contribuição do trabalho se dá pela utilização de um comitê de SVMs com diferentes núcleos (*kernels*) aplicados ao conjunto de dados de mama com casos postivos e negativos de IDC. Foi realizada a comparação entre o método proposto e as seguintes técnicas da literatura: KNN, Árvore de decisão, Floresta aleatória, CNN e comitê de CNNs. Os resultados obtidos demonstram a eficiência do comitê de SVMs proposto, destacada também quando comparada com os trabalhos [1], [15] e [9].

O trabalho encontra-se estruturado em 4 seções. A seção 2 trata da abordagem proposta, a seção 3 aborda uma discussão dos resultados, e, por fim, a seção 4 apresenta a conclusão e trabalhos futuros.

2 Metodologia

A metodologia para desenvolvimento do trabalho baseia-se em 3 etapas: (I) aquisição de dados, ou seja, conjunto de dados com imagens médicas. (II) Em seguida, vem o pré-processamento, nessa etapa ocorre o tratamento de dados e divisão para o treinamento e teste. (III) Por último, o treinamento do modelo e a validação.

No conjunto de dados existem 2759 exemplos negativos (não IDC) e 2788 exemplos positivos (com IDC), podendo ser considerado um conjunto balanceado. Cada imagem X_k, $k = 1, 2, ..., N$ é uma matriz multidimensional representada por um tensor de dimensões $H \times W \times C$, onde $H = 50$ é a altura da imagem, $W = 50$ é a largura da imagem e $C = 3$ é o número de canais de cor, no caso, R (red),G (green) e B (blue). Algumas imagens do conjunto de dados podem ser visualizadas na Figura 1.

Após a leitura da base de dados, as imagens são normalizadas gerando um novo tensor $\overline{X}_k$, um passo importante no pré-processamento dos dados para o treinamento de modelos de aprendizado de máquina. Foi usada uma abordagem em que fez-se uma divisão pelo valor máximo de cada canal, que é 255; ou seja:

$$\overline{X}_k(i,j) = \left(\frac{R_k(i,j)}{255}, \frac{G_k(i,j)}{255}, \frac{B_k(i,j)}{255} \right).$$

Uma outra etapa do pré-processamento foi a vetorização dos dados a qual transforma cada tensor $\overline{X}_k$ em um vetor $x_k \in \mathbb{R}^m$ onde $m = H \cdot W \cdot C$. Os vetores obtidos são utilizados como entrada para o SVM, KNN, árvore de decisão e floresta aleatória.

DOI: 10.5540/03.2025.011.01.0359

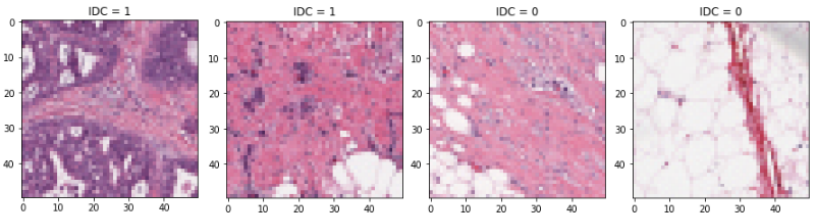

Figura 1: Imagens médicas da base de dados [8] utilizada no trabalho, mostrando a alta variabilidade intra-classe, onde $IDC = 1$ e $IDC = 0$ significam positivo e negativo, respectivamente. Fonte: [8].

2.1 SVM

Dado um conjunto de treinamento $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_N)\} \in \mathbb{R}^m \times \{-1, 1\}$, onde x_i é o vetor de características da amostra i e $y_i \in \{-1, 1\}$ é a classe correspondente, a SVM linear busca encontrar o hiperplano $w^T x + b = 0$ que maximiza a margem de separação entre as classes. A formulação do problema de otimização da SVM linear pode ser escrita como:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i, \\ \text{sujeito a:} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad \forall i, \end{aligned} \tag{1}$$

onde $||w||^2$ é a norma euclidiana ao quadrado de w, C é o parâmetro de regularização que controla o trade-off entre a margem de separação e o erro de classificação, e ξ_i são as variáveis de folga que permitem lidar com casos em que os dados de treinamento não possuem margem de separação definida no espaço original dos dados [7], [5].

A SVM é capaz de realizar classificação não linear por meio do uso de funções *kernel* e do truque do *kernel* (*kernel trick*). Neste método parte-se de uma função simétrica e positiva-definida $K : \mathbb{R}^m \times \mathbb{R}^m \longrightarrow \mathbb{R}$, que é o *kernel* de um operado integral cujas auto-funções permitem formar uma aplicação $\Phi : \mathbb{R}^m \longrightarrow \mathbb{R}^{n_F}$, onde $\mathbb{R}^{n_F}$ é denominado espaço de caracteristicas, tal que $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, onde '$\cdot$' indica o produto interno usual. O truque do *kernel* consiste em utilizar a última expressão para calcular o hiperplano ótimo no espaço de caracteristicas, sem conhecer explicitamente a aplicação Φ [6, 17]. Desta forma, demonstra-se que a função de decisão que permite separar as duas classes pode ser calculada pela expressão:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b, \tag{2}$$

onde α_i são os multiplicadores de Lagrange associados a um problema de otimização análogo ao problema (1), mas envolvendo a função *kernel*, y_i é o rótulo do vetor x_i e $b \in \mathbb{R}$ é o bias. Assim, $f(x) > 0$ indica que x pertence à classe $+1$ enquanto que $f(x) < 0$ indica que x pertence à classe -1. Na formulação do SVM, um vetor x_i para o qual $\alpha_i \neq 0$ é denominado vetor suporte [17]. Essa abordagem permite o uso de diferentes funções *kernel*, tais como: polinomial, sigmoide e RBF (Radial Basis Function), que permitem modelar relações complexas nos dados [17].

2.2 Comitês

Os Comitês são técnicas que combinam a predição de vários modelos de aprendizado de máquina para obter um classificador mais eficiente [3]. Essas abordagens têm se mostrado eficazes em uma variedade de problemas de classificação e regressão. Ao combinar previsões de vários modelos, é possível obter um desempenho superior em relação a um único modelo. A escolha do método de comitê e dos modelos individuais depende do problema e dos dados em questão. Duas abordagens amplamente reconhecidas são o 'Comitê por votação', em que as previsões de vários modelos são combinadas por média ou votação e o 'Comitê por Amostragem', em que vários modelos são treinados em diferentes conjuntos de dados obtidos por amostragem do conjunto de treinamento original.

2.3 Métodos e Técnicas Utilizados para Comparação

O Comitê SVM proposto no trabalho é baseado em "Comitê com votação", explicado na seção 2.2. Assim, seja $\mathcal{M} = \{SVM_1, SVM_2, ..., SVM_m\}$ o conjunto de SVMs treinadas neste trabalho utilizando a técnica de *hold-out*. Cada $SVM_i \in \mathcal{M}$ é treinada no conjunto de treinamento e gera uma previsão $f_i(x)$ para uma instância de entrada x. As previsões das SVMs individuais são combinadas utilizando uma votação por maioria para obter a previsão final $\hat{y}$, ou seja, a classe mais votada pelas SVMs é selecionada como a predição do comitê para a amostra x. Após seu treinamento, o modelo de comitê é testado usando um conjunto de dados de teste separado que não foi usado durante o treinamento. Isso permite avaliar o desempenho do modelo em dados não observados.

Além do método de comitê de SVMs serão utilizados outros métodos para uma análise comparativa. Os métodos são: KNN [21], Floresta aleatória [14], Árvore de decisão [10], CNN [20] e Comitê de CNNs [19].

Todos os métodos serão treinados/testados da mesma forma, usando a validação cruzada *hold-out*, onde o conjunto de dados é dividido em um conjunto de treinamento, que representa 80% dos dados, e um conjunto de teste, que representa os 20% restantes. Ainda será aplicado o *Gridsearch*, o qual é uma técnica utilizada para encontrar a melhor combinação de hiperparâmetros de um modelo de aprendizado de máquina. Essa técnica envolve a criação de uma grade (*grid*) de valores para os hiperparâmetros e a busca exaustiva para determinar a combinação que produz o melhor desempenho de acordo com uma métrica de avaliação pré-definida, no caso, a acurácia.

A implementação foi realizada em Python. Usou-se o scikit-learn para o *GridSearch*, particionamento dos dados em treinamento/validação, classificadores tradicionais e comitê de SVMs. Para a implementação e treinamento das CNNs foi utilizada a biblioteca 'Tensorflow'.

3 Resultados e Discussões

Todos os *kernels* usados nas SVMs foram não-lineares. A Tabela 1, apresenta os melhores hiperparâmentros obtidos pelo *GridSearch* para o comitê de SVMs (ver definição dos hiperparametros em [4]). O comitê de SVMs proposto é formado por três SVMs, cada um usando uma das funções *kernels* da Tabela 1.

Tabela 1: Melhores hiperparâmentros do comitê de SVMs.

kernel	Tolerância	Gamma	C	Coef0	Grau
RBF	0.01	1.0	2.0	–	–
Sigmoide	0.003	–	1.2	0.02	–
Polinomial	0.00002	1.0	1.5	0.004	9

DOI: 10.5540/03.2025.011.01.0359

No caso dos classificadores tradicionais, o *Gridsearch* obteve as seguintes parametrizações: (a) KNN: k=30 e distancia euclidiana; (b) Modelo de árvore de decisão: profundidade máxima da árvore de 16 níveis; (c) Floresta Aleatória: 100 árvores de decisão e uma profundidade máxima de 13 níveis.

No caso das CNNs utilizadas, a Tabela 2 mostra os melhores hiperparâmetros obtidos pelo *Gridsearch*. Nesta etapa, para cada conjunto de hiperparametros do espaco de busca a rede é treinada por 100 épocas. O melhor conjunto é escolhido analisando a melhor acuracia durante as 100 épocas, com o otimizador sendo o Adam.

As arquiteturas foram definidas experimentalmente tendo cada uma delas duas camadas convolucionais (com filtros 5×5), uma camada de pooling com a operação de *max-poolig* e uma camada totalmente conectada, com 256 neurônios, para a classificação. Essas camadas são intercaladas por duas camadas de pooling de dimensão 2×2 e stride igual a dois. Os número de filtros das duas camadas convolucionais são: CNN e CNN_2 com 32 filtros em cada uma; CNN_1 possui 16 e 32 filtros; CNN_3 possui 64 e 32 filtros. As redes CNN_1, CNN_2 e CNN_3 são treinadas separadamente com os hiperparâmetros da Tabela 2. Em seguida, são agrupadas formando o comitê de votação (seção 2.2) proposto.

Tabela 2: Melhores hiperparâmentros das CNNs utilizadas.

Métodos	Taxa de Aprendizado	Lote	Função de Ativação	otimizador	Taxa de dropout
CNN	0.01	128	ReLU	Adam	0.5
CNN_1	0.01	128	ReLU	Adam	0.5
CNN_2	0.01	64	ReLU	Adam	0.4
CNN_3	0.001	64	ReLU	Adam	0.3

Os hiperparementros foram definidos a partir de um dicionário do *Gridsearch* com o números de filtros escolhido entre os valores $\{16, 32, 64\}$, a taxa de aprendizado, escolhida dentre os valores $\{0.001, 0, 01, 0, 1\}$, o tamanho de lote dentre $\{64, 128, 256\}$ e taxa de dropout na lista $\{0.3, 0.4, 0.5\}$.

A Tabela 3 demonstra a acurácia do comitê de SVMs em comparação com outros métodos da literatura que usaram o mesmo conjunto de dados. As acurácias foram estimadas usando a validação cruzada *hold-out* com 20% para teste e 80% para treinamento, executada 10 vezes. No caso da Tabela 3 apresentamos a melhor acurácia dentre as 10 execuções de cada modelo. Observamos que o comitê de SVMs conseguiu acurácia melhor que os demais da literatura.

Tabela 3: Acurácia do modelo proposto em comparação com abordagens publicadas usando o mesmo conjunto de dados..

Métodos	Melhor acurácia
Comitê de SVMs	**0,8691**
CNN VGG-16 [1]	0,8562
Accept-Reject poolin [15]	0,8541
Alexnet [9]	0,8468

A Tabela 4 mostra os resultados de todos os modelos treinados para este trabalho após 10 execuções de cada método. Para avaliação mais completa, apresenta-se a acurácia média, a acurácia mais baixa, a acurácia mais alta e as dispersões das 10 execuções, computadas sobre os resultados do conjunto de teste.

Conforme apresentado na Tabela 4, o Comitê de SVMs possui as melhores acurácias e o desvio padrão mais baixo, indicando que os resultados de acurácia são consistentes e têm uma pequena variação em torno da média. Isso é positivo, indicando que o modelo é estável e produz resultados consistentes em diferentes conjuntos de dados ou divisões de treinamento/teste.

DOI: 10.5540/03.2025.011.01.0359

Tabela 4: Acurácia e desvio padrão dos modelos.

Métodos	% média	% mais baixa	% mais alta	Desvio padrão	Tempo de execução (s)
Comitê de SVMs	**0,8375**	**0,8015**	**0,8691**	**0,021**	192,0
KNN	0,74846	0,7056	0,7860	0,025	**45,0**
Árvore de decisão	0,78142	0,7317	0,8250	0,029	97,0
Floresta aleatória	0,7906	0,7426	0,8327	0,028	286,0
CNN	0,7828	0,7297	0,8295	0,031	317,0
CNN_1-Comitê	0,7992	0,7546	0,8383	0,026	524,0

A Tabela 4 contabiliza também os tempos das 10 execuções realizadas durante a fase de validação cruzada (treinamento/teste) para cada método. O computador que foi usado é um acer Aspire 3, com um processador AMD Ryzen 7, Clock de 2.3 GHZ, e placa gráfica RADEON GRAPHICS, SSD de 256 G, Memória RAM de 8 G, Sistema Windows 10, 64 bits. O ambiente de desenvolvimento é o *Jupyter notebook* versão 6.3.0, com a versão do python 3.8.8. Vê-se que o comitê de SVMs é mais custoso apenas em comparação com o KNN e a Árvore de decisão.

4 Considerações Finais

O presente artigo apresentou uma abordagem de comitê de SVMs usando o voto majoritário aplicado à classificação de imagens de câncer de mama, obtidas por meio de um microscópio de alta resolução. Outros modelos considerados para comparação foram: KNN, Árvore de decisão, Floresta aleatória, CNN e Comitê de CNNs. Os resultados mostraram a eficiência e eficácia do modelo proposto.

Foram avaliadas as acurácias dos modelos (mínima, média, máxima e desvio padrão) após 10 execuções. Observou-se que o comitê de SVMs obteve um desempenho superior em relação aos demais métodos listados acima, bem como em relação a métodos apresentados na literatura, conforme mostrado na Tabela 3. Como proposta futura, busca-se diminuir a complexidade do comitê de SVMs e testar o método em outras bases.

Referências

[1] A. Mohammad. A, S. Mohammad, S. Iman e R. Reza. "Artificial intelligence in automatic classification of invasive ductal carcinoma breast cancer in digital pathology images". Em: **Medical Journal of the Islamic Republic of Iran** 34 (2020), p. 140.

[2] J. Doe. "Breast Cancer: Causes, Diagnosis, and Treatment". Em: **Journal of Oncology** 25.2 (2020), pp. 123–145.

[3] X. Dong, Z. Yu, W. Cao, Y. Shi e Q. Ma. "A survey on ensemble learning". Em: **Frontiers of Computer Science** 14 (2020), pp. 241–258.

[4] K. Duan, S. S. Keerthi e A. N. Poo. "Evaluation of simple performance measures for tuning SVM hyperparameters". Em: **Neurocomputing** 51 (2003), pp. 41–59. ISSN: 0925-2312. DOI: https://doi.org/10.1016/S0925-2312(02)00601-X. URL: https://www.sciencedirect.com/science/article/pii/S092523120200601X.

[5] T Filisbino, G Giraldi e C Thomaz. "Multi-class nonlinear discriminant feature analysis". Em: **38th Ibero-Latin Am. Cong. on Comp. Meth. in Eng.(CILAMCE)** (2017).

[6] T. Filisbino, G. Giraldi, C. Thomaz e D. Leite. "Multi-Class Discriminant Analysis Based on SVM Ensembles for Ranking Principal Components". Em: jan. de 2015. DOI: 10.20906/CPS/CILAMCE2015-0375.

[7] R. Fletcher. "A new variational result for quasi-Newton formulae". Em: **SIAM Journal on Optimization** 1.1 (1991), pp. 18–21. URL: https://doi.org/10.1137/0801002.

[8] A. Janowczyk. **Andrew Janowczyk website**. 2015. URL: https://www.kaggle.com/datasets/simjeg/lymphoma-subtype-classification-fl-vs-c.

[9] A. Janowczyk e A. Madabhushi. "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases". Em: **Journal of Pathology Informatics** 7.1 (2016), p. 29.

[10] S. B. Kotsiantis. "Decision trees: a recent overview". Em: **Artificial Intelligence Review** 39 (2013), pp. 261–283.

[11] D. A. Lamino, D. D. C. F. Mota e C. A. M. Pimenta. **Prevalência e comorbidade de dor e fadiga em mulheres com câncer de mama**. 2011. URL: https://doi.org/10.1590/S0080-62342011000200029.

[12] S. Mani, Y. Chen, X. Li, L. Arlinghaus, A. B. Chakravarthy, V. Abramson, S. R. Bhave, M. A. Levy, H. Xu e T. E. Yankeelov. "Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy". Em: **Journal of the American Medical Informatics Association** 20.4 (2013), pp. 688–695.

[13] R. Mendonca-Neto, J. Reis, L. Okimoto, D. Fenyö, C. Silva, F. Nakamura e E. Nakamura. "Classification of breast cancer subtypes: A study based on representative genes". Em: **Journal of the Brazilian Computer Society** 28.1 (2022), pp. 59–68.

[14] S. J. Rigatti. "Random forest". Em: **Journal of Insurance Medicine** 47.1 (2017), pp. 31–39.

[15] A. M. Romano e A. A. Hernandez. "Enhanced deep learning approach for predicting invasive ductal carcinoma from histopathology images". Em: **2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)**. IEEE. 2019, pp. 142–148.

[16] J. R. Scarff. "The prospects of cariprazine in the treatment of schizophrenia". Em: **Therapeutic Advances in Psychopharmacology** 7.11 (2017), pp. 237–239.

[17] B. Scholkopf e A. J. Smola. **Learning with kernels: support vector machines, regularization, optimization, and beyond**. MIT Press, 2018.

[18] E. A. Shishkina, A. Yu. Volchkova, D. V. Ivanov, P. Fattibene, A. Wieser, V. A. Krivoschapov, M. O. Degteva e B. A. Napier. "Application of EPR tooth dosimetry for validation of the calculated external doses: Experience in dosimetry for the Techa River Cohort". Em: **Radiation Protection Dosimetry** 186.1 (2019), pp. 70–77.

[19] X. Wang e K. Yan. "Gait classification through CNN-based ensemble learning". Em: **Multimedia Tools and Applications** 80 (2021), pp. 1565–1581.

[20] L. Xie e A. Yuille. "Genetic CNN". Em: **Proceedings of the IEEE International Conference on Computer Vision**. 2017, pp. 1379–1388.

[21] S. Zhang, X. Li, M. Zong, X. Zhu e D. Cheng. "Learning k for KNN Classification". Em: **ACM Transactions on Intelligent Systems and Technology (TIST)** 8.3 (2017), pp. 1–19.

